Mean-field Dynamics of Load-Balancing Networks with General Service Distributions

Reza Aghajani [*1], Xingjie Li[†2], and Kavita Ramanan[‡1]

[1]*Division of Applied Mathematics, Brown University, Providence, RI, USA.*
[2]*Department of Mathematics and Statistics, University of North Carolina Charlotte, Charlotte, NC, USA.*

December 2, 2015

Abstract

We introduce a general framework for the mean-field analysis of large-scale load-balancing networks with general service distributions. Specifically, we consider a parallel server network that consists of N queues and operates under the $SQ(d)$ load balancing policy, wherein jobs have independent and identical service requirements and each incoming job is routed on arrival to the shortest of d queues that are sampled uniformly at random from N queues. We introduce a novel state representation and, for a large class of arrival processes, including renewal and time-inhomogeneous Poisson arrivals, and mild assumptions on the service distribution, show that the mean-field limit, as $N \to \infty$, of the state can be characterized as the unique solution of a sequence of coupled partial integro-differential equations, which we refer to as the hydrodynamic PDE. We use a numerical scheme to solve the PDE to obtain approximations to the dynamics of large networks and demonstrate the efficacy of these approximations using Monte Carlo simulations. We also illustrate how the PDE can be used to gain insight into network performance.

1 Introduction

Load balancing is an effective method to improve the performance and reliability of networks by optimizing resource use. With growth in the use of server farms and computer cluster, large-scale load balancing networks appear in variety of applications, including internet services such as high-traffic web sites, high-bandwidth File Transfer Protocol sites, Network News Transfer Protocol (NNTP) servers, Domain Name System (DNS) servers, and databases, as well as cloud computing and communications systems.

An extensively studied problem is the design and analysis of load-balancing algorithms that aim to improve network performance. This is particularly challenging for large-scale networks, where it is not feasible to implement classical algorithms like *join-the-shortest-queue*, which incur high communication overhead and computational cost. In this context, randomized algorithms provide an attractive alternative. A popular algorithm that achieves a better balance between network performance and communication overhead is the so-called $SQ(d)$ (or "supermarket") algorithm. This algorithm was introduced

[*]reza@brown.edu
[†]xli47@uncc.edu
[‡]kavita_ramanan@brown.edu

in the case $d = 2$ by Vydenskaya et al. in [23] in the simple setting of a network comprising N homogeneous parallel servers, each with its own queue, that process a common stream of jobs that must be routed immediately on arrival. In the $SQ(d)$ algorithm, upon arrival of a job, d queues are sampled independently and uniformly at random, and the job is routed to the shortest queue amongst those sampled. Although, when $d \geq 2$ and $\lambda < 1$, the stationary distribution of a typical queue is not computable, the limiting stationary distribution, as $N \to \infty$, was explicitly computed in [23] and shown to have a double exponential tail decay, in contrast to the exponential decay when $d = 1$ (which corresponds to random routing). This dramatic improvement in performance gained by adding just one extra random choice is known as the "power of two choices", and has led to substantial interest in this class of randomized load balancing schemes. The extension to general $d > 2$ was studied by Mitzenmacher [21], and a static balls-and-bins analog was originally studied in Azar et al. [3].

The analysis of the $SQ(d)$ model in the case of exponential service times is carried out using the so-called "ODE method". This proceeds by first representing the dynamics of the N-server network by a Markov process $S^{(N)} = (S_\ell^{(N)}; \ell \geq 1)$, where $S_\ell^{(N)}(t)$ represents the number of queues that have ℓ or more jobs at time t, and then showing that, as $N \to \infty$, the sequence of suitably scaled Markov processes converges weakly (on finite time intervals) to the unique solution of a countable system of coupled $[0, 1]$-valued ordinary differential equations (ODEs). This limit result is obtained by a simple application of Kurtz's theorem (see Theorem 11.2.1 in [11]), generalized to countable state spaces. Insight into the equilibrium behavior is obtained by first showing that, under the stability condition $\lambda < 1$, each Markov process $S^{(N)}$ is ergodic, then characterizing the limit of the sequence of scaled stationary distributions as the unique equilibrium point of this system of ODEs [23], and finally explicitly identifying this equilibrium point.

However, in most real-world applications, service times are typically not exponentially distributed. For example, statistical analyses suggest that service times follow a Log-Normal distribution in (medium-scale) call centers [7], a Gamma distribution in Automatic Teller Machines [16], and studies of content download in Amazon S3 suggest that download times follow a shifted exponential distribution [8, 17]. Moreover, Phase-type distributions are used to approximate more complicated service time distributions [10, 22]. In the case of general service times, in order to describe the evolution of the system it is not sufficient to keep track of the fraction of queues with ℓ jobs at any time. For each job in service, one has also to keep track of its age (the amount of time the job has spent in service) or its residual service time. In the system with N servers, this requires keeping track of N additional random variables, and thus the dimension of the Markovian state representation grows with N, which is not convenient for obtaining a limit theorem.

1.1 Prior Work

The supermarket model and its various modifications have been extensively studied for the case of exponential service distributions. The path-space evolution of the supermarket model was studied by Graham [13], the maximum equilibrium queue length was analyzed in [18], and strong approximations were obtained in [19]. The basic model has also been generalized to incorporate features of relevance in applications such as load migration, load stealing and thresholds (see, e.g., [12], [20] and [15]).

Although results on the exponential service time model were obtained almost two decades ago, to the best of our knowledge, there appears to be no prior work that characterizes the transient behavior of the supermarket model with general service time distributions, and until recently, there was almost no work on the equilibrium distribution under the sub-criticality condition $\lambda < 1$. Recent progress on the equilibrium behavior was made in a nice series of papers by Bramson et al. [4, 5, 6], using the so-called "cavity method". In particular, in [6], in was shown that for the sub-class of power law distributions with exponent $-\beta$ with $\beta > 1$, the limiting stationary distribution has a doubly exponential tail if $\beta >$

$d/(d-1)$, an exponential tail if $\beta = d/(d-1)$ and a power law tail if $\beta < d/(d-1)$. However, it should be noted that the tails of the limiting stationary distribution provide only limited information about the stationary distribution of the N-server system because limits do not interchange, that is, the tail of the limiting stationary distribution is not the limit of the tails of the N-server stationary distributions.

Moreover, the results in [6] assume that the cumulative arrivals are Poisson and the service distribution has a decreasing hazard rate function. According to the authors [6, Page 3], extending their results beyond these assumptions using their framework appears to be a difficult problem.

1.2 Our Contributions

In this paper, we introduce a new framework for the study of load balancing networks with general service time distributions. The work combines stochastic modeling (choosing a suitable state representation), with techniques from probability (convergence results), PDEs (analysis of the limit), numerical analysis (stable schemes for solving the PDE) and simulations (validation) to provide engineering insights into the performance of the network under the SQ(2) load balancing algorithm in the presence of general service times. Our specific contributions include the following:

1. Development of the PDE method for analyzing load-balancing networks with general service distributions, which can be viewed as a generalization of the well known ODE method [23, 21] used to study large-scale networks with exponential service distributions. Our method applies to the large class of service distributions that have a density and finite mean, including Pareto, Log-Normal, Gamma and Phase type distributions. In particular, it does not require the decreasing hazard rate function assumption imposed in [6].

2. The PDE can be used to approximate not only the queue length distribution but also other quality of service (QoS) parameters such as the virtual waiting time.

3. In contrast with previous work, our framework also enables the study of transient behavior of load-balancing networks with general service distributions, in time-homogeneous networks as well as networks with time-inhomogeneous (Poisson) arrivals, both of which seem relevant for real-world applications.

4. The stable numerical scheme that we use to approximate the PDE provides a computationally efficient alternative to Monte Carlo simulations that could be useful for studying the performance and design of large networks, as illustrated in Sections 6 and 7.

5. Although, it is known that networks with heavy-tailed service distributions have worse steady-state performances [6], using our method, we identified the somewhat surprising phenomenon that they can lead to better performance with respect to some transient QoS parameters, as detailed in Section 7.

The rest of the paper is organized as follows. The structure and representation of the load-balancing network are described in Section 2. The hydrodynamic PDE is presented in Section 3. Section 4 contains the main results of the paper, and their proofs are given in Section 5. In Section 6, we present the numerical scheme to solve the PDE and validate the main results using Monte Carlo simulations. Engineering insights gained by the PDE are illustrated in Section 7. Finally, conclusions and future works are discussed in Section 8.

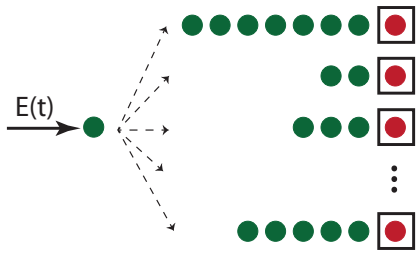


Figure 1: The load-balancing network.

2 A Load Balancing Network

2.1 Model Description

Consider a network of N homogeneous parallel servers, each with its own infinite capacity queue, that processes a common stream of arriving jobs (see Figure 1) that are routed immediately on arrival according to a load-balancing algorithm. Each server follows a first-in-first-out (FIFO) service policy, and is non-idling in the sense that it cannot be idle if there is a job waiting in the queue. Hence, if a job is routed to an idle server, it immediately starts receiving service, and otherwise, it is placed at the end of the queue. We index servers by $i = 1, ..., N$.

Let $E^{(N)}$ be the cumulative arrival process, that is, for every $t \geq 0$, $E^{(N)}(t)$ is the number of job arrivals during the interval $[0, t]$. Jobs are indexed by $j \in \mathbb{Z}$, and each job j has a service time requirement of v_j. The sequence of service times $\{v_j\}_{j\in\mathbb{Z}}$ is assumed to be i.i.d., independent of the arrival process, and distributed according to a general distribution function G. Let $\overline{G}$ represent the complementary service distribution, i.e., $\overline{G}(x) = 1 - G(x)$.

For each server i and at each time $t \geq 0$, the number of jobs in server i (including the one receiving service and the ones waiting in the queue) is called the queue length of server i at time t, and is denoted by $X^{(N),i}(t)$. The total number of jobs in system is denoted by $X^{(N)}(t)$, whence $X^{(N)}(t) \doteq \sum_{i=1}^{N} X^{(N),i}(t)$.

We study the performance of the network described above under a randomized load balancing algorithm that we refer to as the $SQ(d)$ algorithm. Based on this algorithm, upon arrival of a job, d queues are sampled independently and uniformly at random, and the job is then routed to the queue with minimum length (with ties broken uniformly at random). In this paper, we present our results for the case $d = 2$; however, the extension for $d \geq 2$ is almost immediate.

2.2 Assumptions

The following assumptions are imposed on the cumulative arrival process and service time distribution.

Assumption I. (Arrival Process) For every $N \in \mathbb{N}$, $E^{(N)}$ is a time-inhomogeneous Poisson process with rate $\lambda^{(N)}(\cdot) \doteq a^{(N)}\lambda(\cdot)$, where $\lambda(\cdot)$ is a non-negative, locally square integrable function (that is, $\int_0^T \lambda(t)dt < \infty$ for every $T < \infty$) and $\{a^{(N)}\}_{N\in\mathbb{N}}$ is a sequence of positive real numbers satisfying $a^{(N)}/N \to 1$ as $N \to \infty$.

