informs
http://pubsonline.informs.org/journal/stsy/

STOCHASTIC SYSTEMS
Vol. 8, No. 1, March 2018, pp. 45–74
ISSN 1946-5238 (print), ISSN 1946-5238 (online)

Delay, Memory, and Messaging Tradeoffs in Distributed Service Systems

David Gamarnik,[a] John N. Tsitsiklis,[b] Martin Zubeldia[b]

[a] Sloan School of Management, Massachusetts Institute of Technology, Cambridge, Massachusetts, 02139; [b] Laboratory for Information and Decision Systems, Massachusetts Institute of Technology, Cambridge, Massachusetts, 02139

Contact: gamarnik@mit.edu, http://orcid.org/0000-0003-2658-8239 (DG); jnt@mit.edu, http://orcid.org/0000-0003-1320-9893 (JNT); zubeldia@mit.edu (MZ)

Received: February 2017
Accepted: September 2017
Published Online in Articles in Advance: February 23, 2018

MSC2010 Subject Classification: 60K25

https://doi.org/10.1287/stsy.2017.0008

Copyright: © 2018 The Author(s)

Abstract. We consider the following distributed service model: jobs with unit mean, exponentially distributed, and independent processing times arrive as a Poisson process of rate λn, with $0 < \lambda < 1$, and are immediately dispatched by a centralized dispatcher to one of n First-In-First-Out queues associated with n identical servers. The dispatcher is endowed with a finite memory, and with the ability to exchange messages with the servers.

We propose and study a resource-constrained "pull-based" dispatching policy that involves two parameters: (i) the number of memory bits available at the dispatcher, and (ii) the average rate at which servers communicate with the dispatcher. We establish (using a fluid limit approach) that the asymptotic, as $n \to \infty$, expected queueing delay is zero when either (i) the number of memory bits grows logarithmically with n and the message rate grows superlinearly with n, or (ii) the number of memory bits grows superlogarithmically with n and the message rate is at least λn. Furthermore, when the number of memory bits grows only logarithmically with n and the message rate is proportional to n, we obtain a closed-form expression for the (now positive) asymptotic delay.

Finally, we demonstrate an interesting phase transition in the resource-constrained regime where the asymptotic delay is non-zero. In particular, we show that for any given $\alpha > 0$ (no matter how small), if our policy only uses a linear message rate αn, the resulting asymptotic delay is upper bounded, uniformly over all $\lambda < 1$; this is in sharp contrast to the delay obtained when no messages are used ($\alpha = 0$), which grows as $1/(1-\lambda)$ when $\lambda \uparrow 1$, or when the popular power-of-d-choices is used, in which the delay grows as $\log(1/(1-\lambda))$.

Funding: Research supported in part by an MIT Jacobs Presidential Fellowship and the NSF [Grant CMMI-1234062].

Keywords: load balancing • performance tradeoffs • large-scale systems • queueing theory

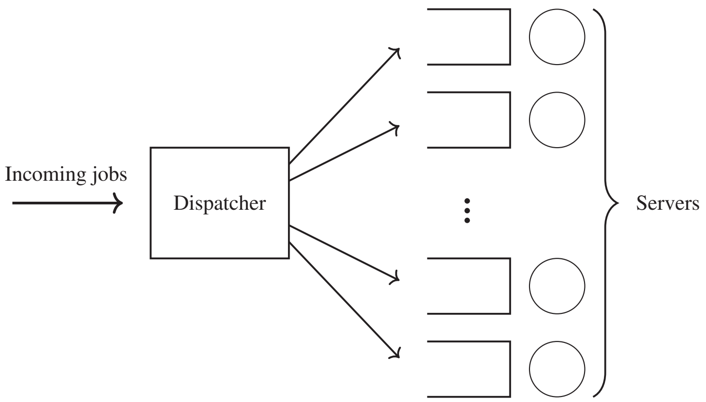
1. Introduction

This paper addresses the tradeoffs between performance (delay) and resources (local memory and communication overhead) in large-scale queueing systems. More specifically, we study such tradeoffs in the context of the supermarket model (Mitzenmacher 1996), which describes a system in which incoming jobs are to be dispatched to one of several queues associated with different servers (see Figure 1).

There is a variety of ways that the system in the supermarket model can be operated, which correspond to different decision making architectures and policies, with different delay performance. At one extreme, incoming jobs can be sent to a random queue. This policy has no informational requirements but incurs a substantial delay because it does not take advantage of resource pooling. At the other extreme, incoming jobs can be sent to a shortest queue, or to a server with the smallest workload. The latter policies have very good performance (small queueing delay), but rely on substantial information exchange.

Many intermediate policies have been explored in the literature, and they achieve different performance levels while using varying amounts of resources, including local memory and communication overhead. For example, the power-of-d-choices (Mitzenmacher 1996, Vvedenskaya et al. 1996) and its variations (Mitzenmacher et al. 2002,

Figure 1. The basic setting.



Ying et al. 2015, Mukherjee et al. 2016) have been extensively studied, including the case of non-exponential service time distributions (Bramson et al. 2013, Aghajani and Ramanan 2017). More recently, pull-based policies like Join-Idle-Queue (Badonnel and Burgess 2008, Lu et al. 2011) have been getting more attention, including extensions for heterogeneous servers (Stolyar 2015), multiple dispatchers (Mitzenmacher 2016, Van Der Boor et al. 2017, Foss and Stolyar 2017), and general service time distributions (Stolyar 2017).

1.1. Our Contribution

Our purpose is to study the effect of different resource levels (local memory and communication overhead), and to understand the amount of resources required for the asymptotic (as $n \to \infty$) delay to become negligible, in the context of the supermarket model. We adopt the average rate at which messages are exchanged between the dispatcher and the servers as our measure of the communication overhead, because of its simplicity and the fact that it applies to any kind of policy. We accomplish our purpose in two steps.

(a) In this paper, we propose a pull-based dispatching policy parameterized by the amount of resources involved, namely, the size of the memory used by the dispatcher and the average message rate. We carry out a thorough analysis in different regimes and show that we obtain vanishing asymptotic delay if and only if the resources are above a certain level.

(b) In a companion paper (see also Gamarnik et al. 2016), we show that in the regime (i.e., level of resources) where our policy fails to result in vanishing asymptotic delay, the same is true for every other policy within a broad class of "symmetric" policies that treat all servers in the same manner.

More concretely, our development relies on a fluid limit approach. As is common with fluid-based analyses, we obtain two types of results: (i) qualitative results obtained through a deterministic analysis of a fluid model, and (ii) technical results on the convergence of the actual stochastic system to its fluid counterpart.

On the qualitative end, we establish the following:

(a) If the message rate is superlinear in n and the number of memory bits is at least logarithmic in n, then the asymptotic delay is zero.

(b) If the message rate is at least λn and the number of memory bits is superlogarithmic in n, then the asymptotic delay is zero.

(c) If the message rate is αn and the number of memory bits is $c \log_2(n)$, we derive a closed form expression for the (now positive) asymptotic delay, in terms of λ, α, and c.

(d) For the same amount of resources as in (c), we show an interesting phase transition in the asymptotic delay as the load approaches capacity ($\lambda \uparrow 1$). As long as a nontrivial linear message rate αn, with $\alpha > 0$, is used, the asymptotic delay is uniformly upper bounded, over all $\lambda < 1$. This is in sharp contrast to the delay obtained if no messages are used ($\alpha = 0$), which grows as $1/(1-\lambda)$ when $\lambda \uparrow 1$. This suggests that for large systems, even a small linear message rate provides significant improvements in the system's delay performance when $\lambda \uparrow 1$.

(e) Again for the same amount of resources as in (c), we show a phase transition in the scaling of the asymptotic delay as a function of the memory parameter c, as we vary the message rate parameter α.

(i) If $\alpha < \lambda$, then the asymptotic delay is uniformly bounded away from zero, for any $c \geq 0$.

(ii) If $\alpha = \lambda$, then the asymptotic delay decreases as $1/c$, when $c \to \infty$.

(iii) If $\alpha > \lambda$, then the queueing delay decreases as $(\lambda/\alpha)^c$, when $c \to \infty$.

This suggests that a message rate of at least λn is required for the memory to have a significant impact on the asymptotic delay.

On the technical end, and for each one of three regimes corresponding to cases (a), (b), and (c) above, we show the following:

(a) The queue length process converges (as $n \to \infty$, and over any finite time interval) almost surely to the unique solution to a certain fluid model.

(b) For any initial conditions that correspond to starting with a finite average number of jobs per queue, the fluid solution converges (as time tends to ∞) to a unique invariant state.

(c) The steady-state distribution of the finite system converges (as $n \to \infty$) to the invariant state of the fluid model.

1.2. Outline of the Paper

The rest of the paper is organized as follows. In Section 2 we introduce some notation. In Section 3 we present the model and the main results, and also compare a particular regime of our policy to the so-called "power-of-d-choices" policy. In Sections 4–6 we provide the proofs of the main results. Finally, in Section 7 we present our conclusions and suggestions for future work.

2. Notation

In this section we introduce some notation that will be used throughout the paper. First, we define the notation for the asymptotic behavior of positive functions. In particular,

$$\begin{aligned}
f(n) \in o(g(n)) \quad &\Leftrightarrow \quad \limsup_{n\to\infty} \frac{f(n)}{g(n)} = 0,\\
f(n) \in O(g(n)) \quad &\Leftrightarrow \quad \limsup_{n\to\infty} \frac{f(n)}{g(n)} < \infty,\\
f(n) \in \Theta(g(n)) \quad &\Leftrightarrow \quad 0 < \liminf_{n\to\infty} \frac{f(n)}{g(n)} \le \limsup_{n\to\infty} \frac{f(n)}{g(n)} < \infty,\\
f(n) \in \Omega(g(n)) \quad &\Leftrightarrow \quad \liminf_{n\to\infty} \frac{f(n)}{g(n)} > 0,\\
f(n) \in \omega(g(n)) \quad &\Leftrightarrow \quad \liminf_{n\to\infty} \frac{f(n)}{g(n)} = \infty.
\end{aligned}$$

We let $[\,\cdot\,]^+ \triangleq \max\{\cdot, 0\}$, and denote by $\mathbb{Z}_+$ and $\mathbb{R}_+$ the sets of non-negative integers and real numbers, respectively. The indicator function is denoted by $\mathbb{1}$, so that $\mathbb{1}_A(x)$ is 1 if $x \in A$, and is 0 otherwise. The Dirac measure δ concentrated at a point x is defined by $\delta_x(A) \triangleq \mathbb{1}_A(x)$. We also define the following sets:

$$\begin{aligned}
\mathcal{S} &\triangleq \{s \in [0,1]^{\mathbb{Z}_+}\colon s_0 = 1; s_i \ge s_{i+1}, \forall\, i \ge 0\},\\
\mathcal{S}^1 &\triangleq \left\{s \in \mathcal{S}\colon \sum_{i=0}^{\infty} s_i < \infty\right\},\\
\mathcal{I}_n &\triangleq \left\{x \in [0,1]^{\mathbb{Z}_+}\colon x_i = \frac{k_i}{n}, \text{ for some } k_i \in \mathbb{Z}_+, \forall\, i\right\}.
\end{aligned} \tag{1}$$

We define the weighted ℓ_2 norm $\|\cdot\|_w$ on $\mathbb{R}^{\mathbb{Z}_+}$ by

$$\|x - y\|_w^2 \triangleq \sum_{i=0}^{\infty} \frac{|x_i - y_i|^2}{2^i}.$$

Note that this norm comes from an inner product, so $(\ell_w^2, \|\cdot\|_w)$ is actually a Hilbert space, where

$$\ell_w^2 \triangleq \{s \in \mathbb{R}^{\mathbb{Z}_+}\colon \|s\|_w < \infty\}.$$

We also define a partial order on $\mathcal{S}$ as follows:

$$\begin{aligned}
x \ge y \quad &\Leftrightarrow \quad x_i \ge y_i, \quad \forall\, i \ge 1,\\
x > y \quad &\Leftrightarrow \quad x_i > y_i, \quad \forall\, i \ge 1.
\end{aligned}$$

We will work with the Skorokhod spaces of functions

$$D[0,T] \triangleq \{f\colon [0,T] \to \mathbb{R}\colon f \text{ is right-continuous with left limits}\},$$

endowed with the uniform metric

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

and

$$D^{\infty}[0,T] \triangleq \{f\colon [0,T] \to \mathbb{R}^{\mathbb{Z}_+}\colon f \text{ is right-continuous with left limits}\},$$

with the metric

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

3. Model and Main Results

In this section we present our main results. In Section 3.1 we describe the model and our assumptions. In Section 3.2 we introduce three different regimes of a certain pull-based dispatching policy. In Sections 3.3 and 3.4 we introduce a fluid model and state the validity of fluid approximations for the transient and the steady-state regimes, respectively. In Section 3.5, we discuss the asymptotic delay, and show a phase transition in its behavior when $\lambda \uparrow 1$.

3.1. Modeling Assumptions

We consider a system consisting of n parallel servers, where each server has a processing rate equal to 1. Furthermore, each server is associated with an infinite capacity FIFO queue. We use the convention that a job that is being served remains in queue until its processing is completed. We assume that each server is work conserving: a server is idle if and only if the corresponding queue is empty.

Jobs arrive to the system as a single Poisson process of rate λn (for some fixed $\lambda < 1$). Job sizes are i.i.d., independent from the arrival process, and exponentially distributed with mean 1.

There is a central controller (dispatcher), responsible for routing each incoming job to a queue, immediately upon arrival. The dispatcher makes decisions based on limited information about the state of the queues, as conveyed through messages from idle servers to the dispatcher, and which is stored in a limited local memory. See the next subsection for the precise description of the policy.

We will focus on the steady-state expectation of the time between the arrival of a typical job and the time at which it starts receiving service (to be referred to as "*queueing delay*" or just "*delay*" for short) and its limit as the system size n tends to infinity (to be referred to as "*asymptotic delay*"). Furthermore, we are interested in the amount of resources (memory size and message rate) required for the asymptotic delay to be equal to zero.

3.2. Policy Description and High-Level Overview of the Results

In this section we introduce our policy and state in a succinct form our results for three of its regimes.

3.2.1. Policy Description. For any fixed value of n, the policy that we study operates as follows.

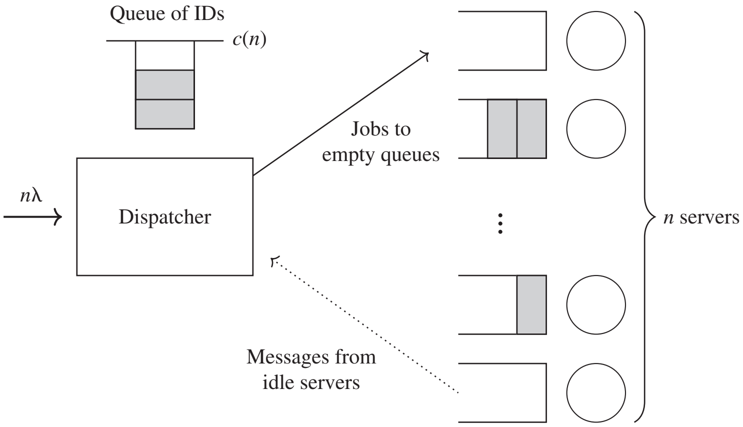
(a) *Memory*: The dispatcher maintains a virtual queue comprised of up to $c(n)$ server identity numbers (IDs), also referred to as *tokens*, so that the dispatcher's memory size is of order $c(n)\log_2(n)$ bits. Since there are only n distinct servers, we will assume throughout the rest of the paper that $c(n) \le n$.

(b) *Spontaneous messages from idle servers*: While a server is idle, it sends messages to the dispatcher as a Poisson process of rate $\mu(n)$, to inform or remind the dispatcher of its idleness. We assume that $\mu(n)$ is a nondecreasing function of n. Whenever the dispatcher receives a message, it adds the ID of the server that sent the message to the virtual queue of tokens, unless this ID is already stored or the virtual queue is full, in which cases the new message is discarded.

(c) *Dispatching rule*: Whenever a new job arrives, if there is at least one server ID in the virtual queue, the job is sent to the queue of a server whose ID is chosen uniformly at random from the virtual queue, and the corresponding token is deleted. If there are no tokens present, the job is sent to a queue chosen uniformly at random.

Note that under the above described policy, which is also depicted in Figure 2, no messages are ever sent from the dispatcher to the servers. Accordingly, following the terminology of Badonnel and Burgess (2008), we will refer to it as the Resource Constrained Pull-Based (*RCPB*) policy or *Pull-Based* policy for short.

Figure 2. Resource constrained pull-based policy. Jobs are sent to queues associated with idle servers, based on tokens in the virtual queue. If no tokens are present, a queue is chosen at random.



3.2.2. High-Level Summary of the Results.

We summarize our results for the RCPB policy, for three different regimes, in Table 1, where we also introduce some mnemonic terms that we will use to refer to these regimes. Formal statements of these results are given later in this section. Furthermore, we provide a pictorial representation of the total resource requirements and the corresponding asymptotic delays in Figure 3.

The more interesting subcase of the High Memory regime is when $\mu \geq \lambda/(1-\lambda)$, which results in zero asymptotic delay with superlogarithmic memory and linear overall message rate. Note that if we set $\mu = \lambda/(1-\lambda)$, and use the fact that servers are idle a fraction $1-\lambda$ of the time, the resulting time-average message rate becomes exactly λn, i.e., one message per arrival.

3.3. Stochastic and Fluid Descriptions of the System

In this subsection, we define a stochastic process that corresponds to our model under the RCPB policy, as well as an associated fluid model.

3.3.1. Stochastic System Representation.

Let $Q_i^n(t)$ be the number of jobs in queue i (including the job currently being served, if any), at time t, in a n-server system. We can model the system as a continuous-time Markov process whose state is the queue length vector, $Q^n(t) = (Q_i^n(t))_{i=1}^n \in \mathbb{Z}_+^n$, together with the number of tokens, denoted by $M^n(t) \in \{0, 1, \ldots, c(n)\}$. However, as the system is symmetric with respect to the queues, we will use instead the more convenient representation $S^n(t) = (S_i^n(t))_{i=0}^\infty$, where

$$S_i^n(t) \triangleq \frac{1}{n}\sum_{j=1}^n \mathbb{1}_{[i,\infty)}(Q_j^n(t)), \quad i \in \mathbb{Z}_+,$$

is the fraction of queues with at least i jobs at time t. Once more, the pair $(S^n(t), M^n(t))$ is a continuous-time Markov process, with a countable state space.

Finally, another possible state representation involves $V^n(t) = (V_i^n(t))_{i=1}^\infty$, where

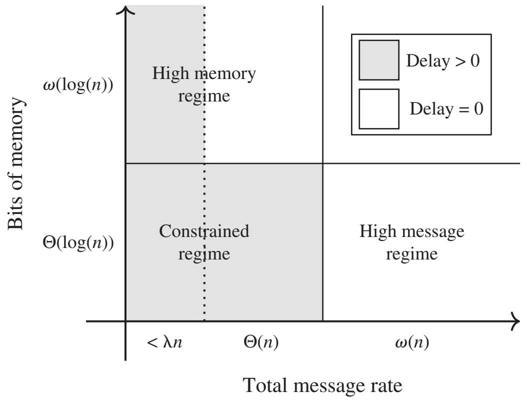
$$V_i^n(t) \triangleq \sum_{j=i}^\infty S_j^n(t)$$

can be interpreted as the average amount by which a queue length exceeds $i-1$ at time t. In particular, $V_1^n(t)$ is the total number of jobs at time t divided by n, and is finite, with probability 1.

Table 1. The three regimes of our policy, and the resulting asymptotic delays.

Regime	Memory	Idle message rate	Delay
High memory	$c(n) \in \omega(1)$ and $c(n) \in o(n)$	$\mu(n) = \mu \geq \lambda/(1-\lambda)$	0
		$\mu(n) = \mu < \lambda/(1-\lambda)$	>0
High message	$c(n) = c \geq 1$	$\mu(n) \in \omega(1)$	0
Constrained	$c(n) = c \geq 1$	$\mu(n) = \mu > 0$	>0

Figure 3. Resource requirements of the three regimes, and the resulting asymptotic delays.



3.3.2. Fluid Model. We now introduce the fluid model of $S^n(t)$, associated with our policy. Recall the definition of the set $\mathcal{S}^1$ in Equation (1).

Definition 3.1 (Fluid Model). Given an initial condition $s^0 \in \mathcal{S}^1$, a continuous function $s(t)\colon [0,\infty) \to \mathcal{S}^1$ is said to be a solution to the fluid model (or fluid solution) if:

1. $s(0) = s^0$.
2. For all $t \geq 0$, $s_0(t) = 1$.
3. For all $t \geq 0$ outside of a set of Lebesgue measure zero, and for every $i \geq 1$, $s_i(t)$ is differentiable and satisfies

$$\frac{ds_1}{dt}(t) = \lambda(1 - P_0(s(t))) + \lambda(1 - s_1(t))P_0(s(t)) - (s_1(t) - s_2(t)), \tag{2}$$

$$\frac{ds_i}{dt}(t) = \lambda(s_{i-1}(t) - s_i(t))P_0(s(t)) - (s_i(t) - s_{i+1}(t)) \quad \forall\, i \geq 2, \tag{3}$$

where $P_0(s)$ is given, for the three regimes considered, by:

(i) High Memory: $P_0(s) = [1 - \mu(1 - s_1)/\lambda]^+$;

(ii) High Message: $P_0(s) = [1 - (1 - s_2)/\lambda]^+ \mathbb{1}_{\{s_1=1\}}$;

(iii) Constrained: $P_0(s) = [\sum_{k=0}^{c} (\mu(1 - s_1)/\lambda)^k]^{-1}$.

We use the convention $0^0 = 1$, so that the case $s_1 = 1$ yields $P_0(s) = 1$.

A solution to the fluid model, $s(t)$, can be thought of as a deterministic approximation to the sample paths of the stochastic process $S^n(t)$, for n large enough. Note that the fluid model does not include a variable associated with the number of tokens. This is because, as we will see, the virtual queue process $M^n(t)$ evolves on a faster time scale than the processes of the queue lengths and does not have a deterministic limit. We thus have a process with two different time scales: on the one hand, the virtual queue evolves on a fast time scale (at least n times faster) and from its perspective the queue process $S^n(t)$ appears static; on the other hand, the queue process $S^n(t)$ evolves on a slower time scale and from its perspective, the virtual queue appears to be at stochastic equilibrium. This latter property is manifested in the drift of the fluid model: $P_0(s(t))$ can be interpreted as the probability that the virtual queue is empty when the rest of the system is fixed at the state $s(t)$. Moreover, the drift of $s_1(t)$ is qualitatively different from the drift of the other components $s_i(t)$, for $i \geq 2$, because our policy treats empty queues differently.

We now provide some intuition for each of the drift terms in Equations (2) and (3).

(i) $\lambda(1 - P_0(s(t)))$: This term corresponds to arrivals to an empty queue while there are tokens in the virtual queue, taking into account that the virtual queue is nonempty with probability $1 - P_0(s(t))$, in the limit.

(ii) $\lambda(s_{i-1}(t) - s_i(t))P_0(s(t))$: This term corresponds to arrivals to a queue with exactly $i - 1$ jobs while there are no tokens in the virtual queue. This occurs when the virtual queue is empty and a queue with $i - 1$ jobs is drawn, which happens with probability $P_0(s(t))(s_{i-1}(t) - s_i(t))$.

(iii) $-(s_i(t) - s_{i+1}(t))$: This term corresponds to departures from queues with exactly i jobs, which after dividing by n, occur at a rate equal to the fraction $s_i(t) - s_{i+1}(t)$ of servers with exactly i jobs.

(iv) Finally, the expressions for $P_0(s)$ are obtained through an explicit calculation of the steady-state distribution of $M^n(t)$ when $S^n(t)$ is fixed at the value s, while also letting $n \to \infty$.

Let us give an informal derivation of the different expressions for $P_0(s)$. Recall that $P_0(s)$ can be interpreted as the probability that the virtual queue is empty when the rest of the system is fixed at the state s. Under this interpretation, for any fixed state s, and for any fixed n, the virtual queue would behave like an $M/M/1$ queue with capacity $c(n)$, arrival rate $\mu(n)n(1-s_1)$, and departure rate λn. In this $M/M/1$ queue, the steady-state probability of being empty is

$$P_0^{(n)}(s) = \left[\sum_{k=0}^{c(n)} \left(\frac{\mu(n)(1-s_1)}{\lambda}\right)^k\right]^{-1}.$$

By taking the limit as $n \to \infty$, we obtain the correct expressions for $P_0(s)$, except in the case of the High Message regime with $s_1 = 1$. In that particular case, this simple interpretation does not work. However, we can go one step further and note that when all servers are busy (i.e., when $s_1 = 1$), servers become idle at rate $1 - s_2$, which is the proportion of servers with exactly one job left in their queues. Since the high message rate assures that messages are sent almost immediately after the server becomes idle, only a fraction $[\lambda - (1 - s_2)]/\lambda$ of incoming jobs will go to a non-empty queue, which is exactly the probability of finding an empty virtual queue in this case.

3.4. Technical Results

In this section we provide precise statements of our technical results.

3.4.1. Properties of the Fluid Solutions. The existence of fluid solutions will be established by showing that, almost surely, the limit of every convergent subsequence of sample paths of $S^n(t)$ is a fluid solution (Proposition 5.4). In addition, the theorem that follows establishes uniqueness of fluid solutions for all initial conditions $s^0 \in \mathcal{S}^1$, characterizes the unique equilibrium of the fluid model, and states its global asymptotic stability. The regimes mentioned in the statement of the results in this section correspond to the different assumptions on memory and message rates described in the 2nd and 3rd columns of Table 1, respectively.

Theorem 3.1 (Existence, Uniqueness, and Stability of Fluid Solutions). *A fluid solution, as described in Definition 3.1, exists and is unique for any initial condition $s^0 \in \mathcal{S}^1$. Furthermore, the fluid model has a unique equilibrium s^*, given by*

$$s_i^* = \lambda(\lambda P_0^*)^{i-1}, \quad \forall\, i \geq 1,$$

where $P_0^* = P_0(s^*)$ is given, for the three regimes considered, by:
(i) *High Memory: $P_0^* = [1 - \mu(1-\lambda)/\lambda]^+$;*
(ii) *High Message: $P_0^* = 0$;*
(iii) *Constrained: $P_0^* = [\sum_{k=0}^{c}(\mu(1-\lambda)/\lambda)^k]^{-1}$.*
This equilibrium is globally asymptotically stable, i.e.,

$$\lim_{t\to\infty} \|s(t) - s^*\|_w = 0,$$

for any initial condition $s^0 \in \mathcal{S}^1$.

The proof is given in Sections 4 (uniqueness and stability) and 5 (existence).

Remark 3.1. Note that, if $\mu \geq \lambda/(1-\lambda)$, the High Memory regime also has $P_0^* = 0$ in equilibrium.

3.4.2. Approximation Theorems. The three results in this section justify the use of the fluid model as an approximation to the finite stochastic system. The first one states that the evolution of the process $S^n(t)$ is almost surely uniformly close, over any finite time horizon $[0, T]$, to the unique fluid solution $s(t)$.

Theorem 3.2 (Convergence of Sample Paths). *Fix $T > 0$ and $s^0 \in \mathcal{S}^1$. Under each of the three regimes, if*

$$\lim_{n\to\infty} \|S^n(0) - s^0\|_w = 0, \quad a.s.,$$

then

$$\lim_{n\to\infty} \sup_{0\leq t\leq T} \|S^n(t) - s(t)\|_w = 0, \quad a.s.,$$

where $s(t)$ is the unique fluid solution with initial condition s^0.

The proof is given in Section 5.

Remark 3.2. On the technical side, the proof is somewhat involved because the process $(S^n(t), M^n(t))$ is not the usual density-dependent Markov process studied by Kurtz (1981) and which appears in the study of several dispatching policies (e.g., Mitzenmacher 1996, Stolyar 2015, Ying et al. 2015). This is because $M^n(t)$ is not scaled by n, and consequently evolves in a faster time scale. We are dealing instead with an infinite-level infinite-dimensional jump Markov process, which is a natural generalization of its finite-level finite-dimensional counterpart studied in Chapter 8 of Shwartz and Weiss (1995). The fact that our process may have infinitely many levels (memory states) and is infinite-dimensional prevents us from directly applying known results. Furthermore, even if we truncated $S^n(t)$ to be finite-dimensional as in Mitzenmacher et al. (2002), our process still would not satisfy the more technical hypotheses of the corresponding result in Shwartz and Weiss (1995) (Theorem 8.15). Finally, the large deviations techniques used to prove Theorem 8.15 in Shwartz and Weiss (1995) do not directly generalize to infinite dimensions. For all of these reasons, we will prove our fluid limit result directly, by using a coupling approach, as in Bramson (1998) and Tsitsiklis and Xu (2012). Our results involve a separation of time scales similar to the ones in Xu and Yun (2017) and Hunt and Kurtz (1994).

If we combine Theorems 3.2 and 3.1, we obtain that after some time, the state of the finite system $S^n(t)$ can be approximated by the equilibrium of the fluid model s^*, because

$$S^n(t) \xrightarrow{n\to\infty} s(t) \xrightarrow{t\to\infty} s^*,$$

almost surely. If we interchange the order of the limits over n and t, we obtain the limiting behavior of the invariant distribution π_s^n of $S^n(t)$ as n increases. In the next proposition and theorem, we show that the result is the same, i.e., that

$$S^n(t) \xrightarrow{t\to\infty} \pi_s^n \xrightarrow{n\to\infty} s^*,$$

in distribution, so that the interchange of limits is justified.

The first step is to show that for every finite n, the stochastic process of interest is positive recurrent.

Proposition 3.3 (Stochastic Stability). *For every n, the Markov process $(S^n(t), M^n(t))$ is positive recurrent and therefore has a unique invariant distribution π^n.*

The proof is given in Section 6.1.

Given π^n, the unique invariant distribution of the process $(S^n(t), M^n(t))$, let

$$\pi_s^n(\cdot) = \sum_{m=0}^{c(n)} \pi^n(\cdot, m)$$

be the marginal for S^n. We have the following result concerning the convergence of this sequence of marginal distributions.

Theorem 3.4 (Convergence of Invariant Distributions). *We have*

$$\lim_{n\to\infty} \pi_s^n = \delta_{s^*}, \quad \text{in distribution.}$$

The proof is given in Section 6.2.

Putting everything together, we conclude that when n is large, the fluid model is an accurate approximation to the stochastic system, for both the transient regime (Theorems 3.2 and 3.1) and the steady-state regime (Theorem 3.4). The relationship between the convergence results is depicted in the commutative diagram of Figure 4.

3.5. Asymptotic Delay and Phase Transitions

In this section we use the preceding results to conclude that in two of the regimes considered, the asymptotic delay is zero. For the third regime, the asymptotic delay is positive and we examine its dependence on various policy parameters.

Figure 4. Relationship between the stochastic system and the fluid model.

$$\begin{tikzcd}
S^n(t) \arrow[r, "\text{Theorem 3.2}", "n \to \infty"'] \arrow[d, "\text{Proposition 3.3}"', "t \to \infty"] & s(t) \arrow[d, "\text{Theorem 3.1}", "t \to \infty"'] \\
\pi_s^n \arrow[r, "\text{Theorem 3.4}", "n \to \infty"'] & s^*
\end{tikzcd}$$

3.5.1. Queueing Delay. Having shown that we can approximate the stochastic system by its fluid model for large n, we can analyze the equilibrium of the latter to approximate the queueing delay under our policy.

For any given n, we define the *queueing delay* (more precisely, the waiting time) of a job, generically denoted by $\mathbb{E}[W^n]$, as the mean time that a job spends in queue until its service starts. Here the expectation is taken with respect to the steady-state distribution, whose existence and uniqueness is guaranteed by Proposition 3.3. Then, the *asymptotic delay* is defined as

$$\mathbb{E}[W] \triangleq \limsup_{n\to\infty} \mathbb{E}[W^n].$$

This asymptotic delay can be obtained from the equilibrium s^* of the fluid model as follows. For a fixed n, the expected number of jobs in the system in steady-state is

$$\mathbb{E}\left[\sum_{i=1}^{\infty} nS_i^n\right].$$

Furthermore, the delay of a job is equal to the total time it spends in the system minus the expected service time (which is 1). Using Little's Law, we obtain that the queueing delay is

$$\mathbb{E}[W^n] = \frac{1}{\lambda n}\mathbb{E}\left[\sum_{i=1}^{\infty} nS_i^n\right] - 1 = \frac{1}{\lambda}\mathbb{E}\left[\sum_{i=1}^{\infty} S_i^n\right] - 1.$$

Taking the limit as $n \to \infty$, and interchanging the limit, summation, and expectation, we obtain

$$\mathbb{E}[W] = \frac{1}{\lambda}\left(\sum_{i=1}^{\infty} s_i^*\right) - 1. \tag{4}$$

The validity of these interchanges is established in Appendix A.

As a corollary, we obtain that if we have a superlinear message rate or a superlogarithmic number of memory bits, the RCPB policy results in zero asymptotic delay.

Corollary 3.5. *For the High Memory regime with $\mu \geq \lambda/(1-\lambda)$, and for the High Message regime, the asymptotic delay is zero, i.e., $\mathbb{E}[W] = 0$.*

Proof. From Theorem 3.1, we have $P_0^* = 0$ and therefore, $s_1^* = \lambda$ and $s_i^* = 0$, for $i \geq 2$. The result follows from Equation (4). □

3.5.2. The Asymptotic Delay in the Constrained Regime. According to Equation (4) and Theorem 3.1, the asymptotic delay is given by

$$\mathbb{E}[W] = \frac{1}{\lambda}\sum_{i=1}^{\infty} s_i^* - 1 = \sum_{i=1}^{\infty} (\lambda P_0^*)^{i-1} - 1 = \frac{\lambda P_0^*}{1-\lambda P_0^*}, \tag{5}$$

and is positive in the Constrained regime. Nevertheless, the dependence of the delay on the various parameters has some remarkable properties, which we proceed to study.

Suppose that the message rate of each idle server is $\mu = \alpha/(1-\lambda)$ for some constant $\alpha > 0$. Since a server is idle (on average) a fraction $1-\lambda$ of the time, the resulting average message rate at each server is α, and the overall (system-wide) average message rate is αn. We can rewrite the equilibrium probability P_0^* in Theorem 3.1 as

$$P_0^* = \left[1 + \frac{\alpha}{\lambda} + \cdots + \left(\frac{\alpha}{\lambda}\right)^c\right]^{-1}.$$

This, together with Equation (5) and some algebra, implies that

$$\mathbb{E}[W] = \lambda\left[1-\lambda+\frac{\alpha}{\lambda}+\cdots+\left(\frac{\alpha}{\lambda}\right)^{c}\right]^{-1}. \tag{6}$$

Phase transition of the delay for $\lambda \uparrow 1$. We have a phase transition between $\alpha = 0$ (which corresponds to uniform random routing) and $\alpha > 0$. In the first case, we have the usual $M/M/1$-queue delay: $\lambda/(1-\lambda)$. However, when $\alpha > 0$, the delay is upper bounded uniformly in λ as follows:

$$\mathbb{E}[W] \le \left(\sum_{k=1}^{c}\alpha^{k}\right)^{-1}. \tag{7}$$

This is established by noting that the expression in Equation (6) is monotonically increasing in λ and then setting $\lambda = 1$. Note that when α is fixed, the total message rate is the same, αn, for all $\lambda < 1$. This is a key qualitative improvement over all other resource constrained policies in the literature; see our discussion of the power-of-d-choices policy at the end of this subsection.