Remark 2.1. Assumption I is satisfied by most interesting cases including the heavy traffic limit $\lambda^N(\cdot) = N\lambda(\cdot) - \beta\sqrt{N}$. However, as shown in forthcoming work, the results of this paper hold for more general sequences of arrival rates.

Recall that when a distribution G has a density g, the hazard rate function h associated with that distribution is defined by

$$h(x) \doteq \frac{g(x)}{1-G(x)}, \qquad x \in [0, L), \tag{1}$$

where $L \doteq \sup\{x \in [0,\infty) : G(x) < 1\}$.

Assumption II. (Service Distribution) The distribution function G has a finite mean equal to 1, and density g. Also, the hazard rate h is uniformly bounded by a constant H, that is,

$$\sup_{x\in[0,L)} h(x) \leq H.$$

Remark 2.2. Assumption II is satisfied by a wide class of service time distributions, including any Pareto distribution with finite mean, Phase-Type distributions, Log-Normal distributions and Gamma distributions. Although the boundedness assumption on h can be relaxed, it makes the representation of the results in Sections 4 considerably easier. Note that since for every $a, b \in [0, L)$,

$$\int_a^b h(x)dx = \ln(\overline{G}(a)) - \ln(\overline{G}(b)),$$

h is always locally integrable on $[0, L)$, but not integrable. Therefore, Assumption II implies that $L = \infty$, that is, the distribution G is supported on the whole $[0,\infty)$.

2.3 An Important State Descriptor

For every $\ell \geq 1$ and time $t \geq 0$, let $\mathcal{S}_\ell^{(N)}(t)$ denote the subset of all servers with queue length at least ℓ at time t, that is,

$$\mathcal{S}_\ell^{(N)}(t) \doteq \left\{i = 1, ..., N : X^{(N),i}(t) \geq \ell\right\}, \tag{2}$$

and denote by $S_\ell^{(N)}$ the number of such servers, that is,

$$S_\ell^{(N)}(t) \doteq \#\left\{i = 1, ..., N : X^{(N),i}(t) \geq \ell\right\} = \#\mathcal{S}_\ell^{(N)}(t). \tag{3}$$

Note that $\mathcal{S}_1^{(N)}(t)$ is the set of servers with at least one job, to which we refer as "busy servers". As each server uses a FIFO policy and is non-idling, there is a unique job being served at each busy server. Hence, at each time t, we can assign to every busy server i an *age*, denoted by $a^{(N),i}(t)$, which is defined to be the amount of time that the current job has been receiving service.

As discussed in the introduction, if the service times are exponentially distributed, the process $\{(S_\ell^{(N)}(t); \ell \geq 1); t \geq 0\}$ is a convenient state descriptor for the N-server network. But, since other service distributions do not have the memoryless property (exponential is unique), the dynamics of the network depends on the ages of all jobs in service. Hence, in the case of general serviced distributions, the following state descriptor turns out to be convenient. For every $N \in \mathbb{N}$, $\ell \geq 1$ and $t \geq 0$, define the functions $Z_\ell^{(N)}(t, \cdot)$ as

$$Z_\ell^{(N)}(t, r) \doteq \sum_{i\in\mathcal{S}_\ell^{(N)}} \frac{\overline{G}(a^{(N),i}(t) + r)}{\overline{G}(a^{(N),i}(t))}, \qquad r \geq 0, \tag{4}$$

where note that the summation is over servers at queues with length at least ℓ. For an intuitive understanding of this quantity, note that for any $N \in \mathbb{N}$, the conditional probability that a job which is being

served at time t in a busy server i will still be in system at time $t+r$ for $r \geq 0$, is $\overline{G}(a^{(N),i}(t)+r)/\overline{G}(a^{(N),i}(t))$. Therefore, $Z_\ell^{(N)}(t,r)$ is the expected number of jobs that are being served at time t in a server with queue length at least ℓ and that will still be in the system at time $t+r$. Note in particular that $Z_\ell^{(N)}(t,0)$ is equal to the cardinality of $\mathcal{S}_\ell^{(N)}(t)$, that is

$$Z_\ell^{(N)}(t,0) = S_\ell^{(N)}(t). \tag{5}$$

3 A PDE Approximation

In this section, we introduce a set of partial integro-differential equations that we call the hydrodynamic PDE. As shown in Section 4, these coupled equations uniquely characterize the limit, as $N \to \infty$, of the scaled state descriptors $\overline{Z}_\ell^{(N)}$ defined as

$$\overline{Z}_\ell^{(N)}(t,r) \doteq \frac{1}{N} Z_\ell^{(N)}(t,r).$$

The equations are described in Section 3.1. Then, in Section 3.2, we show how these set of PDEs reduce to the set of ODEs obtained in [23] when the service distribution is exponential. Finally, we provide an intuitive interpretation of the hydrodynamic PDE in Section 3.3.

3.1 Description of the Hydrodynamic PDE

Define $\mathbb{C}^1_{\leq 1}[0,\infty)$ to be the space of continuously differentiable functions f with bounded derivative, such that $|f|$ is bounded by 1. Also, define $\mathbb{X}$ to be the set of functions φ on $[0,\infty)\times[0,\infty)$ such that for every $t \geq 0$, $\varphi(t,\cdot) \in \mathbb{C}^1_{\leq 1}[0,\infty)$ with $\partial_r\varphi(t,\cdot)$ denoting the derivative with respect to the second variable, and for every $r \geq 0$, the mapping $t \mapsto \varphi(t,r)$ is measurable. Now we define the PDE.

Definition 3.1. For every function λ and set of functions $Z_\ell^0 \in \mathbb{C}^1_{\leq 1}[0,\infty)$, $\ell \geq 1$, a set of functions $(Z_\ell; \ell \geq 1)$ with $Z_\ell \in \mathbb{X}$ for all $\ell \geq 1$ is said to solve the hydrodynamic PDE associated to $(\lambda, Z_\ell^0; \ell \geq 1)$ if for every $t, r \geq 0$, Z_1 satisfies

$$Z_1(t,r) = Z_1^0(t+r) - \int_0^t \overline{G}(t+r-u)\partial_r Z_2(u,0)du + \int_0^t \lambda(u)\overline{G}(t+r-u)(1-Z_1(u,0)^2)du, \tag{6}$$

with the boundary condition

$$Z_1(t,0) = Z_1^0(0) + \int_0^t \left(\partial_r Z_1(u,0) - \partial_r Z_2(u,0)\right)du + \int_0^t \lambda(u)\left(1-Z_1(u,0)^2\right)du, \tag{7}$$

and for $\ell \geq 2$, Z_ℓ satisfies

$$\begin{aligned} Z_\ell(t,r) = Z_\ell^0(t+r) &- \int_0^t \overline{G}(t+r-u)\partial_r Z_{\ell+1}(u,0)du \\ &+ \int_0^t \lambda(u)(Z_{\ell-1}(u,0)+Z_\ell(u,0))(Z_{\ell-1}(u,t+r-u)-Z_\ell(u,t+r-u))du, \end{aligned} \tag{8}$$

with the boundary conditions

$$Z_\ell(t,0) = Z_\ell^0(0) + \int_0^t \left(\partial_r Z_\ell(u,0) - \partial_r Z_{\ell+1}(u,0)\right)du + \int_0^t \lambda(u)\left(Z_{\ell-1}(u,0)^2 - Z_\ell(u,0)^2\right)du. \tag{9}$$

Note that by equations (6) and (8), Z_ℓ^0 is the initial condition for Z_ℓ, that is $Z_\ell(0,r) = Z_\ell^0(r)$, for all $\ell \geq 1$ and $r \geq 0$.

Remark 3.2. Although the equations (6)-(9) are partial integro-differential equations and not partial differential equations in the classic sense, we still refer to them as hydrodynamic PDE for conciseness.

3.2 Reduction to ODE in the Exponential Case

To better illustrate the hydrodynamic PDE, we show that when the service distribution is exponential and the arrival rate $\lambda(\cdot) \equiv \lambda$ is constant, it reduces to the set of ordinary differential equations (ODEs) obtained in [23, 21]. Substituting the complimentary CDF $\overline{G}(x) = e^{-x}$ for the exponential distribution in definition (4) of $Z_\ell^{(N)}$, we have

$$Z_\ell^{(N)}(t,r) = \sum_{i \in \mathcal{S}_\ell^{(N)}} \frac{e^{-a^{(N),i}(t)-r}}{e^{-a^{(N),i}(t)}} = e^{-r} S_\ell^{(N)}(t).$$

As made rigorous in Theorem 4.2, this suggests the limit satisfies $Z_\ell(t,r) = e^{-r} S_\ell(t)$, with S_ℓ equal to the limit of $S_\ell^{(N)}/N$. Substituting $\overline{G}(r) = e^{-r}$, Z_ℓ and its derivative $\partial_r Z_\ell(t,r) = -e^{-r} S_\ell(t)$ in (8), for every $\ell \geq 2$ we obtain

$$S_\ell(t) = e^{-t} S_\ell(0) + \int_0^t e^{-(t-u)} S_{\ell+1}(u) du + \lambda \int_0^t e^{-(t-u)} (S_{\ell-1}(u)^2 - S_\ell(u)^2) du. \tag{10}$$

Taking derivatives with respect to t of both sides of the equation above yields

$$\begin{aligned} \frac{d}{dt} S_\ell(t) = -e^{-t} S_\ell(0) + S_{\ell+1}(t) - \int_0^t e^{-(t-u)} S_{\ell+1}(u) du - \lambda \left(S_{\ell-1}(t)^2 - S_\ell(t)^2 \right) \\ + \lambda \int_0^t e^{-(t-u)} \left(S_{\ell-1}(t)^2 - S_\ell(u)^2 \right) du. \end{aligned}$$

Using (10) to eliminate the third and fifth terms on the right-hand side of the last equation , we obtain

$$\frac{d}{dt} S_\ell(t) = -(S_\ell(t) - S_{\ell+1}(t)) + \lambda (S_{\ell-1}(t)^2 - S_\ell(t)^2). \tag{11}$$

Exactly analogous calculations can be shown to see that equations (6)

$$\frac{d}{dt} S_1(t) = -(S_1(t) - S_2(t)) + \lambda (1 - S_1(t)^2). \tag{12}$$

Similarly, substituting $\overline{G}$ and Z_ℓ in (8), we obtain and, again by taking derivative, we have Moreover, substituting $Z_\ell(t,0) = S_\ell(t)$ and $\partial_r Z_\ell(t,r) = -S_\ell(t)$ in (7) and (9), we again obtain (12) and (11), respectively. Note that the ODEs (12)-(11) coincide with the equations that were previously obtained, e.g., (1) in [21].

3.3 Interpretation of the Hydrodynamic PDE

Existence and uniqueness of solutions to the PDE are established in Theorems 4.2 and 4.1. Here, we first provide a heuristic explanation of each term in the hydrodynamic PDE (6)-(9). Interpreting Z_ℓ as the limit of the sequence $\{Z_\ell^{(N)}/N\}$ defined in (4), note that $Z_\ell(t,r)$ represents the (limit of the) fraction of jobs that were in service at time t at a queue of length at least ℓ and that were still in service at time $t+r$. There are three sources that contribute to $Z_\ell(t,r)$, corresponding to the three terms on the right-hand side of (8). The first is the expected fraction of jobs that were already in service at time 0 at a queue of length at least ℓ and that will still be in service at time $t+r$, which is given by the first term on the right-hand side of (8), namely $Z_\ell(0,t+r) = Z_\ell^0(t+r)$.