Phase transition in the memory-delay tradeoff. When λ and α are held fixed, the asymptotic delay in Equation (6) decreases with c. This represents a tradeoff between the asymptotic delay $\mathbb{E}[W]$, and the number of memory bits, which is equal to $\lceil c\log_2(n)\rceil$ for the Constrained regime. However, the rate at which the delay decreases with c depends critically on the value of α, and we have a phase transition when $\alpha = \lambda$.

(i) If $\alpha < \lambda$, then

$$\lim_{c\to\infty}\mathbb{E}[W] = \frac{\lambda(\lambda-\alpha)}{(1-\lambda)(\lambda-\alpha)+1}.$$

Consequently, if $\alpha < \lambda$, it is impossible to drive the delay to 0 by increasing the value of c, i.e., by increasing the amount of memory available.

(ii) If $\alpha = \lambda$, we have

$$\mathbb{E}[W] = \frac{1}{1-\lambda+c} \le \frac{1}{c},$$

and thus the delay converges to 0 at the rate of $1/c$, as $c \longrightarrow \infty$.

(iii) If $\alpha > \lambda$, we have

$$\mathbb{E}[W] = \lambda\left[1-\lambda+\frac{\alpha}{\lambda}+\cdots+\left(\frac{\alpha}{\lambda}\right)^{c}\right]^{-1} \le \left(\frac{\lambda}{\alpha}\right)^{c}, \tag{8}$$

and thus the delay converges exponentially fast to 0, as $c \longrightarrow \infty$.

This phase transition is due to the fact that the queueing delay depends critically on P_0^*, the probability that there are no tokens left in the dispatcher's virtual queue. In equilibrium, the number of tokens in the virtual queue evolves as a birth-death process with birth rate α, death rate λ, and maximum population c, and has an invariant distribution which is geometric with ratio α/λ. As a result, as soon as α becomes larger than λ, this birth-death process has an upward drift, and the probability of being at state 0 (no tokens present) decays exponentially with the size of its state space. This argument captures the essence of the phase transition at $\mu = \lambda/(1-\lambda)$ for the High Memory regime.

Comparison with the power-of-d-choices. The power-of-d-choices policy queries d random servers at the time of each arrival and sends the arriving job to the shortest of the queried queues. As such, it involves $2\lambda dn$ messages per unit time. For a fair comparison, we compare this policy to our RCPB policy with $\alpha = 2\lambda d$, so that the two policies have the same average message rate.

The asymptotic delay for the power-of-d-choices policy was shown in Mitzenmacher (1996), Vvedenskaya et al. (1996) to be

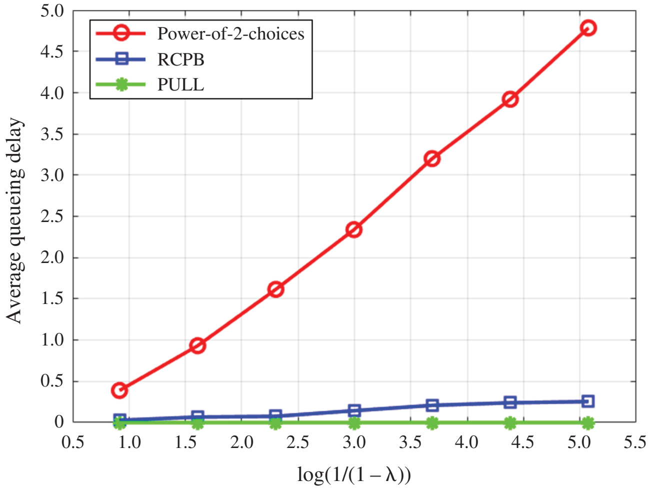
$$\mathbb{E}[W_{\mathrm{Pod}}] = \sum_{i=1}^{\infty}\lambda^{(d^{i}-d)/(d-1)} - 1 \ge \lambda^{d}.$$

Thus, the delay decreases at best exponentially with d, much like the delay decreases exponentially with c in our scheme (cf. Equation (8)). However, increasing d increases the number of messages sent, unlike our policy where the average message rate remains fixed at αn.

Furthermore, the asymptotic delay in the power-of-d-choices when $\lambda \uparrow 1$ is shown in Mitzenmacher (1996) to satisfy

$$\lim_{\lambda\uparrow 1}\frac{\mathbb{E}[W_{\mathrm{Pod}}]}{\log(1/(1-\lambda))} = \frac{1}{\log d}.$$

Figure 5. Average delay of the power-of-2-choices policy (red circles) vs. our policy (blue squares) vs. PULL (green asterisks).



For any fixed d, this is an exponential improvement over the delay of randomized routing, but the delay is still unbounded as $\lambda \uparrow 1$. In contrast, the delay of our scheme has a constant upper bound, independent of λ.

In conclusion, if we set $\alpha = 2d\lambda$, so that our policy and the power-of-d policy use the same number of messages per unit of time, our policy results in much better asymptotic delay, especially when $\lambda \uparrow 1$, even if c is as small as 1.

Numerical results. We implemented three policies in Matlab: the power-of-2-choices (Mitzenmacher 1996, Vvedenskaya et al. 1996), our RCPB policy, and the PULL policy (Stolyar 2015). We evaluate the algorithms in a system with 500 servers. In our algorithm we used $c = 2$, and $\alpha = \lambda$, so it has the same average message rate as the PULL policy (500λ messages per unit of time), which is 4 times less than what the power-of-2-choices utilizes. In Figure 5 we plot the delay as a function of $\log(1/(1-\lambda))$.

As expected, the delay remains uniformly bounded under our RCPB policy (blue squares). This is achieved with only $\lceil 2\log_2(500)\rceil = 18$ bits of memory. Furthermore, with this small amount of memory we are also close to the performance of the PULL algorithm, which requires 500 bits of memory.

4. Fluid Model Analysis—Proof of Part of Theorem 3.1

The proof of Theorem 3.1 involves mostly deterministic arguments; these are developed in Lemmas 4.1 and 4.3, and Proposition 4.5, which establish uniqueness of fluid solutions, existence and uniqueness of a fluid-model equilibrium, and asymptotic stability, respectively. The proof of existence of fluid solutions relies on a stochastic argument and is developed in Section 5, in parallel with the proof of Theorem 3.2.

4.1. Uniqueness of Solutions

Lemma 4.1. *If there exists a fluid solution (cf. Definition 3.1) with initial condition $s^0 \in \mathscr{S}^1$, it is unique.*

Proof. The fluid model is of the form $\dot{s}(t) = F(s(t))$, where the function $F\colon \mathscr{S}^1 \to [-1,\lambda]^{\mathbb{Z}_+}$ is defined by

$$\begin{aligned} F_0(s) &= 0,\\ F_1(s) &= \lambda(1-P_0(s)) + \lambda(1-s_1)P_0(s) - (s_1 - s_2),\\ F_i(s) &= \lambda(s_{i-1} - s_i)P_0(s) - (s_i - s_{i+1}), \quad \forall\, i \geq 2, \end{aligned} \tag{9}$$

and where $P_0(s)$ is given for the three regimes by:

(i) High Memory: $P_0(s) = [1 - \mu(1-s_1)/\lambda]^+$.
(ii) High Message: $P_0(s) = [1 - (1-s_2)/\lambda]^+ \mathbb{1}_{\{s_1=1\}}$.
(iii) Constrained: $P_0(s) = [\sum_{k=0}^{c} (\mu(1-s_1)/\lambda)^k]^{-1}$.

The function $P_0(s)$ for the High Memory regime is continuous and piecewise linear in s_1, so it is Lipschitz continuous in s, over the set $\mathscr{S}^1$. Similarly, $P_0(s)$ for the Constrained regime is also Lipschitz continuous in s, because $P_0(s)$ is a rational function of s_1 and the denominator is lower bounded by 1. However, $P_0(s)$ for the

High Message regime is only Lipschitz continuous "almost everywhere" in $\mathscr{S}^1$; more precisely, it is Lipschitz continuous everywhere except on the lower dimensional set

$$D \triangleq \{s \in \mathscr{S}^1 \colon s_1 = 1 \text{ and } s_2 > 1 - \lambda\}.$$

Moreover, $P_0(s)$ restricted to D is also Lipschitz continuous.

Suppose that $P_0(s)$ is Lipschitz continuous with constant L on some subset $\mathscr{S}_0$ of $\mathscr{S}^1$. Then, for every $s, s' \in \mathscr{S}_0$ and any $i \geq 1$, we have

$$\begin{aligned}|F_i(s) - F_i(s')| &= |-\lambda P_0(s)\mathbb{1}_{i=1} + \lambda(s_{i-1} - s_i)P_0(s) - (s_i - s_{i+1}) + \lambda P_0(s')\mathbb{1}_{i=1} - \lambda(s'_{i-1} - s'_i)P_0(s') + (s'_i - s'_{i+1})| \\ &\leq |P_0(s) - P_0(s')| + |(s_{i-1} - s_i)P_0(s) - (s'_{i-1} - s'_i)P_0(s')| + |s_i - s'_i| + |s_{i+1} - s'_{i+1}| \\ &\leq 2|P_0(s) - P_0(s')| + |s_{i-1} - s'_{i-1}| + 2|s_i - s'_i| + |s_{i+1} - s'_{i+1}| \\ &\leq 2L\|s - s'\|_w + |s_{i-1} - s'_{i-1}| + 2|s_i - s'_i| + |s_{i+1} - s'_{i+1}|.\end{aligned}$$

Then,

$$\begin{aligned}\|F(s) - F(s')\|_w &= \sqrt{\sum_{i=0}^{\infty} \frac{|F_i(s) - F_i(s')|^2}{2^i}} \\ &\leq \sqrt{\sum_{i=1}^{\infty} \frac{(2L\|s - s'\|_w + |s_{i-1} - s'_{i-1}| + 2|s_i - s'_i| + |s_{i+1} - s'_{i+1}|)^2}{2^i}} \\ &\leq \sqrt{12 \sum_{i=1}^{\infty} \frac{4L^2\|s - s'\|_w^2 + |s_{i-1} - s'_{i-1}|^2 + 4|s_i - s'_i|^2 + |s_{i+1} - s'_{i+1}|^2}{2^i}} \\ &\leq \|s - s'\|_w \sqrt{12(4L^2 + 2 + 4 + 1)},\end{aligned}$$

where the second inequality comes from the fact that $(w + x + y + z)^2 \leq 12(w^2 + x^2 + y^2 + z^2)$, for all $(w, x, y, z) \in \mathbb{R}^4$. This means that F is also Lipschitz continuous on the set $\mathscr{S}_0$.

For the High Memory and Constrained regimes, we can set $\mathscr{S}_0 = \mathscr{S}^1$, and by the preceding discussion, F is Lipschitz continuous on $\mathscr{S}^1$. At this point we cannot immediately guarantee the uniqueness of solutions because F is just Lipschitz continuous on a subset ($\mathscr{S}^1$) of the Hilbert space $(\ell_w^2, \|\cdot\|_w)$. However, we can use Kirszbraun's theorem (Kirszbraun 1934) to extend F to a Lipschitz continuous function $\bar{F}$ on the entire Hilbert space. If we have two different solutions to the equation $\dot{s} = F(s)$ which stay in $\mathscr{S}^1$, we would also have two different solutions to the equation $\dot{s} = \bar{F}(s)$. Since $\bar{F}$ is Lipschitz continuous, this would contradict the Picard-Lindelöff uniqueness theorem (Lobanov and Smolyanov 1994). This establishes the uniqueness of fluid solutions for the High Memory and Constrained regimes.

Note that the preceding argument can also be used to show uniqueness of solutions for any differential equation with a Lipschitz continuous drift in an arbitrary subset of the Hilbert space $(\ell_w^2, \|\cdot\|_w)$, as long as we only consider solutions that stay in that set. This fact will be used in the rest of the proof.

From now on, we concentrate on the High Message regime. In this case, the drift $F(s)$ is Lipschitz continuous only "almost everywhere," and a solution will in general be non-differentiable. In particular, results on the uniqueness of classical (differentiable) solutions do not apply. Our proof will rest on the fact that non-uniqueness issues can only arise when a trajectory hits the closure of the set where the drift $F(s)$ is not Lipschitz continuous, which in our case is just the closure of D:

$$\bar{D} = \{s \in \mathscr{S}^1 \colon s_1 = 1 \text{ and } s_2 \geq 1 - \lambda\}.$$

We now partition the space $\mathscr{S}^1$ into three subsets, $\mathscr{S}^1 \backslash \bar{D}$, D, and $\bar{D} \backslash D$, and characterize the behavior of potential trajectories depending on the initial condition. Note that we only consider fluid solutions, and these always stay in the set $\mathscr{S}^1$, by definition. Therefore, we only need to establish the uniqueness of solutions that stay in $\mathscr{S}^1$.

Claim 4.2. *For any fluid solution $s(t)$ in the High Message regime, and with initial condition $s^0 \in \bar{D}$, we have the following.*

(i) *If $s^0 \in D$, then $s(t)$ either stays in D forever or hits $\bar{D} \backslash D$ at some finite time. In particular, it cannot go directly from D to $\mathscr{S}^1 \backslash \bar{D}$.*

(ii) *If $s^0 \in \bar{D} \backslash D$, then $s(t)$ stays in $\mathscr{S}^1 \backslash D$ forever. In particular, it can never return to D.*

Proof.

(i) Suppose that $s^0 \in D$, i.e., $s_1^0 = 1$ and $s_2^0 > 1 - \lambda$. Let t_{D^c} be the exit time from D, and suppose that it is finite. Note that, by continuity of solutions, $s_1(t_{D^c}) = 1$. We will show that $s_2(t_{D^c}) = 1 - \lambda$, so that the trajectory hits $\bar{D} \backslash D$. Suppose, in order to derive a contradiction, that this is not the case and, therefore, $s_2(t_{D^c}) > 1 - \lambda$. Then, due to the continuity of solutions, there exists some time $t_1 > t_{D^c}$ such that $s_1(t_1) < 1$ and $s_2(t) > 1 - \lambda$, for all $t \in [t_{D^c}, t_1]$. Let

$$t_0 \triangleq \sup\{t \le t_1 \colon s_1(t) = 1\}$$

be the last time before t_1 that $s_1(t)$ is equal to 1. Then we have $s_1(t_0) = 1$, and $s_1(t) < 1$ for all $t \in (t_0, t_1]$. Since the drift F is continuous for all $s_1 < 1$, all times in $(t_0, t_1]$ are regular. On the other hand, for all $t \in (t_0, t_1]$, we have $s_1(t) < 1$ and thus $P_0(s(t)) = 0$, which together with $s_2(t) > 1 - \lambda$ implies that

$$\frac{ds_1(t)}{dt} = \lambda - (s_1(t) - s_2(t)) > 0,$$

for all $t \in (t_0, t_1]$. This contradicts the relations $s_1(t_1) < 1 = s_1(t_0)$, and establishes that $s_1(t_D) = 1$. Therefore the fluid solution s either stays in D forever or it exits D with $s_2 = 1 - \lambda$.

(ii) Suppose now that $s^0 \in \bar{D} \backslash D$, i.e., $s_1^0 = 1$ and $s_2^0 = 1 - \lambda$. It is enough to show that $s_2(t) \le 1 - \lambda$, for all $t \ge 0$. Let

$$\tau_2(\epsilon) \triangleq \min\{t \ge 0 \colon s_2(t) = 1 - \lambda + \epsilon\}$$

be the first time s_2 reaches $1 - \lambda + \epsilon$. Suppose, in order to derive a contradiction, that there exists $\epsilon^* > 0$ such that $\tau_2(\epsilon^*) < \infty$. Then, due to the continuity of s_2, we also have $\tau_2(\epsilon) < \infty$, for all $\epsilon \le \epsilon^*$. Since s_2 is differentiable almost everywhere, we can choose ϵ such that $\tau_2(\epsilon)$ is a regular time with $F_2(s(\tau_2(\epsilon))) > 0$. Using the expression (9) for F_2, we obtain

$$\begin{aligned}
0 &< \lambda(s_1(\tau_2(\epsilon)) - s_2(\tau_2(\epsilon)))\left(1 - \frac{1 - s_2(\tau_2(\epsilon))}{\lambda}\right)\mathbb{1}_{\{s_1(\tau_2(\epsilon))=1\}} - (s_2(\tau_2(\epsilon)) - s_3(\tau_2(\epsilon)))\\
&\le \lambda(1 - s_2(\tau_2(\epsilon)))\left(1 - \frac{1 - s_2(\tau_2(\epsilon))}{\lambda}\right) - (s_2(\tau_2(\epsilon)) - s_3(\tau_2(\epsilon)))\\
&= \lambda - 1 + s_3(\tau_2(\epsilon)) + s_2(\tau_2(\epsilon))(1 - \lambda - s_2(\tau_2(\epsilon)))\\
&< \lambda - 1 + s_3(\tau_2(\epsilon)),
\end{aligned}$$

or $s_3(\tau_2(\epsilon)) > 1 - \lambda$. On the other hand, we have $s_3(0) \le s_2(0) = 1 - \lambda$. Combining these two facts, we obtain that $s_3(\tau_2(\epsilon)) > s_3(0)$, i.e., that s_3 increased between times 0 and $\tau_2(\epsilon)$. As a result, and since s_3 is differentiable almost everywhere, there exists another regular time $\tau_3(\epsilon) \le \tau_2(\epsilon)$ such that $s_3(\tau_3(\epsilon)) > 1 - \lambda$ and $F_3(s(\tau_3(\epsilon))) > 0$. Proceeding inductively, we can obtain a sequence of nonincreasing regular times $\tau_2(\epsilon) \ge \tau_3(\epsilon) \ge \cdots \ge 0$ such that $s_k(\tau_k(\epsilon)) > 1 - \lambda$, for all $k \ge 2$. Let $\tau_\infty(\epsilon)$ be the limit of this sequence of regular times. Since all coordinates of the fluid solutions are Lipschitz continuous with the same constant L, we have

$$s_k(\tau_\infty) > 1 - \lambda - L(\tau_k(\epsilon) - \tau_\infty),$$

for all $k \ge 2$. Since $\tau_k(\epsilon) \to \tau_\infty$, there exists some $k^* \ge 2$ such that $s_k(\tau_\infty) > (1 - \lambda)/2 > 0$, for all $k \ge k^*$. But then,

$$\|s(\tau_\infty)\|_1 \ge \sum_{k=k^*}^{\infty} \frac{1 - \lambda}{2} = \infty.$$

This contradicts the fact that $s(\tau_\infty) \in \mathscr{S}^1$, and it follows that we must have $s_2(t) \le 1 - \lambda$ for all $t \ge 0$. □

The uniqueness of a solution over the whole time interval $[0, \infty)$ for the High Message regime can now be obtained by concatenating up to three unique trajectories, depending on the initial condition s^0.

(a) Suppose that $s^0 \in \mathscr{S}^1 \backslash \bar{D}$, and let $t_{\bar{D}}$ be the hitting time of $\bar{D}$, i.e.,

$$t_{\bar{D}} = \inf\{t \ge 0 \colon s(t) \in \bar{D} \text{ with } s(0) = s^0\}.$$

Since $F|_{\mathscr{S}^1 \backslash \bar{D}}$ (the restriction of the original drift F to the set $\mathscr{S}^1 \backslash \bar{D}$) is Lipschitz continuous, we have the uniqueness of a solution over the time interval $[0, t_{\bar{D}})$, by using the same argument as for the other regimes. If $t_{\bar{D}} = \infty$, then we are done. Otherwise, we have $s(t_{\bar{D}}) \in \bar{D}$; the uniqueness of a solution over the time interval $[t_{\bar{D}}, \infty)$ will immediately follow from the uniqueness of a solution with initial condition in $\bar{D}$.

(b) Suppose that $s^0 \in D$. Due to part (i) of Claim 4.2, a solution can only exit the set D by hitting $\bar{D} \backslash D$, and never by going back directly into $\mathscr{S}^1 \backslash \bar{D}$. Let $t_{\bar{D} \backslash D}$ be the hitting time of $\bar{D} \backslash D$. Since $F|_D$ is Lipschitz continuous, we have uniqueness of a solution over the time interval $[0, t_{\bar{D}\backslash D})$. As in case (a), if $t_{\bar{D}\backslash D} = \infty$ we are done. Otherwise, the uniqueness of a solution over the time interval $[t_{\bar{D}\backslash D}, \infty)$ will immediately follow from the uniqueness of a solution with initial condition in $\bar{D}\backslash D$.

(c) Suppose that $s^0 \in \bar{D}\backslash D$. Due to part (ii) of Claim 4.2, a solution stays in $\mathcal{S}^1\backslash D$ forever. As a result, since $F|_{\mathcal{S}^1\backslash D}$ is Lipschitz continuous, uniqueness follows. □

The intuition behind the preceding proof, for the High Message regime, is as follows. A non-differentiable solution may arise if the system starts with a large fraction of the servers having at least two jobs. In that case, the rate $s_1(t) - s_2(t)$ at which the servers become idle is smaller than the rate λ at which idle servers become busy. As a consequence, the fraction $s_1(t)$ of busy servers increases until it possibly reaches its maximum of 1, and stays there until the fraction of servers with exactly one job, which is now $1 - s_2(t)$, exceeds the total arrival rate λ; after that time servers become idle at a rate faster than the arrival rate. This scenario is illustrated in Figure 6.

4.2. Existence, Uniqueness, and Characterization of an Equilibrium

Lemma 4.3. *The fluid model has a unique equilibrium $s^* \in \mathcal{S}^1$, given by*

$$s_i^* = \lambda(\lambda P_0^*)^{i-1}, \quad \forall\, i \geq 1,$$

where $P_0^* \triangleq P_0(s^*)$ is given by

(i) *High Memory: $P_0^* = [1 - \mu(1-\lambda)/\lambda]^+$.*
(ii) *High Message: $P_0^* = 0$.*
(iii) *Constrained: $P_0^* = [\sum_{k=0}^{c} (\mu(1-\lambda)/\lambda)^k]^{-1}$.*

Proof. A point $s^* \in \mathcal{S}^1$ is an equilibrium if and only if

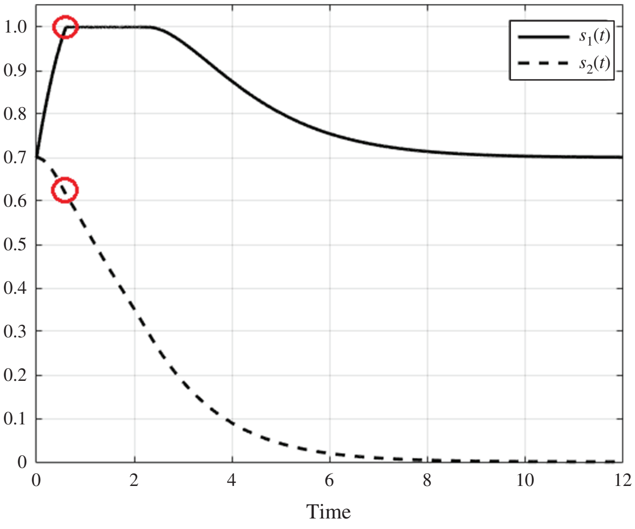
$$0 = \lambda(1 - P_0(s^*)) + \lambda(1 - s_1^*)P_0(s^*) - (s_1^* - s_2^*), \qquad 0 = \lambda(s_{i-1}^* - s_i^*)P_0(s^*) - (s_i^* - s_{i+1}^*), \quad \forall\, i \geq 2.$$

Since $s^* \in \mathcal{S}^1$, the sum $\sum_{i=0}^{\infty}(s_i^* - s_{i+1}^*)$ is absolutely convergent, even when we consider all the terms separately, i.e., when we consider s_i^* and $-s_{i+1}^*$ as separate terms, for each $i \geq 0$. Thus, we can obtain equivalent equilibrium conditions by summing these equations over all coordinates $j \geq i$, for any fixed $i \geq 1$. We then obtain that s^* is an equilibrium if and only if

$$0 = \lambda(1 - P_0(s^*)) + \lambda P_0(s^*) \sum_{j=1}^{\infty} (s_{j-1}^* - s_j^*) - \sum_{j=1}^{\infty} (s_j^* - s_{j+1}^*), \tag{10}$$

$$0 = \lambda P_0(s^*) \sum_{j=i}^{\infty} (s_{j-1}^* - s_j^*) - \sum_{j=i}^{\infty} (s_j^* - s_{j+1}^*), \quad \forall\, i \geq 2. \tag{11}$$

Figure 6. An example of a non-differentiable solution for the high message regime, with $\lambda = 0.9$, $s_1(0) = s_2(0) = s_3(0) = 0.7$, and $s_i(0) = 0$ for all $i \geq 4$.



Note. The solution is non-differentiable at the points indicated by the circles.

Since the sums are absolutely convergent, we can rearrange the terms in Equations (10) and (11) to obtain that $s^* \in \mathscr{S}^1$ is an equilibrium if and only if

$$0 = \lambda - s_1^*, \qquad 0 = \lambda P_0(s^*) s_{i-1}^* - s_i^*, \quad \forall\, i \geq 2.$$

These conditions yield $s_1^* = \lambda < 1$, and

$$s_i^* = \lambda(\lambda P_0(s^*))^{i-1}, \quad \forall\, i \geq 1,$$

which concludes the proof. □

4.3. Asymptotic Stability of the Equilibrium

We will establish global asymptotic stability by sandwiching a fluid solution between two solutions that converge to s^*, similar to the argument in Vvedenskaya et al. (1996). Towards this purpose, we first establish a monotonicity result.

Lemma 4.4. *Suppose that s^1 and s^2 are two fluid solutions with $s^1(0) \geq s^2(0)$. Then $s^1(t) \geq s^2(t)$, for all $t \geq 0$.*

Proof. It is known that uniqueness of solutions implies their continuous dependence on initial conditions, not only for the classical solutions in the High Memory and Constrained regimes, but also for the non-differentiable solutions of the High Message regime (see Chapter 8 of Filippov 1988). Using this fact, it can be seen that it is enough to verify that $s^1(t) \geq s^2(t)$ when $s^1(0) > s^2(0)$, which we henceforth assume, under our particular definition of ">" in Section 2. Let us define

$$t_1 = \inf\{t \geq 0\colon\ s_k^1(t) < s_k^2(t), \text{ for some } k \geq 1\}.$$

If $t_1 = \infty$, then $s^1(t) \geq s^2(t)$ for all $t \geq 0$, and the result holds. It remains to consider the case where $t_1 < \infty$, which we assume from now on.

By the definition of t_1, we have $s_i^1(t) \geq s_i^2(t)$ for all $i \geq 1$, and for all $t \leq t_1$. Since $P_0(s)$ is nondecreasing in s, this implies that $P_0(s^1(t)) \geq P_0(s^2(t))$, for all $t \leq t_1$. Then, for all regular times $t \leq t_1$ and any $i \geq 2$, and also using the fact that s_i is nonincreasing in i, we have

$$\begin{aligned} F_i(s^1(t)) - F_i(s^2(t)) &= \lambda[s_{i-1}^1(t) - s_i^1(t)]P_0(s^1(t)) + [s_{i+1}^1(t) - s_{i+1}^2(t)] - \lambda[s_{i-1}^2(t) - s_i^2(t)]P_0(s^2(t)) - [s_i^1(t) - s_i^2(t)] \\ &\geq \lambda[s_{i-1}^1(t) - s_i^1(t)]P_0(s^2(t)) - \lambda[s_{i-1}^2(t) - s_i^2(t)]P_0(s^2(t)) - [s_i^1(t) - s_i^2(t)] \\ &\geq -\lambda P_0(s^2(t))[s_i^1(t) - s_i^2(t)] - [s_i^1(t) - s_i^2(t)] \\ &\geq -2[s_i^1(t) - s_i^2(t)]. \end{aligned}$$

Then, by Grönwall's inequality we have

$$s_i^1(t) - s_i^2(t) \geq e^{-2t}[s_i^1(0) - s_i^2(0)], \quad \forall\, i \geq 2, \tag{12}$$

for all $t \leq t_1$. This implies that $s_i^1(t) - s_i^2(t) > 0$, for all $i \geq 2$ and for all $t \leq t_1$. It follows that, at time t_1, we must have $s_1^1(t_1) = s_1^2(t_1)$. The rest of the proof considers separately two different cases.

Case 1: Suppose that we are dealing with the High Memory or the Constrained regime, or with the High Message regime with $s_1^1(t_1) = s_1^2(t_1) < 1$. Since $s_1^1(t_1) = s_1^2(t_1)$, we have $P_0(s^1(t_1)) = P_0(s^2(t_1))$. Then, due to the continuity of s^1, s^2, and of P_0 (local continuity for the High Message regime), there exists $\epsilon > 0$ such that

$$\lambda s_1^2(t)P_0(s^2(t)) - \lambda s_1^1(t)P_0(s^1(t)) - [s_1^1(t) - s_1^2(t)] > -\epsilon,$$

and (using Equation (12)) $s_2^1(t) - s_2^2(t) > \epsilon$, for all $t \leq t_1$ sufficiently close to t_1. As a result, we have

$$F_1(s^1(t)) - F_1(s^2(t)) = \lambda s_1^2(t)P_0(s^2(t)) - \lambda s_1^1(t)P_0(s^1(t)) - [s_1^1(t) - s_1^2(t)] + [s_2^1(t) - s_2^2(t)] > 0, \tag{13}$$

for all $t < t_1$ sufficiently close to t_1. Therefore, $s_1^1 - s_1^2$ was increasing just before t_1. On the other hand, from the definition of t_1, we have $s_1^1(t_1) = s_1^2(t_1)$ and $s_1^1(t) \geq s_1^2(t)$ for all $t < t_1$. This is a contradiction, and therefore this case cannot arise.

Case 2: Suppose now that we are dealing with the High Message regime, and that $s_1^1(t_1) = s_1^2(t_1) = 1$. Since $t_1 < \infty$, we can pick a time $t_2 > t_1$, arbitrarily close to t_1, such that $s_1^1(t_2) < s_1^2(t_2)$. Let us define

$$t_1' = \sup\{t \le t_2 \colon s_1^1(t) = s_1^2(t)\}.$$

Due to the continuity of s^1 and s^2, and since $s_1^1(t_1') = s_1^2(t_1')$ and $s_2^1(t_1) > s_2^2(t_1)$, there exists $\epsilon > 0$ such that $s_1^2(t) - s_1^1(t) < \epsilon$ and $s_2^1(t) - s_2^2(t) > \epsilon$, for all $t \in [t_1', t_2]$ (we can always take a smaller t_2, if necessary, so that this holds). Furthermore, since $s_1^1(t) < 1$ for all $t \in [t_1', t_2]$, we have $P_0(s^1(t)) = 0$, for all $t \in [t_1', t_2]$. Using these facts in Equation (13), we obtain $F_1(s^1(t)) - F_1(s^2(t)) \ge 0$, for all $t \in [t_1', t_2]$. Therefore, $s_1^1 - s_1^2$ is nondecreasing in that interval. This is a contradiction, because we have $s_1^1(t_1') = s_1^2(t_1')$ and $s_1^1(t_2) < s_1^2(t_2)$. Therefore, this case cannot arise either. □

We will now show that we can "sandwich" any given trajectory $s(t)$ between a smaller one $s^l(t)$ and a larger one $s^u(t)$ (according to our partial order $\ge$) and prove that both $s^l(t)$ and $s^u(t)$ converge to s^*, to conclude that $s(t)$ converges to s^*.

Proposition 4.5. *The equilibrium s^* of the fluid model is globally asymptotically stable, i.e.,*

$$\lim_{t\to\infty} \|s(t) - s^*\|_w = 0,$$

for all fluid solutions $s(\cdot)$.

Proof. Suppose that $s(0) = s^0 \in \mathcal{S}^1$. We define initial conditions $s^u(0)$ and $s^l(0)$ by letting

$$s_i^u(0) = \max\{s_i(0), s_i^*\}, \qquad \text{and} \qquad s_i^l(0) = \min\{s_i(0), s_i^*\},$$

for all i. We then have $s^u(0) \ge s^0 \ge s^l(0)$, $s^u(0) \ge s^* \ge s^l(0)$, and $s^u(0), s^l(0) \in \mathcal{S}^1$. Due to monotonicity (Lemma 4.4), we obtain that $s^u(t) \ge s(t) \ge s^l(t)$ and $s^u(t) \ge s^* \ge s^l(t)$ for all $t \ge 0$. Thus it suffices to prove that $\|s^u(t) - s^*\|_w$ and $\|s^l(t) - s^*\|_w$ converge to 0 as $t \to \infty$.

For any $s \in \mathcal{S}^1$, we introduce an equivalent representation in terms of a vector v with components v_i defined by

$$v_i \triangleq \sum_{j=i}^{\infty} s_j, \quad i \ge 1.$$

Note that any $s \in \mathcal{S}^1$ can be fully recovered from v. Therefore, we can work with a representation $v^u(t)$, $v^l(t)$, and v^*, of the vectors $s^u(t)$, $s^l(t)$, and s^*, respectively.

From the proof of Lemma 4.1, we know that a trajectory can be non-differentiable at most at a single point in time. This can occur only for the High Message regime, and only if the trajectory hits the set

$$D = \{s \in \mathcal{S}^1 \colon s_1 = 1 \text{ and } s_2 > 1 - \lambda\},$$

where the drift is discontinuous. In all other cases, the trajectories are not only differentiable, but also Lipschitz continuous (in time), with the same Lipschitz constant for all coordinates. Therefore, in order to prove the asymptotic stability of the solutions, which is a property of the limiting behavior as $t \to \infty$, we can assume that the trajectories are everywhere differentiable and Lipschitz continuous.

Our first step is to derive a differential equation for v_i. This requires the interchange of summation and differentiation, which we proceed to justify. For any $i \ge 1$, we define a sequence of functions $\{f_k^{(i)}\}_{k=1}^{\infty}$, as follows:

$$f_k^{(i)}(t) \triangleq \sum_{j=i}^{k} \frac{ds_j^u}{dt}(t).$$

Using Equations (2) and (3), we obtain

$$\begin{aligned} f_k^{(1)}(t) &= \lambda - s_1^u(t) + [s_{n+1}^u(t) - \lambda s_k^u(t) P_0(s^u(t))], \\ f_k^{(i)}(t) &= \lambda s_{i-1}^u(t) P_0(s^u(t)) - s_i^u(t) + [s_{k+1}^u(t) - \lambda s_k^u(t) P_0(s^u(t))], \quad \forall\, i \ge 2. \end{aligned}$$

Since $s^u(t) \in \mathcal{S}^1$, for all t, we have the pointwise limits

$$\lim_{k\to\infty} f_k^{(1)}(t) = \lambda - s_1^u(t), \qquad \lim_{k\to\infty} f_k^{(i)}(t) = \lambda s_{i-1}^u(t) P_0(s^u(t)) - s_i^u(t), \quad \forall\, i \ge 2.$$

On the other hand, since all components of $s^u(\cdot)$ are Lipschitz continuous with the same constant, and since $P_0(s)$ is also Lipschitz-continuous, the functions in the sequence $\{f_k^{(i)}\}_{k=1}^{\infty}$ are equicontinuous, for any given i. Then, the Arzelà-Ascoli theorem allows us to conclude that $f_k^{(i)}(\cdot)$ also converges uniformly, over any compact interval of time, to their pointwise limits. Using the uniform convergence, and the fact that $s^u(0) \in \mathscr{S}^1$, we can interchange summation and differentiation (Theorem 7.17 in Rudin 1976) to obtain

$$\frac{dv_1^u}{dt}(t) = \frac{d}{dt}\sum_{j=1}^{\infty} s_j^u(t) = \sum_{j=1}^{\infty}\frac{ds_j^u}{dt}(t) = \lambda - s_1^u(t)$$
$$\frac{dv_i^u}{dt}(t) = \frac{d}{dt}\sum_{j=i}^{\infty} s_j^u(t) = \sum_{j=i}^{\infty}\frac{ds_j^u}{dt}(t) = \lambda s_{i-1}^u(t)P_0(s^u(t)) - s_i^u(t), \quad \forall\, i \ge 2.$$

Turning the above differential equations into integral equations, and using the facts $s_1^* = \lambda$ and $\lambda s_{i-1}^* P_0^* - s_i^* = 0$, we have

$$v_1^u(t) - v_1^u(0) = \int_0^t (s_1^* - s_1^u(\tau))\, d\tau,$$
$$v_i^u(t) - v_i^u(0) = \int_0^t (\lambda(s_{i-1}^u(\tau)P_0(s^u(\tau)) - s_{i-1}^* P_0^*) - (s_i^u(\tau) - s_i^*))\, d\tau, \quad \forall\, i \ge 2.$$

Note that from the definition of v_i, we have $v_1^u(t) \ge v_i^u(t)$. Furthermore, from Lemma 4.4, we have $s_1^u(t) \ge s_1^*$, so that $\dot{v}_1^u(t) \le 0$, for all $t \ge 0$. It follows that

$$v_1^u(0) \ge v_1^u(t) \ge v_i^u(t) \ge v_i^u(t) - v_i^u(0) \ge -v_i^u(0),$$

for all t.