The second source accounts for any job that entered service at some time $u \in [0,t]$ due to completion of the job ahead of it in the queue, and that would still not have completed service at time $t+r$ (equivalently, jobs that have a service time greater than $t+r-u$). For such a job to be in a queue of length ℓ or more at time t, it must have been a queue of length $\ell+1$ just prior to the departure of the job ahead of it. Now, to estimate the expected rate of entry into service of such jobs, consider a server that is busy serving a job j with age $a(u)$ at time u. Given this information, the probability that the job will depart within the next ϵ units of time is equal to

$$\mathbb{P}\left\{v_j < a(u)+\epsilon | v_j > a(u)\right\} = \frac{G(a(u)+\epsilon) - G(a(u))}{1-G(a(u))} \approx h(a(u))\epsilon.$$

Therefore, the expected departure rate of jobs from servers with queues of length $\ell+1$ or greater (conditional on their ages) is roughly

$$\sum_{i \in \mathcal{S}_{\ell+1}^{(N)}(u)} h(a^{(N),i}(u)),$$

which can be rewritten as

$$\sum_{i \in \mathcal{S}_{\ell+1}^{(N)}(u)} \frac{g(a^{(N),i}(u))}{1-G(a^{(N),i}(u))} = -\partial_r Z_{\ell+1}^{(N)}(u,0).$$

Thus, $-\partial_r Z_\ell(u,0)$ is (the limit of) the expected departure rate at time u from servers with a queue of length at least $\ell+1$, which is also the expected entry rate into service into a queue of length ℓ. Multiplying this by $\overline{G}(t+r-u)$, the fraction of such jobs that will still be in service at time $t+r$, and integrating over all possible values of $u \in [0,t]$, we obtain the second term on the right-hand side of (8).

The last contribution is due to the routing of new jobs at some time $u \in [0,t]$ to a queue of length $\ell-1$ such that the job in service at that queue at time u is still in service at time $t+r$. Note that such a queue has length ℓ or more at time t and the job in service at that queue in time u is in service both at times t and $t+r$, and hence contributes to $Z_\ell(t,t+r)$. To compute the number of such jobs, first note that the total arrival rate of jobs at time u is $\lambda(u)$, and the expected fraction of such jobs that get routed to a queue of length $\ell-1$ under the $SQ(2)$ routing algorithm is computed in Section 5.3 and is equal to $\overline{S}_{\ell-1}^{(N)}(u)^2 - \overline{S}_\ell^{(N)}(u)^2$, which by (5) is also equal to $\overline{Z}_{\ell-1}^{(N)}(u,0)^2 - \overline{Z}_\ell^{(N)}(u,0)^2$ (here, we use the convention $S_0^{(N)} \equiv 1$.) Now, the total number of jobs in service at a queue of exactly length $\ell-1$ just prior to time u is $\overline{Z}_{\ell-1}^{(N)}(u-,0) - \overline{Z}_\ell^{(N)}(u-,0)$ and the number of these that will still be in service at time $t+r$ is $\overline{Z}_{\ell-1}^{(N)}(u-,t+r-u) - \overline{Z}_\ell^{(N)}(u-,t+r-u)$, which implies the fraction of such jobs still in service at time $t+r$ is given by the ratio. Multiplying the ratio by the previously computed arrival rate of jobs into queues of length $\ell-1$ at time u and integrating over $u \in [0,t]$, we obtain the third term.

The boundary conditions (7) and (9) are basically mass balance equations. Recall that $Z_\ell^{(N)}(t,0)$ is the number of jobs receiving service at a queue of length at least ℓ at time t, and similarly, $Z_\ell^{(N)}(0,0)$ is the number of jobs that were receiving service at time 0 at such a queue. Note that $S_\ell^{(N)} = Z_\ell^{(N)}(\cdot,0)$ decreases only due to departures from servers with queue length equal to ℓ and increases only due to arrivals routed to servers with queue length equal to $\ell-1$. As discussed above, $-\partial_r Z_{\ell+1}^{(N)}(u,r)$ is the departures rate at time u of jobs from a queue of length $\ell+1$ or more at the time of departure, and hence the second term on the right-hand side of (9) represents the limit of the cumulative number of such departures in the interval $[0,t]$ from queues of length exactly ℓ. Finally, the third term on the right-hand side of (9) represents the total number of arrivals to servers at queues of length exactly $\ell-1$ in the interval $[0,t]$, whose form can be deduced from the routing probabilities computed above.

4 Main Results

In this section we state the main results of the paper; the proofs are deferred to Section 5. First in Section 4.1, we show that the PDE (6)-(9) have at most one solution in a suitable space. Then in Section 4.2, we show that under suitable assumptions on the sequence of initial distributions, the scaled sequence $\{\overline{Z}_\ell^{(N)}\}_{N\in\mathbb{N}}$ has a limit which satisfies the PDE (and hence, is the unique solution). Using this convergence result, we show in Theorem 4.3 and Theorem 4.4 that the queue length distribution of a typical queue and the mean virtual waiting time in the N-server network converges to certain functionals of the solution to the hydrodynamic PDE. Moreover, we establish a "propagation of chaos" result, showing that, at any finite time, the queue length distributions of any finite set of queues are asymptotically independent.

4.1 Uniqueness of the Solution to the PDE

We now state our first result. We denote by $\mathbb{L}^1_{\text{loc}}(0,\infty)$ the space of locally integrable functions (that is, functions f on $[0,\infty)$ that satisfy $\int_0^T f(t)dt < \infty$ for every $T < \infty$).

Theorem 4.1. Suppose Assumptions I-II hold. Then, for every non-negative function $\lambda \in \mathbb{L}^1_{\text{loc}}(0,\infty)$ and $Z_\ell^0 \in \mathbb{C}^1_{\le 1}[0,\infty)$, $\ell \ge 1$, then the hydrodynamic PDE (6)-(9) has at most one solution.

4.2 Convergence of the State Descriptor $Z^{(N)}$

Theorem 4.2 below shows that the solution to the hydrodynamic PDE characterizes the limit of $\overline{Z}^{(N)}$ as N gets large. To state the theorem, we need to impose an additional assumption on the initial conditions.

Assumption III. (Initial Condition)

a. Almost surely, for every $\ell \ge 1$ and every function $f \in \mathbb{C}_b[0,\infty)$, the limit

$$\lim_{N\to\infty} \frac{1}{N} \sum_{i\in \mathcal{S}_\ell^{(N)}(0)} f(a^{(N),i}(0)).$$

exist. Moreover, $\limsup_N \mathbb{E}[\overline{X}^{(N)}(0)] < \infty$.

b. For every N, the initial conditions of the N-server network are exchangeable, in the sense that for every permutation σ on the set of server indices $\{1,...,N\}$, the distribution of the vector

$$\left(X^{(N),i}(0), \mathbb{1}(X^{(N),i}(0) > 0)a^{(N),i}(0); i = 1,...,N\right)$$

does not depend on the choice of σ.

Note that in particular, substituting $f = \overline{G}(\cdot + r)/\overline{G}(\cdot)$ in Assumption III.a shows that for every $r \ge 0$, the following limit exists:

$$Z_\ell^0(r) \doteq \lim_{N\to\infty} \overline{Z}_\ell^{(N)}(0,r). \tag{13}$$

This theorem is established in a companion paper [1].

Theorem 4.2. Suppose λ and $\overline{G}$ satisfy Assumptions I and II, let Assumption III.a hold and let $Z_\ell^0, \ell \geq 1$, be as in (13). Then almost surely, for $\ell \geq 1$ and $t, r \geq 0$, the limit

$$Z_\ell(t,r) \doteq \lim_{N\to\infty} \overline{Z}_\ell^{(N)}(t,r), \tag{14}$$

exists and is the unique solution of the hydrodynamic PDE (6)-(9) associated to $(\lambda, Z_\ell^0, \ell \geq 1)$. if Assumption III.b also holds and

$$\sum_{\ell\geq 1} \sup_N \mathbb{P}\{X^{(N),i}(0) \geq \ell\} < \infty, \tag{15}$$

then for every $t \geq 0$ we have

$$\sum_{\ell\geq 1} \sup_N \mathbb{E}\left[\overline{Z}_\ell^{(N)}(t,0)\right] < \infty. \tag{16}$$

4.3 Typical Queue Length Distribution

As stated below, Theorem 4.2 also allows us to characterize the distribution of a typical queue in the network and establish an asymptotic independence result.

Theorem 4.3. Suppose λ, $\overline{G}$, Z_ℓ^0 are as in Theorem 4.2, and let $(Z_\ell; \ell \geq 1)$ be the unique solution to the hydrodynamic PDE associated to $(\lambda, Z_\ell^0; \ell \geq 1)$. Then, for every $\ell \geq 1$ and $t \geq 0$,

$$\lim_{N\to\infty} \mathbb{P}\left\{X^{(N),1}(t) \geq \ell\right\} = Z_\ell(t,0). \tag{17}$$

and the queue length of different servers are asymptotically independent, that is, for every $t \geq 0$, $K \geq 1$ and $\ell_1, ..., \ell_K \geq 1$,

$$\lim_{N\to\infty} \mathbb{P}\left\{X^{(N),1}(t) \geq \ell_1, ..., X^{(N),K}(t) \geq \ell_K\right\} = \prod_{k=1}^{K} Z_{\ell_k}(t,0). \tag{18}$$

4.4 Convergence of Virtual Waiting Times

The virtual waiting time at any time t is the time that a virtual customer that hypothetically arrives at t has to wait in order to receive service. As another application of Theorem 4.2, our next result shows that for large N, the hydrodynamic PDE also provides an approximation to the mean virtual waiting time in an N-server network.

Theorem 4.4. Suppose λ, $\overline{G}$, Z_ℓ^0 are as in Theorem 4.2, and let $(Z_\ell; \ell \geq 1)$ be the unique solution to the hydrodynamic PDE associated to $(\lambda, Z_\ell^0; \ell \geq 1)$. If, in addition, the initial queue lengths satisfy (15) and $a^{(N),i}(0) < T_0$ for some $T_0 < \infty$ and every $i = 1, ..., N$, then

$$\lim_{N\to\infty} \mathbb{E}\left[W^{(N)}(t)\right] = \sum_{\ell\geq 2} Z_\ell(t,0)^2 + \sum_{\ell\geq 1} \left[Z_\ell(t,0) + Z_{\ell+1}(t,0)\right] \int_0^\infty \left[Z_\ell(t,r) - Z_{\ell+1}(t,r)\right] dr. \tag{19}$$

Remark 4.5. The uniform boundedness assumption on the initial ages is reasonable for transient analysis since it will be satisfied by any network that started empty a finite time interval ago. However, the assumption can be relaxed, as illustrated in Figure 3(b) in Section 6.0.3. We impose it only to simplify the proofs.