We will now use induction on i to prove coordinate-wise convergence, i.e., that $|s_i^u(t) - s_i^*|$ converges to 0 for all $i \ge 1$. We start with the base case, $i = 1$. We have $s_1^u(\tau) - s_1^* \ge 0$, for all $\tau \ge 0$. Using the fact $\dot{v}_1^u(t) \le 0$, we see that $v_1^u(t)$ converges to some limit, which we denote by $v_1^u(\infty)$. Then,

$$0 \le \int_0^{\infty} (s_1^u(\tau) - s_1^*)\, d\tau = v_1^u(0) - v_1^u(\infty) \le v_1^u(0) < \infty,$$

which, together with the fact that s_1 is Lipschitz continuous, implies that $(s_1^u(\tau) - s_1^*) \to 0$ as $\tau \to \infty$.

We now consider some $i \ge 2$ and make the induction hypothesis that

$$\int_0^{\infty} (s_k^u(\tau) - s_k^*)\, d\tau < \infty, \quad \forall\, k \le i-1. \tag{14}$$

Then,

$$-\, v_i^u(0) \le v_i^u(t) - v_i^u(0) = \int_0^t (\lambda(s_{i-1}^u(\tau)P_0(s^u(\tau)) - s_{i-1}^* P_0^*) - (s_i^u(\tau) - s_i^*))\, d\tau. \tag{15}$$

Adding and subtracting $\lambda s_{i-1}^* P_0(s^u(\tau))$ inside the integral, we obtain

$$-v_1^u(0) \le \int_0^t (\lambda[s_{i-1}^u(\tau) - s_{i-1}^*]P_0(s^u(\tau)) + \lambda[P_0(s^u(\tau)) - P_0^*]s_{i-1}^* - (s_i^u(\tau) - s_i^*))\, d\tau. \tag{16}$$

Using Lemma 4.4, we have $s_{i-1}^u(\tau) \ge s_{i-1}^*$ for all $i \ge 1$, and for all $\tau \ge 0$, which also implies that $P_0(s^u(\tau)) \ge P_0^*$ for all $\tau \ge 0$. Therefore, the two terms inside brackets are nonnegative. Using the facts $\lambda < 1$, $s_{i-1}^* \le 1$, and $P_0(s^u(\tau)) \le 1$, Equation (16) implies that

$$-v_i^u(0) \le \int_0^t ([s_{i-1}^u(\tau) - s_{i-1}^*] + [P_0(s^u(\tau)) - P_0^*] - [s_i^u(\tau) - s_i^*])\, d\tau,$$

or

$$\int_0^t (s_i^u(\tau) - s_i^*)\, d\tau \le v_i(0) + \int_0^t (s_{i-1}^u(\tau) - s_{i-1}^*)\, d\tau + \int_0^t (P_0(s^u(\tau)) - P_0^*)\, d\tau. \tag{17}$$

The first integral on the right-hand side of Equation (17) is upper-bounded uniformly in t, by the induction hypothesis (Equation (14)). We now derive an upper bound on the last integral, for each one of the three regimes.

(i) *High Memory regime*: By inspecting the expression for $P_0(s)$ for the High-Memory variant, we observe that it is monotonically nondecreasing and Lipschitz continuous in s_1. Therefore, there exists a constant L such that

$$\int_0^t (P_0(s^u(\tau)) - P_0^*)\,d\tau \le \int_0^t L(s_1^u(\tau) - s_1^*)\,d\tau.$$

Using the induction hypothesis for $k = 1$, we conclude that the last integral on the right-hand side of Equation (17) is upper bounded, uniformly in t.

(ii) *Constrained regime*: For the Constrained regime, the function $P_0(s)$ is again monotonically nondecreasing and, as remarked at the beginning of the proof of Lemma 4.1, it is also Lipschitz continuous in s_1. Thus, the argument is identical to the previous case.

(iii) *High Message regime*: We have an initial condition $s^0 \in \mathcal{S}^1$, and therefore $0 \le v_1^0 < \infty$. As already remarked, we have $\dot{v}_1^u(t) = \lambda - s_1^u(t) \le 0$. It follows that s_1^u can be equal to 1 for at most $v_1^0/(1-\lambda)$ units of time. Therefore, $P_0(s^u(t)) = [1 - (1 - s_2^u(t))/\lambda]^+ \mathbb{1}_{\{s_1^u(t)=1\}}$ can be positive only on a set of times of Lebesgue measure at most $v_1^0/(1-\lambda)$. This implies the uniform (in t) upper bound

$$\int_0^t (P_0(s^u(\tau)) - P_0^*)\,d\tau = \int_0^t P_0(s^u(\tau))\,d\tau \le \frac{v_1^0}{1-\lambda}.$$

For all three cases, we have shown that the last integral in Equation (17) is upper bounded, uniformly in t. It follows from Equation (17) and the induction hypothesis that

$$\int_0^\infty (s_i^u(\tau) - s_i^*)\,d\tau < \infty.$$

This completes the proof of the induction step. Using the Lipschitz-continuity of $s_i^u(\cdot)$, it follows that $s_i^u(t)$ converges to s_i^* for all $i \ge 1$. It is straightforward to check that this coordinate-wise convergence, together with boundedness ($s_i^u(t) \le 1$, for all i and t), implies that also

$$\lim_{t\to\infty} \|s^u(t) - s^*\|_w = 0.$$

An analogous argument gives us the convergence

$$\lim_{t\to\infty} \|s^l(t) - s^*\|_w = 0,$$

which concludes the proof. □

5. Stochastic Transient Analysis—Proof of Theorem 3.2 and of the Rest of Theorem 3.1

We will now prove the convergence of the stochastic system to the fluid solution. The proof involves three steps. We first define the process using a coupled sample path approach, as in Tsitsiklis and Xu (2012). We then show the existence of limiting trajectories under the fluid scaling (Proposition 5.3). We finally show that any such limit trajectory must satisfy the differential equations in the definition of the fluid model (Proposition 5.4).

5.1. Probability Space and Coupling

We will first define a common probability space for all n. We will then define a coupled sequence of processes $\{(S^n(t), M^n(t))\}_{n=1}^\infty$. This approach will allow us to obtain almost sure convergence in the common probability space.

5.1.1. Fundamental Processes and Initial Conditions. All processes of interest (for all n) will be driven by certain common fundamental processes.

(a) Driving Poisson processes: Independent Poisson counting processes $\mathcal{N}_\lambda(t)$ (process of arrivals, with rate λ), and $\mathcal{N}_1(t)$ (process of potential departures, with rate 1). A coupled sequence $\{\mathcal{N}_{\mu(n)}(t)\}_{n=1}^\infty$ (processes of potential messages, with nondecreasing rates $\mu(n)$), independent of $\mathcal{N}_\lambda(t)$ and $\mathcal{N}_1(t)$, such that the events in $\mathcal{N}_{\mu(n)}(t)$ are a subset of the events in $\mathcal{N}_{\mu(n+1)}(t)$ almost surely, for all $n \ge 1$. These processes are defined on a common probability space $(\Omega_D, \mathcal{A}_D, \mathbb{P}_D)$.

(b) Selection processes: Three independent discrete time processes $U(k)$, $V(k)$, and $W(k)$, which are all i.i.d. and uniform on $[0,1]$, defined on a common probability space $(\Omega_S, \mathcal{A}_S, \mathbb{P}_S)$.

(c) Initial conditions: A sequence of random variables $\{(S^{(0,n)}, M^{(0,n)})\}_{n=1}^{\infty}$ defined on a common probability space $(\Omega_0, \mathscr{A}_0, \mathbb{P}_0)$ and taking values in $(\mathscr{S}^1 \cap \mathscr{I}_n) \times \{0, 1, \ldots, c(n)\}$.
The whole system will be defined on the probability space

$$(\Omega, \mathscr{A}, \mathbb{P}) = (\Omega_D \times \Omega_S \times \Omega_0, \mathscr{A}_D \times \mathscr{A}_S \times \mathscr{A}_0, \mathbb{P}_D \times \mathbb{P}_S \times \mathbb{P}_0).$$

All of the randomness in the system (for any n) will be specified by these fundamental processes, and everything else will be a deterministic function of them.

5.1.2. A Coupled Construction of Sample Paths. Recall that our policy results in a Markov process $(S^n(t), M^n(t)) \in (\mathscr{S}^1 \cap \mathscr{I}_n) \times \{0, 1, \ldots, c(n)\}$, where $S_i^n(t)$ is the fraction of servers with at least i jobs and $M^n(t)$ is the number of tokens stored in memory, at time t. We now describe a particular construction of the process, as a deterministic function of the fundamental processes. We decompose the process $S^n(t)$ as the sum of two non-negative and non-decreasing processes, $A^n(t)$ and $D^n(t)$, that represent the (scaled by n) total cumulative arrivals to and departures from the queues, respectively, so that

$$S^n(t) = S^{(0,n)} + A^n(t) - D^n(t).$$

Let $t_j^{\lambda,n}$, $t_j^{1,n}$, and $t_j^{\mu,n}$ be the time of the j-th arrival of $\mathscr{N}_\lambda(nt)$, $\mathscr{N}_1(nt)$, and $\mathscr{N}_{\mu(n)}(nt)$, respectively. In order to simplify notation, we will omit the superscripts λ, 1, and μ, when the corresponding process is clear. We denote by $S^n(t^-)$ the left limit $\lim_{s \uparrow t} S^n(s)$, and similarly for $M^n(t^-)$. Then, the first component of $A^n(t)$ is

$$A_1^n(t) = \frac{1}{n} \sum_{j=1}^{\mathscr{N}_\lambda(nt)} [\mathbb{1}_{[1,c(n)]}(M^n(t_j^{n-})) + \mathbb{1}_{\{0\}}(M^n(t_j^{n-}))\mathbb{1}_{[0,1-S_1^n(t_j^{n-}))}(U(j))]. \tag{18}$$

The above expression is interpreted as follows. We have an upward jump of size $1/n$ in A_1^n every time that a job joins an empty queue, which happens every time that there is an arrival and either (i) there are tokens in the virtual queue (i.e., $M^n > 0$) or, (ii) there are no tokens and an empty queue is drawn uniformly at random, which happens with probability $1 - S_1^n$. Similarly, for $i \geq 2$,

$$A_i^n(t) = \frac{1}{n} \sum_{j=1}^{\mathscr{N}_\lambda(nt)} \mathbb{1}_{\{0\}}(M^n(t_j^{n-}))\mathbb{1}_{[1-S_{i-1}^n(t_j^{n-}),1-S_i^n(t_j^{n-}))}(U(j)).$$

In this case we have an upward jump in A_i^n of size $1/n$ every time that there is an arrival, there are no tokens in the virtual queue (i.e., $M^n = 0$), and a queue with exactly $i - 1$ jobs is drawn uniformly at random, which happens with probability $S_{i-1}^n - S_i^n$. Moreover, for all $i \geq 1$,

$$D_i^n(t) = \frac{1}{n} \sum_{j=1}^{\mathscr{N}_1(nt)} \mathbb{1}_{[1-S_i^n(t_j^{n-}),1-S_{i+1}^n(t_j^{n-}))}(W(j)).$$

We have an upward jump in D_i^n of size $1/n$ when there is a departure from a queue with exactly i jobs, which happens with rate $(S_i^n - S_{i+1}^n)n$.

Recall that $\mu(n)$ is the message rate of an empty server. In the High Memory and Constrained regimes, we have $\mu(n) = \mu$, while in the High Message regime $\mu(n)$ is a nondecreasing and unbounded sequence. Potential messages are generated according to the process $\mathscr{N}_{\mu(n)}(nt)$, but an actual message is generated only if a randomly selected queue is empty. Thus, the number of tokens in the virtual queue evolves as follows:

$$M^n(t) = M^{(0,n)} - \sum_{j=1}^{\mathscr{N}_\lambda(nt)} \mathbb{1}_{[1,c(n)]}(M^n(t_j^{n-})) + \sum_{j=1}^{\mathscr{N}_{\mu(n)}(nt)} \mathbb{1}_{[0,c(n)-1]}(M^n(t_j^{n-}))\mathbb{1}_{[0,1-S_1(t_j^{n-})-M^n(t_j^{n-})/n)}(V(j)). \tag{19}$$

To see this, if the virtual queue is not empty, a token is removed from the virtual queue each time there is an arrival. Furthermore, if the virtual queue is not full, a new token is added each time a new message arrives from one of the $n(1 - S_1^n) - M^n$ queues that do not have corresponding tokens in the virtual queue.

Remark 5.1. The desired result only concerns the convergence of the projection of the Markov process $(S^n(t), M^n(t))$ onto its first component. However, the process of tokens M^n will still have an impact on that limit.

As mentioned earlier, the proof involves the following two steps:

1. We show that there exists a measurable set $\mathscr{C} \subset \Omega$ with $\mathbb{P}(\mathscr{C}) = 1$ such that for all $\omega \in \mathscr{C}$, any sequence of sample paths $S^n(\omega, t)$ contains a further subsequence that converges to a Lipschitz continuous trajectory $s(t)$, as $n \to \infty$.

2. We characterize the derivative of $s(t)$ at any regular point and show that it is identical to the drift of our fluid model. Hence $s(t)$ must be a fluid solution for some initial condition s^0, yielding also, as a corollary, the existence of fluid solutions.

5.2. Tightness of Sample Paths

We start by finding a set of "nice" sample paths ω for which any subsequence of the sequence $\{S^n(\omega, t)\}_{n=1}^{\infty}$ contains a further subsequence $\{S^{n_k}(\omega, t)\}_{k=1}^{\infty}$ that converges to some Lipschitz continuous function s. The arguments involved here are fairly straightforward and routine.

Lemma 5.1. *Fix $T > 0$. There exists a measurable set $\mathscr{C} \subset \Omega$ such that $\mathbb{P}(\mathscr{C}) = 1$ and for all $\omega \in \mathscr{C}$,*

$$\lim_{n\to\infty} \sup_{t\in[0,T]} \left| \frac{1}{n}\mathscr{N}_\lambda(\omega, nt) - \lambda t \right| = 0, \tag{20}$$

$$\lim_{n\to\infty} \sup_{t\in[0,T]} \left| \frac{1}{n}\mathscr{N}_1(\omega, nt) - t \right| = 0, \tag{21}$$

$$\lim_{n\to\infty} \frac{1}{n} \sum_{i=1}^{n} \mathbb{1}_{[a,b)}(U(\omega, i)) = b - a, \quad \text{for all } [a,b) \subset [0,1], \tag{22}$$

$$\lim_{n\to\infty} \frac{1}{n} \sum_{i=1}^{n} \mathbb{1}_{[c,d)}(W(\omega, i)) = d - c, \quad \text{for all } [c,d) \subset [0,1]. \tag{23}$$

Proof. Using the Functional Strong Law of Large Numbers for Poisson processes, we obtain a subset $\mathscr{C}_D \subset \Omega_D$ such that $\mathbb{P}_D(\mathscr{C}_D) = 1$ on which Equations (20) and (21) hold. Furthermore, the Glivenko-Cantelli lemma gives us another subset $\mathscr{C}_S \subset \Omega_S$ such that $\mathbb{P}_S(\mathscr{C}_S) = 1$ and on which Equations (22) and (23) hold. Taking $\mathscr{C} = \mathscr{C}_D \times \mathscr{C}_S \times \Omega_0$ concludes the proof. □

Let us fix an arbitrary $s^0 \in [0,1]$, sequences $R_n \downarrow 0$ and $\gamma_n \downarrow 0$, and a constant $L > 0$. For $n \geq 1$, we define the following subsets of $D[0,T]$:

$$E_n(R_n, \gamma_n) \triangleq \{s \in D[0,T]\colon |s(0) - s^0| \leq R_n, \text{ and } |s(a) - s(b)| \leq L|a-b| + \gamma_n, \forall\, a, b \in [0,T]\}. \tag{24}$$

We also define

$$E_c \triangleq \{s \in D[0,T]\colon s(0) = s^0, |s(a) - s(b)| \leq L|a-b|, \forall\, a, b \in [0,T]\},$$

which is the set of L-Lipschitz continuous functions with fixed initial conditions, and which is known to be sequentially compact, by the Arzelà-Ascoli theorem.

Lemma 5.2. *Fix $T > 0$, $\omega \in \mathscr{C}$, and some $s^0 \in \mathscr{S}^1$. Suppose that*

$$\|S^n(\omega, 0) - s^0\|_w \leq \tilde{R}_n,$$

for some sequence $\tilde{R}_n \downarrow 0$. Then, there exist sequences $\{R_n^{(i)} \downarrow 0\}_{i=0}^{\infty}$ and $\gamma_n \downarrow 0$ such that

$$S_i^n(\omega, \cdot) \in E_n(R_n^{(i)}, \gamma_n), \quad \forall\, i \in \mathbb{Z}_+, \forall\; n \geq 1,$$

with the constant L in the definition of E_n equal to $1 + \lambda$.

Proof. Fix some $\omega \in \mathscr{C}$. Based on our coupled construction, each coordinate of A^n (the process of cumulative arrivals) and D^n (the process of cumulative departures) is non-decreasing, and can have a positive jump, of size $1/n$, only when there is an event in $\mathscr{N}_\lambda$ or $\mathscr{N}_1$, respectively. As a result, for every i and n, we have

$$|A_i^n(\omega, a) - A_i^n(\omega, b)| \leq \frac{1}{n}|\mathscr{N}_\lambda(\omega, na) - \mathscr{N}_\lambda(\omega, nb)|, \quad \forall\, a, b \in [0,T],$$

and

$$|D_i^n(\omega, a) - D_i^n(\omega, b)| \leq \frac{1}{n}|\mathscr{N}_1(\omega, na) - \mathscr{N}_1(\omega, nb)|, \quad \forall\, a, b \in [0,T].$$

Therefore,

$$|S_i^n(\omega,a)-S_i^n(\omega,b)| \le \frac{1}{n}|\mathcal{N}_\lambda(\omega,na)-\mathcal{N}_\lambda(\omega,nb)| + \frac{1}{n}|\mathcal{N}_1(\omega,na)-\mathcal{N}_1(\omega,nb)|.$$

Since $\omega\in\mathscr{C}$, Lemma 5.1 implies that $(1/n)\mathcal{N}_\lambda(\omega,nt)$ and $(1/n)\mathcal{N}_1(\omega,nt)$ converge uniformly on $[0,T]$ to λt and to t, respectively. Thus, there exists a pair of sequences $\gamma_n^1\downarrow 0$ and $\gamma_n^2\downarrow 0$ (which depend on ω) such that for all $n\ge 1$,

$$\frac{1}{n}|\mathcal{N}_\lambda(\omega,na)-\mathcal{N}_\lambda(\omega,nb)| \le \lambda|a-b|+\gamma_n^1,$$

and

$$\frac{1}{n}|\mathcal{N}_1(\omega,na)-\mathcal{N}_1(\omega,nb)| \le |a-b|+\gamma_n^2,$$

which imply that

$$|S_i^n(\omega,a)-S_i^n(\omega,b)| \le (1+\lambda)|a-b|+(\gamma_n^1+\gamma_n^2).$$

The proof is completed by setting $R_n^{(i)}=2^i\tilde{R}_n$, $\gamma_n=\gamma_n^1+\gamma_n^2$, and $L=1+\lambda$. □

We are now ready to prove the existence of convergent subsequences of the process of interest.

Proposition 5.3. *Fix $T>0$, $\omega\in\mathscr{C}$, and some $s^0\in\mathscr{S}^1$. Suppose (as in Lemma 5.2) that $\|S^n(\omega,0)-s^0\|_w\le\tilde{R}_n$, where $\tilde{R}_n\downarrow 0$. Then, every subsequence of $\{S^n(\omega,\cdot)\}_{n=1}^\infty$ contains a further subsequence $\{S^{n_k}(\omega,\cdot)\}_{k=1}^\infty$ that converges to a coordinate-wise Lipschitz continuous function $s(t)$ with $s(0)=s^0$ and*

$$|s_i(a)-s_i(b)|\le L|a-b|,\quad \forall\, a,b\in[0,T],\ \ i\in\mathbb{Z},$$

where L is independent of T, ω, and $s(\cdot)$.

Proof. As in Lemma 5.2, let $L=1+\lambda$. A standard argument, similar to the one in Bramson (1998) and Tsitsiklis and Xu (2012), based on the sequential compactness of E_c and the "closeness" of $E_n(R_n^{(i)},\gamma_n)$ to E_c establishes the following. For any $i\ge 1$, every subsequence of $\{S_i^n(\omega,\cdot)\}_{n=1}^\infty$ contains a further subsequence that converges to a Lipschitz continuous function $y_i(t)$ with $y_i(0)=s_i^0$.

Starting with the existence of coordinate-wise limit points, we now argue the existence of a limit point of S^n in $D^\infty[0,T]$. Let s_1 be a Lipschitz continuous limit point of $\{S_1^n(\omega,\cdot)\}_{n=1}^\infty$, so that there is a subsequence such that

$$\lim_{k\to\infty} d(S_1^{n_k^1}(\omega,\cdot),s_1)=0.$$

We then proceed inductively and let s_{i+1} be a limit point of a subsequence of $\{S_{i+1}^{n_k^i}(\omega,\cdot)\}_{k=1}^\infty$, where $\{n_k^i\}_{k=1}^\infty$ are the indices of the subsequence of S_i^n.

We now argue that s is indeed a limit point of S^n in $D^\infty[0,T]$. Fix a positive integer i. Because of the construction of s, $S_j^{n_k^i}(\omega,\cdot)$ converges to s_j, as $k\to\infty$, for $j=1,\dots,i$. In particular, there exists some $n^i>i$, for which

$$d(S_j^{n^i}(\omega,\cdot),s_j)\le\frac{1}{i},\quad j=1,\dots,i.$$

We then have

$$d^{\mathbb{Z}_+}(S^{n^i}(\omega,\cdot),s)=\sup_{t\in[0,T]}\sqrt{\sum_{j=1}^\infty 2^{-j}|S_j^{n^i}(\omega,t)-s_j(t)|^2}\le\frac{1}{i}+\sqrt{\sum_{j=n^i+1}^\infty 2^{-j+2}}.$$

We now let i increase to infinity (in which case n^i also increases to infinity), and we conclude that $d^{\mathbb{Z}_+}(S^{n^i}(\omega,\cdot),s)\to 0$. □

This concludes the proof of the tightness of the sample paths. It remains to prove that any possible limit point is a fluid solution.

5.3. Derivatives of the Fluid Limits

Proposition 5.4. *Fix $\omega \in \mathscr{C}$ and $T > 0$. Let s be a limit point of some subsequence of $\{S^n(\omega, \cdot)\}_{n=1}^{\infty}$. As long as ω does not belong to a certain zero-measure subset of $\mathscr{C}$, s satisfies the differential equations that define a fluid solution (cf. Definition 3.1).*

Proof. We fix some $\omega \in \mathscr{C}$ and for the rest of this proof we suppress the dependence on ω in our notation. Let $\{S^{n_k}\}_{k=1}^{\infty}$ be a subsequence that converges to s, i.e.,

$$\lim_{k\to\infty} \sup_{0\le t\le T} \|S^{n_k}(t) - s(t)\|_w = 0.$$

After possibly restricting, if necessary, to a further subsequence, we can define Lipschitz continuous functions $a_i(t)$ and $d_i(t)$ as the limits of the subsequences of cumulative arrivals and departures processes $\{A_i^{n_k}(t)\}_{k=1}^{\infty}$ and $\{D_i^{n_k}(t)\}_{k=1}^{\infty}$ respectively. Because of the relation $S_i^n(t) = S^{(0,n)} + A_i^n(t) - D_i^n(t)$, it is enough to prove the following relations, for almost all t:

$$\begin{aligned}
\frac{da_1}{dt}(t) &= \lambda[1 - P_0(s(t))] + \lambda[1 - s_1(t)]P_0(s(t)),\\
\frac{da_i}{dt}(t) &= \lambda[s_{i-1}(t) - s_i(t)]P_0(s(t)), \quad \forall\, i \ge 2,\\
\frac{dd_i}{dt}(t) &= s_i(t) - s_{i+1}(t), \quad \forall\, i \ge 1.
\end{aligned}$$

We will provide a proof only for the first one, as the other proofs are similar. The main idea in the argument that follows is to replace the token process M^n by simpler, time-homogeneous birth-death processes that are easy to analyze.

Let us fix some time $t \in (0, T)$, which is a regular time for both a_1 and d_1. Let $\epsilon > 0$ be small enough so that $t + \epsilon \le T$ and so that it also satisfies a condition to be introduced later. Equation (18) yields

$$A_1^{n_k}(t+\epsilon) - A_1^{n_k}(t) = \frac{1}{n_k} \sum_{j=\mathscr{N}_\lambda(n_k t)+1}^{\mathscr{N}_\lambda(n_k(t+\epsilon))} [\mathbb{1}_{[1,c(n_k)]}(M^{n_k}(t_j^{n_k-})) + \mathbb{1}_{\{0\}}(M^{n_k}(t_j^{n_k-}))\mathbb{1}_{[0,1-S_1^{n_k}(t_j^{n_k-}))}(U(j))]. \tag{25}$$

By Lemma 5.2, there exists a sequence $\gamma_{n_k} \downarrow 0$ and a constant L such that

$$S_1^{n_k}(u) \in [s_1(t) - (\epsilon L + \gamma_{n_k}), s_1(t) + (\epsilon L + \gamma_{n_k})), \quad \forall\, u \in [t, t+\epsilon].$$

Then, for all sufficiently large k, we have

$$S_1^{n_k}(u) \in [s_1(t) - 2\epsilon L, s_1(t) + 2\epsilon L)), \quad \forall\, u \in [t, t+\epsilon]. \tag{26}$$

In particular, for k sufficiently large and for every event time $t_j^{n_k-} \in (t, t+\epsilon]$ of the driving process $\mathscr{N}_\lambda$, we have

$$[0, 1 - S_1^{n_k}(t_j^{n_k-})) \subset [0, 1 - s_1(t) + 2\epsilon L).$$

This implies that

$$A_1^{n_k}(t+\epsilon) - A_1^{n_k}(t) \le \frac{1}{n_k} \sum_{j=\mathscr{N}_\lambda(n_k t)+1}^{\mathscr{N}_\lambda(n_k(t+\epsilon))} [\mathbb{1}_{[1,c(n_k)]}(M^{n_k}(t_j^{n_k-})) + \mathbb{1}_{\{0\}}(M^{n_k}(t_j^{n_k-}))\mathbb{1}_{[0,1-s_1(t)+2\epsilon L)}(U(j))].$$

We wish to analyze this upper bound on $A_1^{n_k}(t+\epsilon) - A_1^{n_k}(t)$, which will then lead to an upper bound on $(da_i/dt)(t)$. Towards this purpose, we will focus on the empirical distribution of $\mathbb{1}_{\{0\}}(M^{n_k}(t_j^{n_k-}))$, which depends on the birth-death process $M^{n_k}(t)$, and which is in turn modulated by $S^{n_k}(t)$. In particular, we will define two coupled time-homogeneous birth-death processes: $M_+^{n_k}$, which is dominated by M^{n_k}; and $M_-^{n_k}$, which dominates M^{n_k} over $(t, t+\epsilon]$, i.e.,

$$M_+^{n_k}(u) \le M^{n_k}(u) \le M_-^{n_k}(u), \quad \forall\, u \in (t, t+\epsilon]. \tag{27}$$

This is accomplished as follows. Using again Equation (26), when n_k is sufficiently large, we get the set inclusion

$$\left[0, 1 - S_1^{n_k}(t_j^{n_k-}) - \frac{M^{n_k}(t_j^{n_k-})}{n_k}\right) \subset [0, 1 - s_1(t) + 2\epsilon L),$$

for all event times $t_j^{n_k} \in [t, t+\epsilon)$. Furthermore, our assumptions on $c(n_k)$ imply that $M^{n_k}(t)/n_k \le c(n_k)/n_k$ goes to zero as $k \to \infty$. Thus, when n_k is sufficiently large,

$$\left[0, 1 - S_1^{n_k}(t_j^{n_k-}) - \frac{M^{n_k}(t_j^{n_k-})}{n_k}\right) \supset [0, 1 - s_1(t) - 3\epsilon L),$$

for all event times $t_j^{n_k} \in [t, t+\epsilon)$. We now define intermediate coupled processes $\tilde{M}_+^{n_k}$ and $\tilde{M}_-^{n_k}$ by replacing the last indicator set in the evolution equation for $M^n(t)$ (cf. Equation (19)), by the deterministic sets introduced above. Furthermore, we set $\tilde{M}_+^{n_k}(t) = 0 \le M^{n_k}(t)$ and $\tilde{M}_-^{n_k}(t) = c(n_k) \ge M^{n_k}(t)$.

More concretely, for all $u \in [t, t+\epsilon]$, we let

$$\tilde{M}_-^{n_k}(u) \triangleq c(n_k) - \sum_{j=\mathcal{N}_\lambda(n_k t)+1}^{\mathcal{N}_\lambda(n_k u)} \mathbb{1}_{[1, c(n_k)]}(\tilde{M}_-^{n_k}(t_j^{n_k-})) + \sum_{j=\mathcal{N}_{\mu(n_k)}(n_k t)+1}^{\mathcal{N}_{\mu(n_k)}(n_k u)} \mathbb{1}_{[0, c(n_k)-1]}(\tilde{M}_-^{n_k}(t_j^{n_k-}))\mathbb{1}_{[0, 1-s_1(t)+2\epsilon L)}(V(j))$$

and

$$\tilde{M}_+^{n_k}(u) \triangleq 0 - \sum_{j=\mathcal{N}_\lambda(n_k t)+1}^{\mathcal{N}_\lambda(n_k u)} \mathbb{1}_{[1, c(n_k)]}(\tilde{M}_+^{n_k}(t_j^{n_k-})) + \sum_{j=\mathcal{N}_{\mu(n_k)}(n_k t)+1}^{\mathcal{N}_{\mu(n_k)}(n_k u)} \mathbb{1}_{[0, c(n_k)-1]}(\tilde{M}_+^{n_k}(t_j^{n_k-}))\mathbb{1}_{[0, 1-s_1(t)-3\epsilon L)}(V(j)).$$

We note that the processes $\tilde{M}_-^{n_k}(u)$ and $\tilde{M}_+^{n_k}(u)$ are plain, time-homogenous birth-death Markov processes, no longer modulated by $S^{n_k}(t)$, and therefore easy to analyze. It can now be argued, by induction on the event times, that $\tilde{M}_-^{n_k}(u) \ge M^{n_k}(u)$ for all u. We omit the details but simply note that (i) this inequality holds at time t; (ii) whenever the process $M^{n_k}(u)$ has an upward jump, the same is true for $\tilde{M}_-^{n_k}(u)$, unless $\tilde{M}_-^{n_k}(u)$ is already at its largest possible value, $c(n_k)$, in which case the desired inequality is preserved; (iii) as long as the desired inequality holds, whenever the process $\tilde{M}_-^{n_k}(u)$ has a downward jump, the same is true for $M^{n_k}(u)$, unless $M^{n_k}(u)$ is already at its smallest possible value, 0, in which case the desired inequality is again preserved. Using also a symmetrical argument for $\tilde{M}_+^{n_k}(u)$, we obtain the domination relationship

$$\tilde{M}_+^{n_k}(u) \le M^{n_k}(u) \le \tilde{M}_-^{n_k}(u), \quad \forall\, u \in (t, t+\epsilon]. \tag{28}$$

Even though $\tilde{M}_+^{n_k}$ and $\tilde{M}_-^{n_k}$ are simple birth-death processes, it is convenient to simplify them even further. We thus proceed to define the coupled processes $M_+^{n_k}$ and $M_-^{n_k}$ by modifying the intermediate processes $\tilde{M}_+^{n_k}$ and $\tilde{M}_-^{n_k}$ in a different way for each regime.

(i) *High Memory regime*: Recall that in this regime we have $\mu(n_k) = \mu$ for all k. Let us fix some l, independently from k, and let $c_l = c(n_l)$. For every k, we define $M_+^{n_k}$ and $M_-^{n_k}$ by replacing the upper bound $c(n_k)$ on the number of tokens in $\tilde{M}_+^{n_k}$ and $\tilde{M}_-^{n_k}$, by c_l and ∞ respectively. More concretely, for $u \in [t, t+\epsilon]$ we let

$$M_-^{n_k}(u) \triangleq c(n_k) - \sum_{j=\mathcal{N}_\lambda(n_k t)+1}^{\mathcal{N}_\lambda(n_k u)} \mathbb{1}_{[1, \infty)}(M_-^{n_k}(t_j^{n_k-})) + \sum_{j=\mathcal{N}_\mu(n_k t)+1}^{\mathcal{N}_\mu(n_k u)} \mathbb{1}_{[0, 1-s_1(t)+2\epsilon L)}(V(j))$$

and

$$M_+^{n_k}(u) \triangleq 0 - \sum_{j=\mathcal{N}_\lambda(n_k t)+1}^{\mathcal{N}_\lambda(n_k u)} \mathbb{1}_{[1, c_l]}(M_+^{n_k}(t_j^{n_k-})) + \sum_{j=\mathcal{N}_\mu(n_k t)+1}^{\mathcal{N}_\mu(n_k u)} \mathbb{1}_{[0, c_l-1]}(M_+^{n_k}(t_j^{n_k-}))\mathbb{1}_{[0, 1-s_1(t)-3\epsilon L)}(V(j)).$$

When k is large enough, we have $c(n_k) \ge c_l$, and as we are replacing $c(n_k)$ by c_l in $\tilde{M}_+^{n_k}$, we are reducing the state space of the homogeneous birth-death process $\tilde{M}_+^{n_k}$. It is easily checked (by induction on the events of the processes) that we have the stochastic dominance $\tilde{M}_+^{n_k} \ge M_+^{n_k}$. Using a similar argument, we obtain $\tilde{M}_-^{n_k} \le M_-^{n_k}$. These facts, together with Equation (28), imply the desired dominance relation in Equation (27).