Remark 4.6. We can show that in fact, as $N \to \infty$, the sequence of virtual waiting time distributions (and not just their means) converge to a limit distribution whose characteristic function can be expressed as a functional of the solution to the hydrodynamic PDE. We do not present the details of the proof, but it is similar to that of Theorem 4.4.

5 Proofs of Main Theorems

Now we prove the main Theorems of the paper stated in Section 4. The uniqueness result of Theorem 4.1 is first proved in Section 5.1. The proof of Theorem 4.2 is rather technical, and requires establishing a limit theorem for a sequence of interactive measure-valued processes describing the N-server network, which is carried out in a companion paper [1]. Theorem 4.3 is proved in Section 5.2. Then in section 5.3, we carry out the calculation to compute the so-called routing probabilities corresponding to the $SQ(d)$ algorithm, which is required for the proof of Theorem 4.4. We have singled out this calculation because it is the only part of the proof which depends on the particular load-balancing algorithm invoked in the network. The proof of Theorem 4.4 is then given in Section 5.4.

5.1 Proof of the Uniqueness Theorem

Throughout this section, we adopt the following notation. For every function f in $[0,\infty)$, that is bounded on finite intervals and every $T \geq 0$, we denote

$$\|f\|_T \doteq \sup_{0\leq t\leq T} |f(t)|.$$

Also, for functions f_1, f_2 on $[0,\infty)$, $f_1 * f_2(t) \doteq \int_0^t f_1(s) f_2(t-s)ds$ is the (one-sided) convolution of f_1 and f_2.

Proof of Theorem 4.1. Fix a non-negative function $\lambda \in \mathbb{L}^1_{\text{loc}}[0,\infty)$ and $Z^0 = (Z^0_\ell; \ell \geq 1)$, and let $Z = (Z_\ell; \ell \geq 1)$ and $\widetilde{Z} = (\widetilde{Z}_\ell; \ell \geq 1)$ both be solutions to the hydrodynamic PDE (6)-(9) associated to (λ, Z^0). Defining the functions

$$D_\ell(t) \doteq -\int_0^t \partial_r Z_\ell(s,0)ds, \qquad t \geq 0,$$

and using integration by parts, we can rewrite

$$\int_0^t \overline{G}(t+r-s)\partial_r Z_{\ell+1}(s,0)ds = -\overline{G}(r)D_{\ell+1}(t) + \int_0^t D_{\ell+1}(s)g(t+r-s)ds. \tag{20}$$

Substituting (20) into (6) and (8), and the definition of D_ℓ into (7) and (9), we see that for every $\ell \geq 1$, Z_ℓ satisfies for $t, r \geq 0$,

$$Z_\ell(t,r) = Z^0_\ell(r+t) + \overline{G}(r)D_{\ell+1}(t) + R_\ell(t,r) - \int_0^t D_{\ell+1}(s)g(t+r-s)ds$$

with boundary condition

$$Z_\ell(t,0) = Z^0_\ell(0) - D_\ell(t) + D_{\ell+1}(t) + \Lambda_\ell(t),$$

where

$$R_\ell(t) = \begin{cases} \int_0^t \lambda(s)\overline{G}(t+r-s)\big(1 - Z^2_1(s,0)\big)ds & \text{if } \ell = 1, \\ \int_0^t \lambda(s)\big(Z_{\ell-1}(s,0) - Z_\ell(s,0)\big)\big(Z_{\ell-1}(s,t+r-s) - Z_\ell(s,t+r-s)\big)ds & \text{if } \ell \geq 2, \end{cases} \tag{21}$$

and

$$\Lambda_\ell(t) \doteq \begin{cases} \int_0^t \lambda(s)\big(1 - Z^2_1(s,0)\big)ds & \text{if } \ell = 1, \\ \int_0^t \lambda(s)\big(Z^2_{\ell-1}(s,0) - Z^2_\ell(s,0)\big)ds & \text{if } \ell \geq 2. \end{cases} \tag{22}$$

Similarly, $\widetilde{Z}$ satisfies analogous equations. Defining $\Delta H_\ell \doteq H_\ell - \widetilde{H}_\ell$ for $H = Z, D, R, \Lambda$ and $\ell \geq 1$, for all $\ell \geq 1$ and $t, r \geq 0$ we have

$$\Delta Z_\ell(t,r) = \overline{G}(r)\Delta D_{\ell+1}(t) + \Delta R_\ell(t,r) - \int_0^t \Delta D_{\ell+1}(s) g(t+r-s)ds, \tag{23}$$

with boundary condition

$$\Delta Z_\ell(t,0) = -\Delta D_\ell(t) + \Delta D_{\ell+1}(t) + \Delta\Lambda_\ell(t). \tag{24}$$

Now for every $t \geq 0$ and $\ell \geq 1$, define

$$V_\ell(t) \doteq \sup_{r\geq 0} |\Delta Z_\ell(t,r)|. \tag{25}$$

Note that V_ℓ is measurable because it is the supremum over measurable functions, and is bounded by 2 because $Z_\ell(t,\cdot)$ and $\widetilde{Z}_\ell(t,\cdot)$ are both in $\mathbb{C}^1_{\leq 1}[0,\infty)$. By definition (21) of R_1, we have,

$$\Delta R_1(t,r) = -\int_0^t \lambda(s)\overline{G}(t+r-s)\big(Z_1(s,0) + \widetilde{Z}_1(s,0)\big)\Delta Z_1(s,0)ds,$$

and hence,

$$|\Delta R_1(t,r)| \leq 2\int_0^t \lambda(s)V_1(s)ds, \qquad \forall t, r \geq 0. \tag{26}$$

Similarly for $\ell \geq 2$, by definition (21) of R_ℓ,

$$\begin{aligned}\Delta R_\ell(t,r) = &\int_0^t \lambda(s)\big(Z_{\ell-1}(s,0) + Z_\ell(s,0)\big)\big(\Delta Z_{\ell-1}(s,t+r-s) - \Delta Z_\ell(s,t+r-s)\big)ds \\ &+ \int_0^t \lambda(s)\big(\Delta Z_{\ell-1}(s,0) + \Delta Z_\ell(s,0)\big)\big(\widetilde{Z}_{\ell-1}(s,t+r-s) - \widetilde{Z}_1(s,t+r-s)\big)ds,\end{aligned}$$

and hence, for all $t, r \geq 0$

$$|\Delta R_\ell(t,r)| \leq 4\int_0^t \lambda(s)\big(V_{\ell-1}(s) + V_\ell(s)\big)ds. \tag{27}$$

Furthermore, to bound ΔD_ℓ for $\ell \geq 2$, we substitute $\Delta D_{\ell+1}$ from (24) in (23) and set $r = 0$ to conclude that ΔD_ℓ satisfies the renewal equation

$$\Delta D_\ell(t) = g * \Delta D_\ell(t) + F_\ell(t), \qquad t \geq 0,$$

with

$$F_\ell(t) \doteq \Delta\Lambda_\ell(t) - g * \Delta\Lambda_\ell(t) + g * \Delta Z_\ell(t,0) - \Delta R_\ell(t,0).$$

Fix $\ell \geq 2$ and note that by definition (22) of Λ_ℓ,

$$\Delta\Lambda_\ell(t) = \int_0^t \lambda(s)\big(Z_{\ell-1}(s,0) + \widetilde{Z}_{\ell-1}(s,0)\big)\Delta Z_{\ell-1}(s,0)ds - \int_0^t \lambda(s)\big(Z_\ell(s,0) + \widetilde{Z}_\ell(s,0)\big)\Delta Z_\ell(s,0)ds,$$

and hence,

$$|\Delta\Lambda_\ell(t)| \leq 4\int_0^t \lambda(s)\big(V_{\ell-1}(s) + V_\ell(s)\big)ds, \qquad \forall t \geq 0. \tag{28}$$

Recall that $V_\ell(t) \leq 2$ for all $t \geq 0$ and $\ell \geq 1$, and due to the local integrability of λ, (28) implies that $\Delta\Lambda_\ell(t)$ is uniformly bounded on any finite interval $t \in [0,T]$. Hence,

$$g * \Delta\Lambda_\ell(t) \leq \|\Delta\Lambda\|_T \int_0^t g(s)ds < \infty, \qquad t \in [0,T].$$

Similarly, the bound (27) shows that $\Delta R_\ell(t)$ is also bounded for finite t. Therefore, F_ℓ is bounded on finite intervals, and hence by the renewal theorem (see Theorem V.2.4 in [2])

$$\Delta D_\ell(t) = \int_0^t F_\ell(t-s) dU_G(s),$$

where U_G is the renewal measure corresponding to the service distribution G. Since G has a density g, U_G satisfies

$$U_G(x) = 1 + \int_0^x u_G(y) dy, \qquad x \geq 0,$$

and the density u_G is bounded on finite intervals and satisfies the equation $u_G = g * u_G + g$ due to Proposition V.2.7 in [2]. Therefore, $\Delta D_\ell(t)$ can be written as

$$\Delta D_\ell(t) = F_\ell(t) + u_G * F_\ell(t) = \Delta\Lambda_\ell(t) + u_G * \Delta Z_\ell(t,0) - \Delta R_\ell(t,0) - u_G * \Delta R_\ell(t,0). \tag{29}$$

For every fixed $T \geq 0$ and all $0 \leq t \leq T$, by the definition of the convolution operator and (25),

$$|u_G * \Delta Z_\ell(t,0)| \leq \int_0^t u_G(s) |\Delta Z_\ell(t-s,0)| ds \leq \|u_G\|_T \int_0^t V_\ell(s) ds. \tag{30}$$

Also, using the bound (27) we have

$$\begin{aligned} |u_G * \Delta R_\ell(t,0)| \leq & 4 \int_0^t u_G(t-s) \int_0^s \lambda(v) \big(V_{\ell-1}(v) + V_\ell(v)\big) dv\, ds \\ = & 4 \int_0^t \lambda(v) \big(V_{\ell-1}(v) + V_\ell(v)\big) \int_v^t u_G(t-s) ds\, dv \\ \leq & 4 U_G(T) \int_0^t \lambda(s) \big(V_{\ell-1}(s) + V_\ell(s)\big) ds. \end{aligned} \tag{31}$$

Bounding the terms on the right-hand side of (29) using inequalities (27), (28), (30) and (31), for every $\ell \geq 2$ and $t \leq T$, we have

$$\|\Delta D_\ell\|_t \leq C_T \int_0^t (1 + \lambda(s))(V_{\ell-1}(s) + V_\ell(s)) ds, \tag{32}$$

with $C_T \doteq 8 + 4U_G(T) + \|u_G\|_T$. Next, substituting the bound (32), but with ℓ replaced by $\ell + 1$, (26) and (27) into (23), we obtain that for all $0 \leq t \leq T$ and $r \geq 0$,

$$|\Delta Z_1(t,r)| \leq 4(C_T \overline{G}(r) + 1) \int_0^t (1 + \lambda(s))(V_1(s) + V_2(s)) ds, \tag{33}$$

and for $\ell \geq 2$,

$$|\Delta Z_\ell(t,r)| \leq 4(C_T \overline{G}(r) + 1) \int_0^t (1 + \lambda(s))(V_{\ell-1} + V_\ell(s) + V_{\ell+1}(s)) ds. \tag{34}$$