(ii) *High Message regime*: Recall that in this regime we have $c(n_k) = c$, for all k. Let us fix some l, independently from k, and let $\mu_l = \mu(n_l)$. We define $M_+^{n_k}$ by replacing the process $\mathcal{N}_{\mu(n_k)}$ that generates the spontaneous messages in $\tilde{M}_+^{n_k}$, by $\mathcal{N}_{\mu_l}$. More concretely, for $u \in [t, t+\epsilon]$ we let

$$M_+^{n_k}(u) \triangleq 0 - \sum_{j=\mathcal{N}_\lambda(n_k t)+1}^{\mathcal{N}_\lambda(n_k u)} \mathbb{1}_{[1, c]}(M_+^{n_k}(t_j^{n_k-})) + \sum_{j=\mathcal{N}_{\mu_l}(n_k t)+1}^{\mathcal{N}_{\mu_l}(n_k u)} \mathbb{1}_{[0, c-1]}(M_+^{n_k}(t_j^{n_k-}))\mathbb{1}_{[0, 1-s_1(t)-3\epsilon L)}(V(j)).$$

Recall that we assumed that the event times in the Poisson process $\mathcal{N}_{\mu(n_k)}$ are a subset of the event times of $\mathcal{N}_{\mu(n_{k+1})}$, for all k. As a result, when $k \ge l$, the process $M_+^{n_k}$ only has a subset of the upward jumps in $\tilde{M}_+^{n_k}$, and thus (using again a simple inductive argument) satisfies $\tilde{M}_+^{n_k} \ge M_+^{n_k}$. Furthermore, we define $M_-^{n_k}(u) \triangleq c$, which clearly satisfies $\tilde{M}_-^{n_k} \le M_-^{n_k}$. Combining these facts with Equation (28), we have again the desired dominance relation in Equation (27).

(iii) *Constrained regime*: Recall that in this regime we have $c(n_k) = c$ and $\mu(n_k) = \mu$, for all $k \ge 1$. For this case, we define $M_-^{n_k} = \tilde{M}_-^{n_k}$ and $M_+^{n_k} = \tilde{M}_+^{n_k}$, which already satisfy the desired dominance relation in Equation (27).

For all three regimes, and having fixed l, the dominance relation in Equation (27) implies that when k is large enough ($k \ge l$), we have

$$\mathbb{1}_{\{0\}}(M_-^{n_k}(t_j^{n_k-})) \le \mathbb{1}_{\{0\}}(M^{n_k}(t_j^{n_k-})) \le \mathbb{1}_{\{0\}}(M_+^{n_k}(t_j^{n_k-}))$$

for all $t_j^{n_k-} \in (t, t+\epsilon]$. Consequently,

$$A_1^{n_k}(t+\epsilon) - A_1^{n_k}(t) \le \frac{1}{n_k} \sum_{j=\mathcal{N}_\lambda(n_k t)+1}^{\mathcal{N}_\lambda(n_k(t+\epsilon))} [1 - \mathbb{1}_{\{0\}}(M_-^{n_k}(t_j^{n_k-})) + \mathbb{1}_{\{0\}}(M_+^{n_k}(t_j^{n_k-}))\mathbb{1}_{[0,1-s_1(t)+2\epsilon L)}(U(j))]. \tag{29}$$

Note that the transition rates of the birth-death processes $M_-^{n_k}$ and $M_+^{n_k}$, for different n_k, involve n_k only as a scaling factor. As a consequence, the corresponding steady-state distributions are the same for all n_k.

Let $P_0^-(s(t))$ and $P_0^+(s(t))$ be the steady-state probabilities of state 0 for $M_-^{n_k}$ and $M_+^{n_k}$, respectively. Then, using the PASTA property, we have that as $n_k \to \infty$, the empirical averages

$$\frac{1}{n_k} \sum_{j=\mathcal{N}_\lambda(n_k t)+1}^{\mathcal{N}_\lambda(n_k(t+\epsilon))} \mathbb{1}_{\{0\}}(M_-^{n_k}(t_j^{n_k-}))$$

and

$$\frac{1}{n_k} \sum_{j=\mathcal{N}_\lambda(n_k t)+1}^{\mathcal{N}_\lambda(n_k(t+\epsilon))} \mathbb{1}_{\{0\}}(M_+^{n_k}(t_j^{n_k-})) \tag{30}$$

converge almost surely to $\epsilon\lambda P_0^-(s(t))$ and $\epsilon\lambda P_0^+(s(t))$, respectively.

We now continue with the explicit calculation of $P_0^-(s(t))$ and $P_0^+(s(t))$.

(i) *High Memory regime*:

$$P_0^-(s(t)) = \left[1 - \frac{\mu \cdot \min\{1 - s_1(t) + 2\epsilon L, 1\}}{\lambda}\right]^+ \qquad \text{and} \qquad P_0^+(s(t)) = \left[\sum_{k=0}^{c_l} \left(\frac{\mu(1 - s_1(t) - 3\epsilon L)^+}{\lambda}\right)^k\right]^{-1},$$

(ii) *High Message regime*: If $s_1(t) < 1$, then we assume that ϵ has been chosen small enough so that $1 - s_1(t) - 3\epsilon L > 0$. We then obtain

$$P_0^-(s(t)) = 0 \qquad \text{and} \qquad P_0^+(s(t)) = \left[\sum_{k=0}^{c} \left(\frac{\mu_l[1 - s_1(t) - 3\epsilon L]^+}{\lambda}\right)^k\right]^{-1}.$$

Suppose now that $s_1(t) = 1$. In this case, the approach based on the processes $M_-^{n_k}$ and $M_+^{n_k}$ is not useful, because it yields $P_0^-(s(t)) = 0$ and $P_0^+(s(t)) = 1$, for all $\epsilon > 0$ and for all μ_l. This case will be considered separately later.

(iii) *Constrained regime*:

$$P_0^-(s(t)) = \left[\sum_{k=0}^{c} \left(\frac{\mu \cdot \min\{1 - s_1(t) + 2\epsilon L, 1\}}{\lambda}\right)^k\right]^{-1} \qquad \text{and} \qquad P_0^+(s(t)) = \left[\sum_{k=0}^{c} \left(\frac{\mu(1 - s_1(t) - 3\epsilon L)^+}{\lambda}\right)^k\right]^{-1},$$

We now continue by considering all three regimes, with the exception of the High Message regime with $s_1(t) = 1$, which will be dealt with separately. We use the fact that the random variables $U(j)$ are independent from the process $M_+^{n_k}$. Using an elementary argument, which is omitted, it can be seen that

$$\frac{1}{n_k} \sum_{j=\mathcal{N}_\lambda(n_k t)+1}^{\mathcal{N}_\lambda(n_k(t+\epsilon))} \mathbb{1}_{\{0\}}(M_+^{n_k}(t_j^{n_k-}))\mathbb{1}_{[0,1-s_1(t)+2\epsilon L)}(U(j))$$

converges to the limit of the empirical average in Equation (30), which is the product of $\epsilon\lambda P_0^+(s(t))$ times the expected value of $\mathbb{1}_{[0,1-s_1(t)+2\epsilon L)}(U(j))$. That is, it converges to $\epsilon P_0^+(s(t))\min\{1-s_1(t)+2\epsilon L, 1\}$, $\mathbb{P}$-almost surely.

Recall that we have fixed some $\epsilon > 0$ and some l and, furthermore, that P_0^- and P_0^+ depend on l for the High Memory and High Message regimes, and on ϵ for all regimes. We will first take limits, as $k \to \infty$, while holding ϵ and l fixed. Using the inequality in Equation (29), and the fact that the left-hand side converges to the fluid limit $a(t+\epsilon) - a(t)$ as $k \to \infty$, we obtain

$$a_1(t+\epsilon) - a_1(t) \le \epsilon\lambda[1 - P_0^-(s(t))] + \epsilon\lambda P_0^+(s(t))\min\{1 - s_1(t) + 2\epsilon L, 1\}.$$

An analogous argument yields

$$a_1(t+\epsilon) - a_1(t) \ge \epsilon\lambda[1 - P_0^+(s(t))] + \epsilon\lambda P_0^-(s(t))[1 - s_1(t) - 2\epsilon L]^+.$$

We now take the limit as $l \to \infty$, so that $c_l \to \infty$ for the High Memory regime and $\mu_l \to \infty$ for the High Message regime, and then take the limit as $\epsilon \to 0$. Some elementary algebra shows that in all cases, $P_0^+(s(t))$ and $P_0^-(s(t))$ both converge to $P_0(s(t))$, as defined in the statement of the proposition. We thus obtain

$$\frac{da_1(t)}{dt} = \lambda[1 - P_0(s(t))] + \lambda[1 - s_1(t)]P_0(s(t)), \tag{31}$$

as desired.

We now return to the exceptional case of the High Message regime with $s_1(t) = 1$, and find the derivative of $a_1(t)$ using a different argument. Recall that we have the hard bound $S_1^n(t) \le 1$, for all t and for all n. This leads to the same bound for the fluid solutions, i.e., $s_1(t) \le 1$ for all t. As a result, since $t > 0$ is a regular time, we must have $\dot{s}_1(t) = 0$. Furthermore, we also have the formula

$$\dot{d}_1(t) = s_1(t) - s_2(t) = 1 - s_2(t),$$

which is established by an independent argument, using the same proof technique as for $\dot{a}_1$, but without the inconvenience of having to deal with M^{n_k}. Then, since $\dot{s}_1(t) = \dot{a}_1(t) - \dot{d}_1(t)$, we must also have

$$\dot{a}_1(t) = 1 - s_2(t). \tag{32}$$

On the other hand, it can be easily checked that $\dot{a}_1(t) \le \lambda$ for all regular t, and thus we must have $s_2(t) \ge 1 - \lambda$. We have thus established that at all regular times $t > 0$ with $s_1(t) = 1$, $s_2(t)$ must be at least $1 - \lambda$. Then it follows (cf. Definition 3.1) that at time t, we have

$$P_0(s(t)) = \left[1 - \frac{1 - s_2(t)}{\lambda}\right]^+ \mathbb{1}_{\{s_1(t)=1\}} = 1 - \frac{1 - s_2(t)}{\lambda}.$$

It is then easily checked that Equation (32) is of the form

$$\dot{a}_1(t) = \lambda(1 - P_0(s(t))) + \lambda(1 - s_1(t))P_0(s(t)),$$

exactly as in Equation (31), where the last equality used the property $s_1(t) = 1$.

The derivatives of a_i, for $i > 1$, and of d_i, for $i \ge 1$, are obtained using similar arguments, which are omitted. □

For every sample path outside a zero-measure set, we have established the following. Proposition 5.3 implies the existence of limit points of the process S^n. Furthermore, according to Proposition 5.4 these limit points verify the differential equations of the fluid model. Since all stochastic trajectories $S^n(t)$ take values in $\mathscr{S}$ (which is a closed set), their limits are functions taking values in $\mathscr{S}$ as well. We will now show that the limit $s(t)$ actually belongs to the smaller set $\mathscr{S}^1$, which is a requirement in our definition of fluid solutions. Using the same argument as in the proof of Proposition 4.5, it can be shown that

$$\frac{d}{dt}\|s(t)\|_1 \le \lambda,$$

for all regular times t. Since the trajectories s are continuous with respect to our weighted norm $\|\cdot\|_w$, but not necessarily with respect to the 1-norm, it now remains to be checked that the 1-norm cannot become infinite at a nonregular time.

Suppose that t_1 is a nonregular time. Recall, from the proof of Proposition 4.5, that such a time may occur only once, and only in the High Message regime, if trajectory hits the set

$$D = \{s \in \mathcal{S}: \, s_1 = 1, s_2 > 1 - \lambda\}.$$

For all $t < t_1$, we have $P_0(s(t)) = 0$, and thus $\dot{s}_i(t) \le 0$, for all $t < t_1$ and all $i \ge 2$. Combining this with the continuity of the coordinates, we obtain $s_i(t_1) \le s_i(0)$, for all $i \ge 2$. It follows that

$$\|s(t_1)\|_1 \le 1 + s_1(t_1) + \sum_{i=2}^{\infty} s_i(0) \le 2 + \|s(0)\|_1.$$

Combining this with the fact that $\|s(0)\|_1 < \infty$, we get that $\|s(t)\|_1 < \infty$, for all $t \ge 0$, and thus $s(t) \in \mathcal{S}^1$, for all $t \ge 0$. This implies the existence of fluid solutions, thus completing the proof of Theorem 3.1.

Moreover, we have already established a uniqueness result in Theorem 3.1: for any initial condition $s^0 \in \mathcal{S}^1$, we have at most one fluid solution. We also have (Proposition 5.3) that every subsequence of S^n has a further subsequence that converges—by necessity to the same (unique) fluid solution. It can be seen that this implies the convergence of S^n to the fluid solution, thus proving Theorem 3.2.

6. Stochastic Steady-State Analysis—Proofs of Proposition 3.3 and Theorem 3.4

In this section, we prove Proposition 3.3 and Theorem 3.4, which assert that for any finite n, the stochastic system is positive recurrent with some invariant distribution π^n and that the sequence of the marginals of the invariant distributions, $\{\pi_s^n\}_{n=1}^{\infty}$, converges in distribution to a measure concentrated on the unique equilibrium of the fluid model. These results guarantee that the properties derived from the equilibrium s^* of the fluid model, and specifically for the asymptotic delay, are an accurate approximation of the steady state of the stochastic system for n large enough.

6.1. Stochastic Stability of the n-th System

We will use the Foster-Lyapunov criterion to show that for any fixed n, the continuous-time Markov process $(S^n(t), M^n(t))$ is positive recurrent.

Our argument is developed by first considering a detailed description of the system:

$$(Q_1(t), \ldots, Q_n(t), M^n(t)),$$

which keeps track of the size of each queue, but without keeping track of the identities of the servers with associated tokens in the virtual queue. As hinted in Section 3.3.1, this is a continuous-time Markov process, on the state space

$$Z_n \triangleq \left\{(q_1, \ldots, q_n, m) \in \mathbb{Z}_+^n \times \{0, 1, \ldots, c(n)\}: \, \sum_{i=1}^{n} \mathbb{1}_{\{q_i = 0\}} \ge m\right\}.$$

The transition rates, denoted by $r^n_{\cdot \to \cdot}$ are as follows, where we use e_i to denote the i-th unit vector in $\mathbb{Z}_+^n$.

1. When there are no tokens available ($m = 0$), each queue sees arrivals with rate λ:

$$r^n_{(q,0) \to (q+e_i,0)} = \lambda, \quad i = 1, \ldots, n.$$

2. When there are tokens available ($m > 0$), the arrival stream, which has rate $n\lambda$, is divided equally between all empty queues:

$$r^n_{(q,m) \to (q+e_i, m-1)} = \frac{n\lambda \mathbb{1}_{\{q_i = 0\}}}{\sum_{j=1}^{n} \mathbb{1}_{\{q_j = 0\}}} \mathbb{1}_{\{m > 0\}}, \quad i = 1, \ldots, n.$$

3. Transitions due to service completions occur at a uniform rate of 1 at each queue, and they do not affect the token queue:

$$r^n_{(q,m) \to (q-e_i, m)} = \mathbb{1}_{\{q_i > 0\}}, \quad i = 1, \ldots, n.$$

4. Messages from idling servers are sent to the dispatcher (hence resulting in additional tokens) at a rate equal to $\mu(n)$ times the number of empty servers that do not already have associated tokens in the virtual queue:

$$r^n_{(q,m) \to (q, m+1)} = \mu(n) \left(\sum_{i=1}^{n} \mathbb{1}_{\{q_i = 0\}} - m \right) \mathbb{1}_{\{m < c(n)\}}.$$

Note that the Markov process of interest, $(S^n(t), M^n(t))$, is a function of the process $(Q(t), M^n(t))$. Therefore, to establish positive recurrence of the former, it suffices to establish positive recurrence of the latter, as in the proof that follows.

Proof of Proposition 3.3. The Markov process $(Q(t), M^n(t))$ on the state space Z_n is clearly irreducible, with all states reachable from each other. To show positive recurrence, we define the quadratic Lyapunov function

$$\Phi(q,m) \triangleq \frac{1}{n}\sum_{i=1}^{n} q_i^2, \tag{33}$$

and note that

$$\sum_{(q',m')\neq(q,m)} \Phi(q',m') r^n_{(q,m)\to(q',m')} < \infty, \quad \forall\, (q,m)\in Z_n.$$

We also define the finite set

$$F_n \triangleq \left\{ (q,m)\in Z_n \colon \sum_{i=1}^{n} q_i < \frac{n(\lambda+2)}{2(1-\lambda)} \right\}.$$

As q_i can change but at most 1 during a transition, we use the relations $(q_i+1)^2 - q_i^2 = 2q_i + 1$ and $(q_i-1)^2 - q_i^2 = -2q_i + 1$. For any (q,m) outside the set F_n, we have

$$\begin{aligned}\sum_{(q',m')\in Z_n} [\Phi(q',m') - \Phi(q,m)] r^n_{(q,m)\to(q',m')} &= \frac{1}{n}\sum_{i=1}^{n}\left[(2q_i+1)\lambda\left(\frac{n\mathbb{1}_{\{q_i=0\}}}{\sum_{j=1}^{n}\mathbb{1}_{\{q_j=0\}}}\mathbb{1}_{\{m>0\}} + \mathbb{1}_{\{m=0\}} \right) - (2q_i-1)\mathbb{1}_{\{q_i>0\}} \right] \\ &= \lambda + \frac{1}{n}\sum_{i=1}^{n}[\mathbb{1}_{\{q_i>0\}} - 2q_i(1-\lambda\mathbb{1}_{\{m=0\}})] \\ &\le \lambda + 1 - \frac{2(1-\lambda)}{n}\sum_{i=1}^{n} q_i \le -1, \quad \forall\, (q,m)\in Z_n \backslash F_n.\end{aligned}$$

The last equality is obtained through a careful rearrangement of terms; the first inequality is obtained by replacing each indicator function by unity. Then, the Foster-Lyapunov criterion (Foster 1953) implies positive recurrence. □

6.2. Convergence of the Invariant Distributions

As a first step towards establishing the interchange of limits result, we start by establishing some tightness properties, in the form of uniform (over all n) upper bounds for $\mathbb{E}_{\pi^n}[\|S^n\|_1]$ and for $\pi^n(Q_1^n \ge k)$. One possible approach to obtaining such bounds is to use an appropriate coupling and show that our system is stochastically dominated by a system consisting of n independent parallel $M/M/1$ queues. However, we follow an easier approach based on a simple linear Lyapunov function and the results in Hajek (1982) and Bertsimas et al. (2002).