Taking supremum over $r \geq 0$ on both sides of (33) and (34), we have

$$V_1(t) \leq 8 C_T \int_0^t (1 + \lambda(s))(V_1(s) + V_2(s)) ds, \tag{35}$$

and for $\ell \geq 2$,

$$V_\ell(t) \leq 8 C_T \int_0^t (1 + \lambda(s))(V_{\ell-1}(s) + V_\ell(s) + V_{\ell+1}(s)) ds. \tag{36}$$

Now define

$$V(t) \doteq \sum_{\ell \geq 1} 2^{-\ell} V_\ell(s). \tag{37}$$

Note that V is measurable, and $V(t) = 0$ if and only if $V_\ell(t) = 0$ for all $\ell \geq 1$. Considering the weighted sums of both sides of (35) and (36) over $\ell \geq 2$, we obtain

$$V(t) \leq 26 C_T \int_0^t (1 + \lambda(s)) V(s) ds, \qquad 0 \leq t \leq T. \tag{38}$$

An application of Gronwall's inequality shows that $V \equiv 0$, and hence, $Z_\ell = \widetilde{Z}_\ell$ for all $\ell \geq 0$. This completes the proof. □

5.2 Proof of Theorem 4.3

Proof of Theorem 4.3. First note that since the routing algorithm is symmetric with respect to queue indices, the queue lengths and age distributions, which are initially exchangeable by III.b, remain exchangeable at all finite times $t \geq 0$. In particular, the distribution of the vector $(X^{(N),\sigma(i)}(t); i = 1, ..., N)$ is the same for every permutation σ on $\{1, 2, ..., N\}$. Since $Z_\ell^{(N)}(t, 0) = S_\ell^{(N)}(t)$ and by definition (3) of $S_\ell^{(N)}(t)$, we have

$$\mathbb{E}\left[\overline{Z}_\ell^{(N)}(t, 0)\right] = \frac{1}{N}\mathbb{E}\left[\sum_{i=1}^N \mathbb{1}\left(X^{(N),i}(t) \geq \ell\right)\right] = \frac{1}{N}\sum_{i=1}^N \mathbb{P}\left\{X^{(N),i}(t) \geq \ell\right\} = \mathbb{P}\left\{X^{(N),1}(t) \geq \ell\right\}, \tag{39}$$

where the last equality is due to exchangeability. By Theorem 4.2, $\overline{Z}_\ell^{(N)}(t, 0)$ converges to $Z_\ell(t, 0)$ as $N \to \infty$, almost surely, and since $\overline{Z}_\ell^{(N)}(t, 0) = \overline{S}_\ell^{(N)}(t)$ is bounded by 1, the convergence also holds in expectation by the bounded convergence theorem. Therefore, (17) follows on taking the limit as $N \to \infty$ of both sides of (39).

Similarly, for $m = 1, \ldots, n$, since $Z_{\ell_m}^{(N)}(t, 0) = S_{\ell_m}^{(N)}(t)$ by (5) and $S_\ell^{(N)}(t)$ is defined by (3), we have

$$\begin{aligned}
\mathbb{E}\left[\prod_{m=1}^n \overline{Z}_{\ell_m}^{(N)}(t, 0)\right] &= \frac{1}{N^n}\mathbb{E}\left[\sum_{i_1=1}^N ... \sum_{i_n=1}^N \mathbb{1}\left(X^{(N),i_1}(t) \geq \ell_1\right) ... \mathbb{1}\left(X^{(N),i_n}(t) \geq \ell_n\right)\right] \\
&= \frac{1}{N^n}\sum_{i_1=1}^N ... \sum_{i_n=1}^N \mathbb{P}\left\{X^{(N),i_1}(t) \geq \ell_1, X^{(N),i_n}(t) \geq \ell_n\right\} \\
&= \mathbb{P}\left\{X^{(N),1}(t) \geq \ell_1, ..., X^{(N),n}(t) \geq \ell_n\right\},
\end{aligned} \tag{40}$$

where the last equality is again due to exchangeability. By another use of Theorem 4.2, $\prod_{m=1}^n Z_{\ell_m}^{(N)}(t, 0)$, converges to $\prod_{m=1}^n Z_{\ell_m}(t, 0)$ as $N \to \infty$, almost surely, and in expectation, using the bounded convergence theorem. Taking the limit as $N \to \infty$ of both sides of (40), we obtain (18). □

5.3 Routing Probabilities

For every server index i and queue length ℓ, define the routing probability $p(i, \ell; t)$ to be the conditional probability, given the state of the network at time t, that the load-balancing algorithm would route a hypothetical job arriving at time t to server i, when its queue length is ℓ. Defining the vector $\mathbf{X}^{(N)}(t)$ of all queue lengths and ages,

$$\mathbf{X}^{(N)}(t) \doteq \left(X^{(N),i}(t), a^{(N),i}(t); i = 1, ..., N\right). \tag{41}$$

and denoting by $\kappa(t)$ the (random) index to which the virtual job arriving at time t is routed, the routing probability $p(i,\ell;t)$ is defined by

$$p(i,\ell,t) \doteq \mathbb{1}\left(X^{(N),i}(t)=\ell\right)\mathbb{P}\left\{\kappa(t)=i|\mathbf{X}^{(N)}(t)\right\}. \tag{42}$$

Now we compute the routing probabilities $p(i,\ell;t)$ for $\ell \geq 1$ and under the $SQ(2)$ algorithm described in Section 2.1. If κ_1 and κ_2 are indices of queues chosen independently and uniformly at random, then $\kappa(t)$ is the index associated with the shorter queue (with ties being broken uniformly at random). The job is routed to a server of queue length exactly ℓ if and only if both κ_1 and κ_2 have queue lengths at least ℓ, and at least one of them has queue length ℓ. Given the vector of all queue lengths in $\mathbf{X}^{(N)}(t)$, this happens with probability $(\overline{S}_\ell^{(N)}(t))^2-(\overline{S}_{\ell+1}^{(N)}(t))^2$. Since all servers with queue length equal to ℓ are equally likely to be chosen and there are $S_\ell^{(N)}(t)-S_{\ell+1}^{(N)}(t)$ of them, for $\ell \geq 1$ we have

$$p(i,\ell;t)=\mathbb{1}\left(X^{(N),i}(t)=\ell\right)\frac{1}{N}\left(\overline{S}_\ell^{(N)}(t)+\overline{S}_{\ell+1}^{(N)}(t)\right) \tag{43}$$

Remark 5.1. Note that the form of the routing probability for the $SQ(d)$ algorithm is clearly the same as in the exponential case. To apply our framework to other load balancing algorithms, one would have to replace the expression in (43) with the routing probabilities associated with that algorithm.

5.4 Proof of Theorem 4.4

Proof of Theorem 4.4. Suppose that the virtual job arriving at time t is routed to queue i, that is $\kappa(t)=i$. If the server i is idle (i.e., $X^{(N),i}(t)=0$), the virtual waiting time is zero; otherwise if $X^{(N),i}(t)=\ell$ for some $\ell \geq 1$, the virtual waiting time $W^{(N)}(t)$ is the sum of service times v_j of jobs waiting in queue i plus the residual time $b^{(N),i}(t)$ of the job in service at server i at time t, that is,

$$\mathbb{1}\left(\kappa(t)=i\right)W^{(N)}(t)=\sum_{j=1}^{\ell-1}v_j+b^{(N),i}(t).$$

Summing over all possible queue indices i and queue lengths ℓ, we have

$$W^{(N)}(t)=W_1^{(N)}(t)+W_2^{(N)}(t), \tag{44}$$

where

$$W_1^{(N)}(t)\doteq\sum_{i=1}^{N}\sum_{\ell\geq 1}\mathbb{1}\left(\kappa(t)=i,X^{(N),i}(t)=\ell\right)\sum_{j=1}^{\ell-1}v_j, \tag{45}$$

and

$$W_2^{(N)}(t)\doteq\sum_{i=1}^{N}\sum_{\ell\geq 1}\mathbb{1}\left(\kappa(t)=i,X^{(N),i}(t)=\ell\right)b^{(N),i}(t). \tag{46}$$

To compute the expectation of $W_1^{(N)}(t)$, note that by Assumption II, the service times v_j of jobs that are still waiting in queues satisfy $\mathbb{E}[v_j]=1$ and are independent of all queue lengths and ages at time t, as well as $\kappa(t)$. Therefore, taking the conditional expectation given $\mathbf{X}^{(N)}(t)$ of both sides of (45) and using the definition (42) of $p(i,\ell,t)$, we have

$$\begin{aligned}\mathbb{E}\left[W_1^{(N)}(t)|\mathbf{X}^{(N)}(t)\right]&=\sum_{i=1}^{N}\sum_{\ell\geq 1}\mathbb{P}\left\{\kappa(t)=i,X^{(N),i}(t)=\ell|\mathbf{X}^{(N)}(t)\right\}\sum_{j=1}^{\ell-1}\mathbb{E}\left[v_j\right]\\&=\sum_{\ell\geq 1}(\ell-1)\sum_{i=1}^{N}\mathbb{1}\left(X^{(N),i}(t)=\ell\right)p(i,\ell;t).\end{aligned}$$

Taking expectations of both sides of the last equation, substituting $p(i,\ell;t)$ from (43), recalling that the number of servers with queue length equal to ℓ is $S_\ell^{(N)}(t) - S_{\ell+1}^{(N)}(t)$, and using the equality $S_\ell^{(N)}(t) = Z_\ell^{(N)}(t,0)$ from (5), we see that

$$\begin{aligned}
\mathbb{E}[W_1^{(N)}(t)] &= \frac{1}{N}\mathbb{E}\left[\sum_{\ell\geq 1}(\ell-1)\left(\overline{S}_\ell^{(N)}(t)+\overline{S}_{\ell+1}^{(N)}(t)\right)\sum_{i=1}^{N}\mathbb{1}\left(X^{(N),i}(t)=\ell\right)\right] \\
&= \frac{1}{N}\sum_{\ell\geq 1}(\ell-1)\mathbb{E}\left[\left(\overline{S}_\ell^{(N)}(t)+\overline{S}_{\ell+1}^{(N)}(t)\right)\left(S_\ell^{(N)}(t)-S_{\ell+1}^{(N)}(t)\right)\right] \\
&= \sum_{\ell\geq 1}(\ell-1)\mathbb{E}\left[\left(\overline{Z}_\ell^{(N)}(t,0)\right)^2-\left(\overline{Z}_{\ell+1}^{(N)}(t,0)\right)^2\right] \\
&= \sum_{\ell\geq 2}\mathbb{E}\left[\left(\overline{Z}_\ell^{(N)}(t,0)\right)^2\right]. \qquad (47)
\end{aligned}$$

To compute the expectation of $W_2^{(N)}(t)$, note that according to the $SQ(2)$ algorithm, the random queue indices κ_1 and κ_2 are independent of all other random variables, and therefore, given $\mathbf{X}^{(N)}(t)$, $\kappa(t)$ is independent of all residual service times. Therefore, taking conditional expectations of both sides of (46) and invoking the definition (42) of $p(i,\ell;t)$ again, we have

$$\mathbb{E}\left[W_2^{(N)}(t)|\mathbf{X}^{(N)}(t)\right] = \sum_{\ell\geq 1}\sum_{i=1}^{N}\mathbb{1}\left(X^{(N),i}(t)=\ell\right)p(i,\ell;t)\mathbb{E}\left[b^{(N),i}(t)|\mathbf{X}^{(N)}(t)\right]. \qquad (48)$$

The residual service time $b^{(N),i}(t)$ of the job that begins being processed by server i at time t satisfies

$$b^{(N),i}(t) = v_{J(i;t)} - a^{(N),i}(t),$$

where $J(i;t)$ is the index of the job begin processes in server i at time t. It follows from the i.i.d assumption on the service times and their independence from the arrival process (Assumption II) that the residual service time $b^{(N),i}(t)$ depends on the state variable $\mathbf{X}^{(N)}(t)$ only through the age $a^{(N),i}(t)$. A complete rigorous justification of this intuitive assertion is rather long and technical, and hence will be presented elsewhere. Hence using equation (3.2) of Section V.3 in [2] in the second equality below, for every $r \geq 0$, we have

$$\mathbb{P}\left\{b^{(N),i}(t) > r|\mathbf{X}^{(N)}(t)\right\} = \mathbb{P}\left\{b^{(N),i}(t) > r|a^{(N),i}(t)\right\} = \frac{\overline{G}(a^{(N),i}(t)+r)}{\overline{G}(a^{(N),i}(t))},$$

and therefore,

$$\mathbb{E}\left[b^{(N),i}(t)|\mathbf{X}^{(N)}(t)\right] = \int_0^\infty \frac{\overline{G}(a^{(N),i}(t)+r)}{\overline{G}(a^{(N),i}(t))}\,dr. \qquad (49)$$

Now, substituting equations (43) and (49) in (48), taking expectations of both sides, and using definition (4) of $Z_\ell^{(N)}$ and equation (5), we can see that

$$\begin{aligned}
\mathbb{E}[W_2^{(N)}(t)] &= \sum_{\ell\geq 1}\frac{1}{N}\mathbb{E}\left[\left(\overline{S}_\ell^{(N)}(t)+\overline{S}_{\ell+1}^{(N)}(t)\right)\int_0^\infty\sum_{i=1}^{N}\mathbb{1}\left(X^{(N),i}(t)=\ell\right)\frac{\overline{G}(a^{(N),i}(t)+r)}{\overline{G}(a^{(N),i}(t))}dx\right] \\
&= \sum_{\ell\geq 1}\mathbb{E}\left[\left(\overline{Z}_\ell^{(N)}(t,0)+\overline{Z}_{\ell+1}^{(N)}(t,0)\right)\int_0^\infty\left(\overline{Z}_\ell^{(N)}(t,r)-\overline{Z}_{\ell+1}^{(N)}(t,r)\right)dr\right]. \qquad (50)
\end{aligned}$$

Finally, taking the limit as $N \to \infty$ in (50), we obtain

$$\lim_{N\to\infty} \mathbb{E}\left[W_1^{(N)}(t)\right] = \sum_{\ell\geq 2} \lim_{N\to\infty} \mathbb{E}\left[\overline{Z}_\ell^{(N)}(t,0)^2\right] = \sum_{\ell\geq 2} \left(Z_\ell(t,0)\right)^2. \tag{51}$$

The exchange of limit and summation in the first equality is justified by the bound (16), $\overline{Z}_\ell^{(N)}(t,0) \leq 1$ and the dominated convergence theorem, while the second equality follows from Theorem 4.2 and the bounded convergence theorem. Moreover, by the uniform boundedness assumption on the ages imposed in Theorem (4.4), $a^{(N),i}(t) < T_0 + t$. Therefore, since $\int_0^\infty \overline{G}(r)dr = 1$ because the service time has mean 1, and by definition (4) of $Z_\ell^{(N)}$,

$$\int_0^t \overline{Z}_\ell^{(N)}(t,r)dr \leq \frac{1}{N}\int_0^t \sum_{i\in\mathcal{S}_\ell^{(N)}(t)} \frac{\overline{G}(r)}{\overline{G}(T_0+t)}dr \leq \frac{1}{\overline{G}(T_0+t)}\overline{S}_\ell^{(N)}(t)\int_0^t \overline{G}(r)dr \leq \frac{1}{\overline{G}(T_0+t)}. \tag{52}$$

Therefore, taking the limit as $N \to \infty$ of both sides of (50), we have

$$\begin{aligned}\lim_{N\to\infty} \mathbb{E}\left[W_2^{(N)}(t)\right] &= \sum_{\ell\geq 1} \lim_{N\to\infty} \mathbb{E}\left[\left(\overline{Z}_\ell^{(N)}(t,0) + \overline{Z}_{\ell+1}^{(N)}(t,0)\right)\int_0^\infty \left(\overline{Z}_\ell^{(N)}(t,r) - \overline{Z}_{\ell+1}^{(N)}(t,r)\right)dr\right] \\ &= \sum_{\ell\geq 1} \mathbb{E}\left[\left(Z_\ell(t,0) + Z_{\ell+1}(t,0)\right)\int_0^\infty \left(Z_\ell(t,r) - Z_{\ell+1}(t,r)\right)dr\right].\end{aligned} \tag{53}$$

The exchange of limit and summation in the first equality is justified by (16), $\overline{Z}_\ell^{(N)}(t,r) \leq \overline{Z}_\ell^{(N)}(t,0) \leq 1$ and the dominated convergence theorem, while the second equality holds due to Theorem 4.2, (52) and the bounded convergence theorem. The result follows from (51) and (53). □

In this section we validate the results in Theorem 4.3 and Theorem 4.4 using Monte Carlo (MC) simulations. First in Section 5.5, we present a stable scheme for numerically approximating the solution to the hydrodynamic PDE. Then in Section 5.6, we compare the results obtained from Monte Carlo simulation and numerical solutions to PDE for a variety of different service distributions and in each case, we observe a close match.

5.5 A Numerical Approximation Scheme

Note that the hydrodynamic PDE is comprised of a countable number of integro-differential equations, corresponding to each Z_ℓ. Therefore, in order to numerically solve the hydrodynamic PDE (6)-(9), we truncate the state space and only solve for $Z_\ell(t,r)$, $\ell = 1,\ldots,L_0, 0 \leq t \leq T_0, 0 \leq r \leq R_0$, for suitable $L_0 \in \mathbb{N}$ and $R_0, T_0 < \infty$. Next, we discretize r and t on uniform meshes: $\hat{Z}_\ell(t_m, r_n) \doteq Z_\ell(m\delta, n\delta)$ with $0 \leq m \leq \lfloor T_0/\delta \rfloor$ and $0 \leq n \leq \lfloor R_0/\delta \rfloor$. Then we solve the equations (6)-(9) numerically by the finite difference method and the explicit forward Euler scheme[14]. That is, for fixed $t_m > 0$ and $r_n > 0$, we apply the following approximations

$$\begin{aligned}\hat{Z}_1(t_m+\delta, r_n) &= \hat{Z}_1(t_m, r_n+\delta) - \int_{t_m}^{t_m+\delta} \overline{G}(t_m+\delta+r_n-u)\partial_r \hat{Z}_2(u,0)du \\ &\quad + \int_t^{t+\delta} \lambda(u)\overline{G}(t_m+\delta+r_n-u)\left(1-\hat{Z}_1(u,0)^2\right)du \\ &\approx \hat{Z}_1(t_m, r_n+\delta) - \overline{G}(r_n)\left(\hat{Z}_2(t_m,\delta) - \hat{Z}_2(t_m,0)\right) + \overline{G}(r_n)\left(1-\hat{Z}_1(t_m,0)^2\right)\int_t^{t+\delta}\lambda(u)du.\end{aligned}$$

and for $\ell \geq 2$,

$$\begin{aligned}\hat{Z}_\ell(t_m+\delta, r_n) &= \hat{Z}_\ell(t_m, r_n+\delta) - \int_{t_m}^{t_m+\delta} \overline{G}(t_m+\delta+r_n-u)\partial_r \hat{Z}_{\ell+1}(u,0)du \\ &\quad + \int_{t_m}^{t_m+\delta} \lambda(u)\left(\hat{Z}_{\ell-1}(u,0)+\hat{Z}_\ell(u,0)\right)\left(\hat{Z}_{\ell-1}(u, t_m+\delta+r_n-u) - \hat{Z}_\ell(u, t_m+\delta+r_n-u)\right)du \\ &\approx \hat{Z}_\ell(t_m, r_n+\delta) - \overline{G}(r_n)\left(\hat{Z}_{\ell+1}(t_m,\delta) - \hat{Z}_{\ell+1}(t_m,0)\right) \\ &\quad + \left(\hat{Z}_{\ell-1}(t_m,0)+\hat{Z}_\ell(t_m,0)\right)\left(\hat{Z}_{\ell-1}(t_m,r_n) - \hat{Z}_\ell(t_m,r_n)\right)\int_{t_m}^{t_m+\delta}\lambda(u)du.\end{aligned}$$

The numerical scheme above has been stabilized by applying the upwind scheme [9] to the derivatives with respect to r, otherwise, the numerics will blow up in a short time interval. Using these approximations, we can update $\hat{Z}_\ell\left(t_m+\delta,\cdot\right)$ from $\hat{Z}_\ell(t_m,\cdot)$.

5.6 Validation of Results

We now compare the empirical queue length distribution and mean virtual waiting time obtained from the Monte Carlo simulation of an N-server network, with the corresponding limit quantities (17) and (19), as predicted by the numerical approximation of the PDE.

We consider a sequence of networks indexed by the number of servers N, with a Poisson arrival process with rate $\lambda = 0.5$, and the following initial conditions. Each server has initially one job with initial age equal to zero, that is $X^{(N),i}(0) = 1$ and $a^{(N),i}(0) = 0$ for all $i = 1, ..., N$. Note that this sequence of initial conditions satisfies the conditions of Theorem 4.2, and converges to the initial condition $(Z_\ell^0(\cdot); \ell \geq 1)$ for the hydrodynamic PDE, where for $r \geq 0$,

$$Z_1^0(r) = \overline{G}(r), \qquad Z_\ell^0(r) = 0, \quad \ell \geq 2.$$

5.6.1 Queue Length Distribution

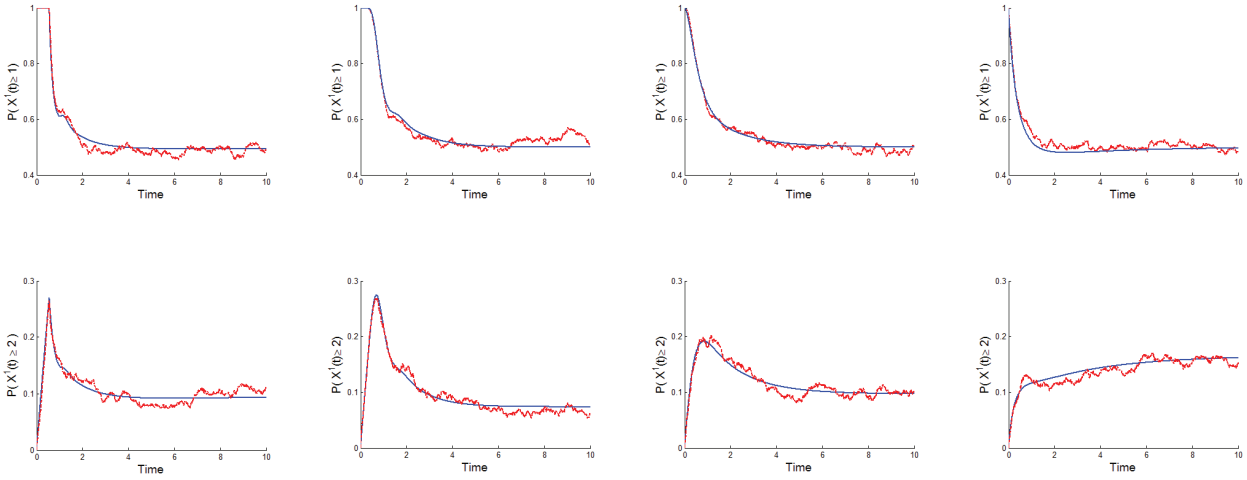
6 Simulation Results

First we compute the probability that a typical (fixed) queue has length at least ℓ at time t, for $t \in [0, 10]$ in a network of $N = 1000$ servers, using Monte Carlo simulation with 1000 realizations. Then we compare this probability with the quantity obtained by numerically solving the PDE using the method described in Section 5.5, with $L_0 = 6$, $R_0 = 20$ and $\delta = 0.001$. We make this comparison for a variety of (unit mean) service time distribution, including Pareto (with shape parameter $\beta = 2.25$), Log-Normal (with shape parameter $\sigma = 0.33$), Gamma (with shape parameter $k = 2$) and Hyper-Exponential (with parameters $\lambda_1 = 0.5$ and $\lambda_2 = 2$), which is a special case of a Phase-type distribution.