Lemma 6.1. *Let π^n be the unique invariant distribution of the process $(Q^n(t), M^n(t))$. We then have the uniform upper bounds*

$$\pi^n(Q_1^n \ge k) \le \left(\frac{1}{2-\lambda}\right)^{k/2}, \quad \forall\, n,\ \forall\, k,$$

and

$$\mathbb{E}_{\pi^n}[\|S^n\|_1] \le 2 + \frac{2}{1-\lambda}, \quad \forall\, n.$$

Proof. Consider the linear Lyapunov function

$$\Psi(q,m) = q_1.$$

Under the terminology in Bertsimas et al. (2002), this Lyapunov function has exception parameter $B = 1$, drift $\gamma = 1-\lambda$, maximum jump $\nu_{\max} = 1$, and maximum rate $p_{\max} \le 1$. Note that this function is not a witness of stability because the set $\{(q,m)\in Z_n \colon \Psi(q,m) < 1\}$ is not finite. However, the boundedness of the upward jumps allows us to use Theorem 2.3 from Hajek (1982) to obtain that $\mathbb{E}_{\pi^n}[Q_1^n] < \infty$. Thus, all conditions in Theorem 1 in Bertsimas et al. (2002) are satisfied, yielding the upper bounds

$$\pi^n(Q_1^n \ge 1 + 2m) \le \left(\frac{1}{2-\lambda}\right)^{m+1}, \quad \forall\, m \ge 1,$$

and

$$\mathbb{E}_{\pi^n}[Q_1^n] \le 1 + \frac{2}{1-\lambda}.$$

The first part of the result is obtained by letting $m = (k-1)/2$ if k is odd or $m = k/2 - 1$ if k is even. Finally, using the definition $\|S^n\|_1 = 1 + (1/n)\sum_{i=1}^n Q_i$, which, together with symmetry yields

$$\mathbb{E}[\|S^n\|_1] = 1 + \frac{1}{n}\sum_{i=1}^{n}\mathbb{E}[Q_i] = \mathbb{E}[Q_1],$$

and concludes the proof. □

We now prove our final result on the interchange of limits.

Proof of Theorem 3.4. Consider the set $\mathbb{Z}_+ \cup \{\infty\}$ endowed with the topology of the Alexandroff compactification, which is known to be metrizable. Moreover, it can be seen that the topology defined by the norm $\|\cdot\|_w$ on $[0,1]^{\mathbb{Z}_+}$ is equivalent to the product topology, which makes $[0,1]^{\mathbb{Z}_+}$ compact. As a result, the product $\{s \in \mathcal{S}^1 \colon \|s\|_1 \le M\} \times (\mathbb{Z}_+ \cup \{\infty\})$ is closed, and thus compact, for all M. Note that, for each n, the invariant distribution π^n is defined over the set $(\mathcal{S}^1 \cap \mathcal{I}_n) \times \{0, 1, \ldots, c(n)\}$. This is a subset of $\mathcal{S}^1 \times (\mathbb{Z}_+ \cup \{\infty\})$, so we can extend the measures π^n to the latter, larger set.

Let $\{S^n(0)\}_{n=1}^\infty$ be a sequence of random variables distributed according to the marginals $\{\pi_s^n\}_{n=1}^\infty$. From Lemma 6.1, we have

$$\mathbb{E}_{\pi^n}[\|S^n(0)\|_1] \le 2 + \frac{2}{1-\lambda}, \quad \forall\, n. \tag{34}$$

Using Markov's inequality, it follows that for every $\epsilon > 0$, there exists a constant M such that

$$\pi_s^n(\{s \in \mathcal{S}^1 \colon \|s\|_1 \le M\}) \ge 1 - \epsilon, \quad \forall\, n,$$

which implies that

$$\pi^n(\{s \in \mathcal{S}^1 \colon \|s\|_1 \le M\} \times (\mathbb{Z}_+ \cup \{\infty\})) \ge 1 - \epsilon, \quad \forall\, n.$$

Thus, the sequence $\{\pi^n\}_{n=1}^\infty$ is tight and, by Prohorov's theorem (Billingsley 1999), it is also relatively compact in the weak topology on the set of probability measures. It follows that any subsequence has a weakly convergent subsequence whose limit is a probability measure over $\mathcal{S}^1 \times (\mathbb{Z}_+ \cup \{\infty\})$.

Let $\{\pi^{n_k}\}_{k=1}^\infty$ be a weakly convergent subsequence, and let π be its limit. Let $S(0)$ be a random variable distributed according to π_s, where π_s is the marginal of π. Since $\mathcal{S}^1 \times (\mathbb{Z}_+ \cup \{\infty\})$ is separable, we invoke Skorokhod's representation theorem to construct a probability space $(\Omega_0, \mathcal{A}_0, \mathbb{P}_0)$ and a sequence of random variables $(S^{n_k}(0), M^{n_k}(0))$ distributed according to π^{n_k}, such that

$$\lim_{k\to\infty} \|S^{n_k}(0) - S(0)\|_w = 0 \quad \mathbb{P}_0\text{-}a.s. \tag{35}$$

We use the random variables $(S^{n_k}(0), M^{n_k}(0))$ as the initial conditions for a sequence of processes $\{(S^{n_k}(t), M^{n_k}(t))\}_{k=1}^\infty$, so that each one of these processes is stationary. Note that the initial conditions, distributed as π^{n_k}, do not necessarily converge to a deterministic initial condition (this is actually part of what we are trying to prove), so we cannot use Theorem 3.2 directly to find the limit of the sequence of processes $\{S^{n_k}(t)\}_{k=1}^\infty$. However, given any $\omega \in \Omega_0$ outside a zero $\mathbb{P}_0$-measure set, we can restrict this sequence of stochastic processes to the probability space

$$(\Omega_\omega, \mathcal{A}_\omega, \mathbb{P}_\omega) = (\Omega_D \times \Omega_S \times \{\omega\}, \mathcal{A}_D \times \mathcal{A}_S \times \{\emptyset, \{\omega\}\}, \mathbb{P}_D \times \mathbb{P}_S \times \delta_\omega)$$

and apply Theorem 3.2 to this new space, to obtain

$$\lim_{k\to\infty} \sup_{0\le t\le T} \|S^{n_k}(t,\omega) - S(t,\omega)\|_w = 0, \quad \mathbb{P}_\omega\text{-}a.s.,$$

where $S(t,\omega)$ is the fluid solution with initial condition $S(0,\omega)$. Since this is true for all $\omega \in \Omega_0$ except for a set of zero P_0-measure, it follows that

$$\lim_{k\to\infty} \sup_{0\le t\le T} \|S^{n_k}(t) - S(t)\|_w = 0, \quad \mathbb{P}\text{-}a.s.,$$

where $\mathbb{P} = \mathbb{P}_D \times \mathbb{P}_S \times \mathbb{P}_0$ and where $S(t)$ is another stochastic process whose randomness is only in the initial condition $S(0)$ (its trajectory is the deterministic fluid solution for that specific initial condition).

We use Lemma 6.1 once again to interchange limit, expectation, and infinite summation in Equation (34) (using the same argument as in Lemma A.1) to obtain

$$\mathbb{E}_{\pi_s}[\|S(0)\|_1] \le 2 + \frac{2}{1-\lambda}.$$

Using Markov's inequality now in the limit, it follows that for every $\epsilon > 0$, there exists a constant M such that

$$\pi_s(\|S(0)\|_1 \le M) \ge 1 - \epsilon. \tag{36}$$

Recall that the uniqueness of fluid solutions (Theorem 3.1) implies the continuous dependence of solutions on initial conditions (Filippov 1988). Moreover, Theorem 3.1 implies that any solution $s(t)$ with initial conditions $s(0) \in \mathcal{S}^1$ converges to s^* as $t \to \infty$. As a result, there exists $T_\epsilon > 0$ such that

$$\sup_{s(0):\, \|s(0)\|_1 \le M} \|s(T_\epsilon) - s^*\|_w < \epsilon.$$

Combining this with Equation (36), we obtain

$$\begin{aligned}\mathbb{E}_{\pi_s}[\|S(T_\epsilon)-s^*\|_w] &= \mathbb{E}_{\pi_s}[\|S(T_\epsilon)-s^*\|_w \mid \|S(0)\|_1 \le M]\pi_s(\|S(0)\|_1 \le M) + \mathbb{E}_{\pi_s}[\|S(T_\epsilon)-s^*\|_w \mid \|S(0)\|_1 > M]\pi_s(\|S(0)\|_1 > M) \\ &< \epsilon + \left(\sup_{s\in\mathcal{S}} \|s - s^*\|_w\right)\epsilon \le 2\epsilon,\end{aligned} \tag{37}$$

where the expectations $\mathbb{E}_{\pi_s}$ are with respect to the random variable $S(0)$, distributed according to π_s, even though the dependence on $S(0)$ is suppressed from our notation and is left implicit. On the other hand, due to the stationarity of $S^{n_k}(\cdot)$, the random variables $S^{n_k}(0)$ and $S^{n_k}(T_\epsilon)$ have the same distribution, for any k. Taking the limit as $k \to \infty$, we see that $S(0)$ and $S(T_\epsilon)$ have the same distribution. Combining this with Equation (37), we obtain

$$\mathbb{E}_{\pi_s}[\|S(0) - s^*\|_w] \le 2\epsilon.$$

Since ϵ was arbitrary, it follows that $S(0) = s^*$, π_s-almost surely, i.e., the distribution π_s of $S(0)$ is concentrated on s^*. We have shown that the limit π_s of a convergent subsequence of π^n is the Dirac measure δ_{s^*}. Since this is true for every convergent subsequence and π^n is tight, this implies that π^n converges to δ_{s^*}, as claimed. □

7. Conclusions and Future Work

The main objective of this paper was to study the tradeoff between the amount of resources (messages and memory) available to a central dispatcher, and the expected queueing delay as the system size increases. We introduced a parametric family of pull-based dispatching policies and, using a fluid model and associated convergence theorems, we showed that with enough resources, we can drive the queueing delay to zero as the system size increases.

We also analyzed a resource constrained regime of our pull-based policies that, although it does not have vanishing delay, it has some remarkable properties. We showed that by wisely exploiting an arbitrarily small message rate (but still proportional to the arrival rate) we obtain a queueing delay which is finite and uniformly upper bounded for all $\lambda < 1$, a significant qualitative improvement over the delay of the $M/M/1$ queue (obtained when we use no messages). Furthermore, we compared it with the popular power-of-d-choices and found that while using the same number of messages, our policy achieves a much lower expected queueing delay, especially when λ is close to 1.

Moreover, in a companion paper we show that *every* dispatching policy (within a broad class of policies) that uses the same amount of resources as our policy in the constrained regime, results in a non-vanishing queueing delay. This implies that our family of policies is able to achieve vanishing delay with the minimum amount of resources in some sense.

There are several interesting directions for future research.

(i) It would be interesting to extend these results to the case of different service disciplines such as processor sharing or LIFO, or to the case of general service time distributions, as these are prevalent in many applications.

(ii) We have focused on a system with homogeneous servers. For the case of nonhomogeneous servers, even stability can become an issue, and there are interesting tradeoffs between the resources used and the stability region.

(iii) Another interesting line of work is to consider a reverse problem, in which we have decentralized arrivals to several queues, a central server, and a scheduler that needs to decide which queue to serve. In this context we expect to find again a similar tradeoff between the resources used and the queueing delay.

Acknowledgments

A preliminary version of this paper appeared in the Proceedings of the 2016 ACM Sigmetrics Conference (Gamarnik et al. 2016).

Appendix A. Interchange of Limit, Expectation, and Infinite Summation

Lemma A.1. *We have*

$$\lim_{n\to\infty} \mathbb{E}\left[\sum_{i=1}^{\infty} S_i^n\right] = \sum_{i=1}^{\infty} s_i^*.$$

Proof. By Fubini's theorem, we have

$$\lim_{n\to\infty} \mathbb{E}\left[\sum_{i=1}^{\infty} S_i^n\right] = \lim_{n\to\infty} \sum_{i=1}^{\infty} \mathbb{E}[S_i^n].$$

Due to the symmetric nature of the invariant distribution π^n, we have

$$\mathbb{E}[S_i^n] = \mathbb{E}\left[\frac{1}{n}\sum_{j=1}^{n} \mathbb{1}_{[i,\infty)}(Q_j^n)\right] = \mathbb{E}[\mathbb{1}_{[i,\infty)}(Q_1^n)] = \pi^n(Q_1^n \geq i) \leq \left(\frac{1}{2-\lambda}\right)^{i/2},$$

where the inequality is established in Lemma 6.1. We can therefore apply the Dominated Convergence Theorem to interchange the limit with the first summation, and obtain

$$\lim_{n\to\infty} \sum_{i=1}^{\infty} \mathbb{E}[S_i^n] = \sum_{i=1}^{\infty} \lim_{n\to\infty} \mathbb{E}[S_i^n],$$

We already know that S_i^n converges to s^*, in distribution (Theorem 3.4). Then, using a variant of the Dominated Convergence Theorem for convergence in distribution, and the fact that we always have $S_i^n \leq 1$, we can finally interchange the limit and the expectation and obtain

$$\sum_{i=1}^{\infty} \lim_{n\to\infty} \mathbb{E}[S_i^n] = \sum_{i=1}^{\infty} s_i^*. \quad \square$$

References

Aghajani R, Ramanan K (2017) The hydrodynamic limit of a randomized load balancing network. Preprint arXiv:1707.02005.

Badonnel R, Burgess M (2008) Dynamic pull-based load balancing for autonomic servers. Brunner N, Becker Westphall C, Zambenedetti Granville L, eds. *Proc. Network Oper. Management Sympos.* (*NOMS*), (IEEE, Piscataway, NJ), 751–754.

Bertsimas D, Gamarnik D, Tsitsiklis JN (2002) Performance of multiclass Markovian queueing networks via piecewise linear Lyapunov functions. *Ann. Appl. Probab.* 11(4):1384–1428.

Billingsley P (1999) *Convergence of Probability Measures*, Second ed. (John Wiley & Sons, New York).

Bramson M (1998) State space collapse with application to heavy traffic limits for multiclass queueing networks. *Queueing Systems: Theory Appl.* 30(1–2):89–148.

Bramson M, Lu Y, Prabhakar B (2013) Decay tails at equilibrium for FIFO join the shortest queue networks. *Ann. Appl. Probab.* 23(5): 1841–1878.

Filippov AF (1988) *Differential Equations with Discontinuous Righthand Sides* (Springer, Dordrecht, Netherlands).

Foss S, Stolyar AL (2017) Large-scale join-idle-queue system with general service times. Preprint arXiv:1605.05968.

Foster FG (1953) On the stochastic matrices associated with certain queueing processes. *Ann. Math. Statist.* 24(3):355–360.

Gamarnik D, Tsitsiklis JN, Zubeldia M (2016) Delay, memory, and messaging tradeoffs in distributed service systems. Gamarnik D, Tsitsiklis JN, Zubeldia M, eds. *Proc. ACM SIGMETRICS Internat. Conf. Measurement Model. Comput. Sci.* (ACM, New York), 1–12.

Hajek B (1982) Hitting-time and occupation-time bounds implied by drift analysis with applications. *Adv. Appl. Probab.* 14(3):502–525.

Hunt PJ, Kurtz TG (1994) Large loss networks. *Stochastic Processes and Their Appl.* 53(2):363–378.

Kirszbraun MD (1934) Über die zusammenziehende und Lipschitzsche Transformationen. *Fund. Math* 22(1):77–108.

Kurtz TG (1981) *Approximation of Population Processes* (SIAM, Philadelphia).

Lobanov SG, Smolyanov OG (1994) Ordinary differential equations in locally convex spaces. *Uspekhi Mat. Nauk* 49(3):93–168.

Lu Y, Xie Q, Kliot G, Geller A, Larus JR, Greenberg A (2011) Join-Idle-queue: A novel load balancing algorithm for dynamically scalable web services. *Performance Evaluation* 68(11):1056–1071.

Mitzenmacher MD (1996) The power of two choices in randomized load balancing. Ph.D. thesis, University of California, Berkeley.

Mitzenmacher M (2016) Analyzing distributed join-idle-queue: A fluid limit approach. *Proc. 54th Annual Allerton Conf. Commun., Control, Comput.* (IEEE, Piscataway, NJ), 312–318.

Mitzenmacher M, Prabhakar B, Shah D (2002) Load balancing with memory. *Proc. 43rd Annual IEEE Sympos. Foundations Comput. Sci.* (*FOCS*) (IEEE Computer Society, Washington, DC), 799–808.

Mukherjee D, Borst S, van Leeuwaarden J, Whiting P (2016) Universality of power-of-d load balancing schemes. *ACM SIGMETRICS Performance Evaluation Rev.* 44(2):36–38.

Rudin W (1976) *Principles of Mathematical Analysis*, 3rd ed. (McGraw-Hill, New York).

Shwartz A, Weiss A (1995) *Large Deviations for Performance Analysis: Queues, Communications, and Computing* (Chapman & Hall, London).

Stolyar AL (2015) Pull-based load distribution in large-scale heterogeneous service systems. *Queueing Systems: Theory Appl.* 80(4):341–361.

Stolyar AL (2017) Pull-based load distribution among heterogeneous parallel servers: The case of multiple routers. *Queueing Systems: Theory Appl.* 85(1–2):31–65.

Tsitsiklis JN, Xu K (2012) On the power of (even a little) resource pooling. *Stochastic Systems* 2(1):1–66.

Van Der Boor M, Borst S, van Leeuwaarden J (2017) Load balancing in large-scale systems with multiple dispatchers. *Proc. IEEE Conf. Comput. Commun.* (*INFOCOM*) (IEEE, Piscataway, NJ).

Vvedenskaya ND, Dobrushin RL, Karpelevich FI (1996) Queueing system with selection of the shortest of two queues: An asymptotic approach. *Problems Inform. Transmission* 32(1):15–27.

Xu K, Yun S-Y (2017) Reinforcement with fading memories. Preprint.

Ying L, Srikant R, Kang X (2015) The power of slightly more than one sample in randomized load balancing. *Proc. IEEE Conf. Comput. Commun.* (*INFOCOM*) (IEEE, Piscataway, NJ), 1131–1139.