The results are illustrated in Figure 2 for $\ell = 1$ and $\ell = 2$, and a close match is observed in all cases. In our setting, the run time for the Monte Carlo method is approximately $2 - 3$ hours, which is orders of magnitude longer than the time taken to numerically approximate the PDE, which is around $7 - 9$ seconds.

6.0.2 Waiting Time

Next, we compare the mean virtual waiting time in the same setting described above, but with the arrival rate now set to $\lambda = 0.7$. We measure the mean virtual waiting from a Monte Carlo simulation as follows: at any time t, we determine which server a virtual arriving job would have been routed to



(a) Pareto (b) Log-Normal (c) Gamma (d) Hyper-Exponential

Figure 2: Comparison of the estimate for $P(X^{(N),1}(t) \geq \ell)$ obtained from MC simulation (in red) versus the numerical approximation of $Z_\ell(t,0)$ (in blue) during $t \in [0,10]$ for $\ell = 1$ (top row) and $\ell = 2$ (bottom row).

by choosing two servers uniformly at random, and picking the one with shorter queue length, and then observe the waiting time in that server. The average is taken over 2000 realizations. We compare this to the following approximation of the limit provided by Theorem 4.4:

$$\mathbb{E}\left[W^{(N)}(t)\right] \approx \sum_{\ell=2}^{L_0} \hat{Z}_\ell(t,0)^2 + \sum_{\ell=1}^{L_0-1} \left[\hat{Z}_\ell(t,0) + \hat{Z}_{\ell+1}(t,0)\right] \sum_{j=0}^{\lfloor R_0/\delta \rfloor} \left[\hat{Z}_\ell(t,r_j) - \hat{Z}_{\ell+1}(t,r_j)\right] \delta. \tag{54}$$

where $\{\hat{Z}_\ell; \ell \geq 1\}$ is the numerical solution of the PDE described in Section 5.5.

The result of the comparison for the Pareto distribution (with mean set to 1 and shape parameter $\beta = 3$) and is plotted in Figure 3(a). We observe good agreement between the two curves. Furthermore, recall that the actual waiting time of a job is the difference between the arrival and service entry times of that job. We also plot the average of the average of the actual waiting time. At each time t, the latter quantity is defined to be the sum of the waiting times of all jobs arrived in that time slot in the mesh, divided by the number of jobs that arrived in that time slot. We observe that the mean virtual waiting time is a good approximation to the average actual waiting time as well.

6.0.3 More General Initial Conditions

Finally, to illustrate that the assumption $a^{(N),i}(0) \leq T_0$ on initial ages is not necessary, we validate our result for another sequence of networks with the following initial condition: there are initially two jobs in each queue, and the initial ages are all independent and distributed according to $\overline{G}(x)dx$ (which is the stationary age distribution in a renewal process with inter-arrival distribution G). The corresponding initial condition for the PDE are given by

$$Z_\ell^0(r) = \int_r^\infty \overline{G}(x)dx, \quad \ell = 1,2, \qquad Z_\ell^0(r) = 0, \quad \ell \geq 3.$$

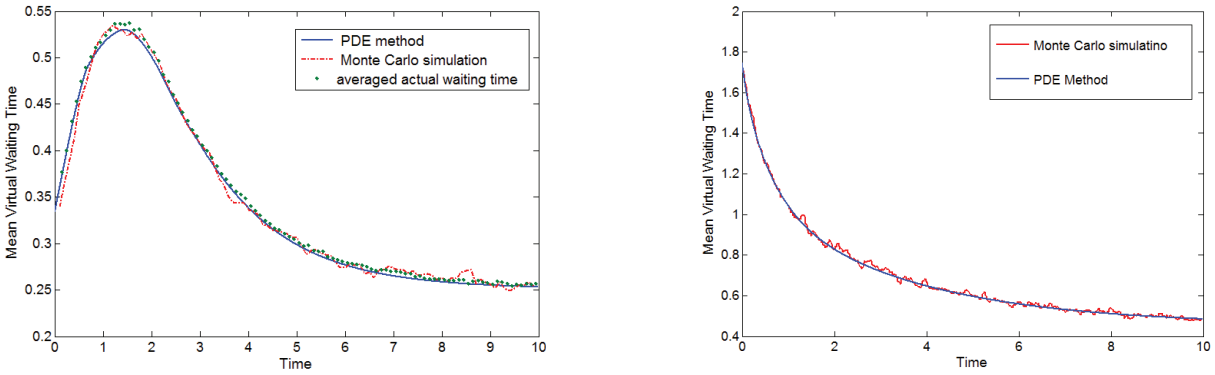


Figure 3: **a.** Mean virtual sojourn times from MC (red) and PDE (blue) as well as the averaged actual sojourn times from MC (green) for the Pareto distribution with $\beta = 3$. **b.** Mean virtual waiting times from MC (red) and PDE (blue) for Gamma service distribution and unbounded initial ages.

We validate the approximation given in (54) using Monte Carlo simulations for a network with $N = 1000$ servers and a Gamma service time distribution (with shape parameter $k = 2$). The result is depicted in Figure 3(b).

7 Engineering Insights

In practice, transient Quality of Service (QoS) parameters are of particular interest in many applications. However, to the best of our knowledge, prior to this work, load-balancing networks with general service distributions have only been studied in steady state and under constant Poisson arrivals [6]. We illustrate how our PDE approximation can be used to shed insight into transient phenomena of practical relevance. In Section 7.1, we study the time it takes for a congested network to get rid of a backlog of jobs, and in Section 7.2 we analyze the performance of a load balancing network with time-varying, periodic, Poisson arrivals, consisting of intermittent "high" and "low" periods.

7.1 Initial Backlog

Given a network that is congested with a backlog of jobs, the system administrator would like to know how long it would take for the system to get rid of the backlog and for the QoS to return closer to the normal operating point. In this section, we consider a network with a large initial backlog, and hence a large initial virtual waiting time, and study the time it takes for the system to unload to the extent that the mean virtual waiting time reaches half of its initial value, which we refer to as "relaxation time". We investigate the effect of service distribution statistics on the relaxation time. In each case, the goal is to illustrate how the PDE approximation may be used to help uncover interesting (and possibly unexpected) network phenomena.

In the presence of general service distributions, in order to capture congestion in the initial distribution of the network, one has to specify not only the distribution of the number of jobs in the different queues, but also the distribution of ages of jobs being served at these queues. There are a number of different configurations that could reflect a congested initial state. In order to generate initial conditions that may naturally occur in practice, we consider an initial state that corresponds to the distribution resulting from a network that has been experiencing a higher than normal arrival rate for a period of time.

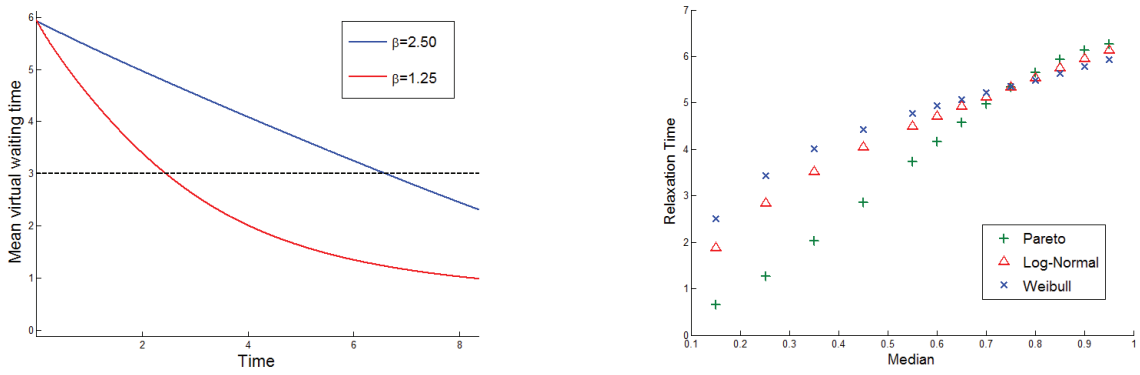


Figure 4: **a.** Mean virtual waiting time for network with initial backlog and Pareto service distribution with $\beta = 1.25$ (red) and $\beta = 2.50$ (blue), using the PDE method. **b.** Relaxation time vs. Median for Pareto, Log-Normal and Weibull service time distributions.

Specifically, we first consider a network that starts empty and runs with nominal arrival rate $\lambda = 0.6$ for $T_0 = 10$ units of time, and then experienced an arrival rate that is about 8 times the nominal arrival rate, namely $\lambda = 5$, for a period of $T_b = 2$ units prior to zero. The distribution of this network at time $T_0 + T_b$ then represents a congested state for the network and is used as the initial condition for our PDE approximations. We then assume that at time 0, the mean arrival rate reverts back to its nominal value $\lambda = 0.6$ and we study the relaxation time, namely the time it takes for the mean virtual waiting time to reach half its initial value.

In Figure 4(a), we plot the evolution in time of the mean virtual waiting time for (nominal) arrival rate $\lambda = 0.6$, both when the service distribution is Pareto with unit mean and parameter $\beta = 1.25$ (heavy-tailed) and Pareto with parameter $\beta = 2.50$ (light-tailed). The curves are obtained using the numerical scheme to solve the PDE with cutoff values $L_0 = 12$ and $R_0 = 20$ and step size $\delta = 0.001$. An interesting observation is that the heavy-tailed case clearly has a smaller relaxation time than the light-tailed case. This phenomenon may seem particularly surprising in light of the steady-state result of Bramson et. al in [6], which shows that for Pareto service time distributions with unit mean, the tail of the limit steady state queue-length distribution has a double exponential decay when $\beta > 2$ (light-tailed), in contrast to the case when $\beta < 2$ (heavy-tailed), when it has only a power law decay. Although the tails of the limit steady-state distribution do not represent the limit of the tails of the steady-state distribution in the N-server system (since the $N \to \infty$ and tail decay limits are typically not interchangeable), the results in [6] suggest that from the point of view of equilibrium queue length, the light-tailed Pareto distribution shows better performance.

The transient phenomenon observed above also persists when the initial condition is instead chosen to be a large number of jobs uniformly distributed amongst different queues, all starting with zero ages, which may roughly correspond to congestion caused due to a sudden large spurt of job arrivals into the network.

A possible heuristic explanation for the contrasting behavior observed in the transient case is that when a Pareto service distribution is fixed to have mean equal to one, the median of the distribution decreases with a decrease in β. As a result, when the heavier the tail of the service distribution, the greater the fraction of initially backlogged (and newly arriving) jobs with smaller service requirements, and the smaller the fraction of jobs with very long service times. This helps servers in a large number of queues to get rid of their backlog and reduce their queue lengths faster, whereas the jobs with long service times lead to large queue lengths in only a small fraction of servers. The latter jobs do not

increase the mean virtual waiting time significantly because a new potential arrival will avoid the few long queues with high probability under the $SQ(2)$ load-balancing policy. This argument suggests that the same trend should hold for different classes of service time distributions. In Figure 4(b), we plot the relaxation times obtained via the PDE approximation for the Pareto, Weibull and Log-Normal service time distributions for different median values, and observe that the relaxation time decreases as the median decreases (and variance increases) for all these distribution families.

7.2 Periodic Arrivals and Effective Arrival Rate

In contrast to prior work, our method also allows us to study the $SQ(d)$ network under time-varying arrivals. In many real-world applications, the arrival process is periodic (over period of a day, for example,) comprised of peak and off-peak periods. In this scenario, we examine the effect of the time-inhomogeneity of the arrival on the performance of the network.

As an illustration, we consider a family of periodic arrival rate functions $\lambda(\cdot) \in L^1_{\text{loc}}(0,\infty)$, with period T, parameterized by the average arrival rate $\overline{\lambda}$ over the period and a constant $\Delta > 0$, which can be viewed as a burstiness parameter. The arrival rate takes the value $\overline{\lambda} + \Delta$ in the first-half of the period and the value $\overline{\lambda} - \Delta$ in the second half of the period, corresponding to peak and off-peak periods. The larger the Δ, the more dramatic the difference between the peak and off-peak arrival rates; the particular case $\Delta = 0$ corresponds to the case of constant arrival rate. Using our PDE approximation, for fixed $\overline{\lambda}$, we study the effect of the burstiness parameter Δ on the mean virtual waiting time of the network.

As one would expect, the burstiness has a negative impact on the network performance as it results in an increase in the mean virtual waiting time. However, the PDE approximation also quantifies this effect, and allows comparisons across different service time distributions. To this end, for a fixed average arrival rate and different values of the burstiness parameter Δ, we compute the mean virtual waiting time averaged over one period, denoted by $\overline{W}(\overline{\lambda}, \Delta)$, and find the *constant rate* Poisson arrival that results in the same averaged virtual waiting time. We denote this constant rate by $\lambda_{\text{eff}} = \lambda_{\text{eff}}(\overline{\lambda}, \Delta)$, and refer to it as the "effective arrival rate" corresponding to $\overline{\lambda}$ and Δ, which hence has the property

$$\overline{W}(\overline{\lambda}, \Delta) = \overline{W}(\lambda_{\text{eff}}, 0).$$

We compare the effect of traffic burstiness on the performance of the network under a heavy tail and a light tail Pareto service distribution. In Figure 5, we plot λ_{eff} as a function of Δ for $\overline{\lambda} = 0.7$ and Pareto service time distributions with shape parameters $\beta = 1.5$ (heavy tail, infinite variance) and $\beta = 3.0$ (light tail, finite variance.) As illustrated in the figure, the burstiness of the arrival rate has a greater effect on the network when the service time has finite variance ($\beta = 3.0$) rather than infinite variance ($\beta = 1.5$). In other words, as far as the average mean virtual waiting time is concerned, the network with the heavy-tailed service time shows less increase in λ_{eff} than the light-tailed service time. This is in line with our observation of backlogged networks that heavy tails seem to have a less deleterious effect on the transient behavior of the network than in equilibrium.

It is worth mentioning that quantitative insights into the network such as those illustrated in Figure 5, may require the solution of inverse problems, such as the computation of λ_{eff} for various values of Δ, which would be incredibly time-consuming, if not infeasible using Monte Carlo simulation.

8 Discussion and Future Work

In this paper, we have introduced a new framework to analyze load-balancing networks that use the $SQ(d)$ algorithm, focusing for simplicity on the $SQ(2)$ algorithm. We introduced the hydrodynamic PDE, which captures the evolution of the scaled state of the network in the limit as the number of

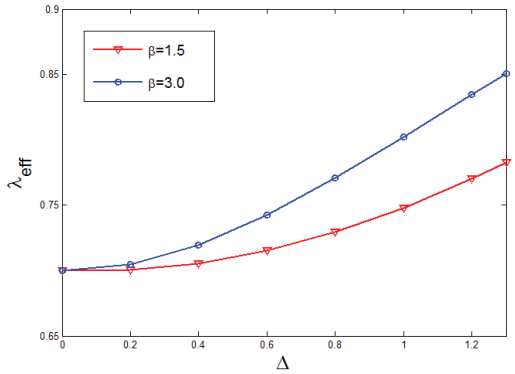


Figure 5: Effective arrival rate $\boldsymbol{\lambda}_{\text{eff}}$ vs. burstiness Δ for Pareto service distributions with $\beta = 1.5$ and $\beta = 3$.

servers N tends to infinity. We proved that the PDE has a unique solution and proposed a numerical scheme for efficiently solving the PDE, thus providing a more computationally effective method than Monte Carlo simulations for analyzing properties of large networks. As an illustration, we applied our approximation to study transient performance measures such as the relaxation time in a backlogged network and the effect of traffic burstiness on a network with periodic time-varying arrivals.

There are many avenues for future research to extend this work. Firstly, it would be worthwhile to continue to study the effect of heavy tails on transient measures of network performance under different scenarios and to also investigate and rigorously establish convergence of the numerical scheme, as the truncation parameters L_0, R_0 go to infinity and the mesh size δ goes to zero, and obtain convergence rates. It would also be of interest to directly study the PDE to analytically establish properties of the limit dynamics. Secondly, as in the ODE method for exponential service time distributions, we would also like to study the equilibrium properties of the PDE. In future work, we hope to show that the PDE has a unique fixed point and characterize its dependence on network parameters.

More broadly, our state representation and overall framework can be applied, with suitable small modifications, to analyze other load-balancing algorithms in the presence of general service distributions, both in the transient and steady state regimes. We hope to use the insight gained from such analyses to also design new algorithms (in both the homogeneous setting considered here, as well as heterogeneous settings) that may lead to better performance.

References

[1] R. Aghajani and K. Ramanan. The hydrodynamic limit of a randomized load balancing network. Preprint, 2015.

[2] S. Asmussen. *Applied Probability and Queues*. Springer-Verlag, 2nd edition edition, 2003.

[3] Y. Azar, A. Z. Broder, A. R. Karlin, and E. Upfal. Balanced allocations. *SIAM J. Comput.*, 29(1):180–200, Sept. 1999.

[4] M. Bramson, Y. Lu, and B. Prabhakar. Randomized load balancing with general service time distributions. *SIGMETRICS Perform. Eval. Rev.*, 38(1):275–286, June 2010.

[5] M. Bramson, Y. Lu, and B. Prabhakar. Asymptotic independence of queues under randomized load balancing. *Queueing Systems*, 71(3):247–292, 2012.

[6] M. Bramson, Y. Lu, and B. Prabhakar. Decay of tails at equilibrium for FIFO join the shortest queue networks. *The Annals of Applied Probability*, 23(5):1841–1878, 10 2013.

[7] L. Brown, N. Gans, A. Mandelbaum, A. Sakov, H. Shen, S. Zeltyn, and L. Zhao. Statistical analysis of a telephone call center. *Journal of the American Statistical Association*, 100(469):36–50, 2005.

[8] S. Chen, Y. Sun, U. Kozat, L. Huang, P. Sinha, G. Liang, X. Liu, and N. Shroff. When queueing meets coding: Optimal-latency data retrieving scheme in storage clouds. In *INFOCOM, 2014 Proceedings IEEE*, pages 1042–1050, April 2014.

[9] R. Courant, E. Isaacson, and M. Rees. On the solution of nonlinear hyperbolic differential equations by finite differences. *Communications on Pure and Applied Mathematics*, 5(3):243–255, 1952.

[10] J. Dai, A. Dieker, and X. Gao. Validity of heavy-traffic steady-state approximations in many-server queues with abandonment. *Queueing Systems*, 78(1):1–29, 2014.

[11] S. Ethier and T. Kurtz. *Markov processes: characterization and convergence*. Wiley series in probability and mathematical statistics. Probability and mathematical statistics. Wiley, 1986.

[12] V. Farias, C. Moallemi, and B. Prabhakar. Load balancing with migration penalties. In *Information Theory, 2005. ISIT 2005. Proceedings. International Symposium on*, pages 558–562, Sept 2005.

[13] C. Graham. Chaoticity on path space for a queueing network with selection of the shortest queue among several. *Journal of Applied Probability*, 37(1):198–211, 03 2000.

[14] C. Grossmann, H. Roos, and M. Stynes. *Numerical Treatment of Partial Differential Equations*. Universitext. Springer Berlin Heidelberg, 2007.

[15] K. Kardassakis. Load balancing in stochastic networks: Algorithms, analysis, and game theory. Undergraduate Honors Thesis, Brown University, 2014.

[16] P. Kolesar. Stalking the endangered CAT: A queueing analysis of congestion at Automatic Teller Machines. *Interfaces*, 14(6):16–26, 1984.

[17] G. Liang and U. Kozat. TOFEC: achieving optimal throughput-delay trade-off of cloud storage using erasure codes. In *INFOCOM, 2014 Proceedings IEEE*, pages 826–834, April 2014.

[18] M. J. Luczak and C. McDiarmid. On the maximum queue length in the supermarket model. *The Annals of Probability*, 34(2):493–527, 03 2006.

[19] M. J. Luczak and J. Norris. Strong approximation for the supermarket model. *The Annals of Applied Probability*, 15(3):2038–2061, 08 2005.

[20] M. Mitzenmacher. Analyses of load stealing models based on families of differential equations. *Theory of Computing Systems*, 34(1):77–98, 2000.

[21] M. Mitzenmacher. The power of two choices in randomized load balancing. *IEEE Trans. Parallel Distrib. Syst.*, 12(10):1094–1104, Oct. 2001.

[22] L. Seelen. An algorithm for Ph/Ph/C queues. *European Journal of Operational Research*, 23(1):118 – 127, 1986.

[23] N. D. Vvedenskaya, R. L. Dobrushin, and F. I. Karpelevich. A queueing system with a choice of the shorter of two queues—an asymptotic approach. *Problemy Peredachi Informatsii*, 32(1):20–34, 1996.