The PDE Method for the Analysis of Randomized Load Balancing Networks

Reza Aghajani
Department of Mathematics,
University of California San Diego
La Jolla, California
maghajani@ucsd.edu

Xingjie Li
University of North Carolina
Charlotte
Charlotte, NC, USA
xli47@uncc.edu

Kavita Ramanan
Brown University
Providence, RI, USA
kavita_ramanan@brown.edu

ABSTRACT

We introduce a new framework for the analysis of large-scale load balancing networks with general service time distributions, motivated by applications in server farms, distributed memory machines, cloud computing and communication systems. For a parallel server network using the so-called $SQ(d)$ load balancing routing policy, we use a novel representation for the state of the system and identify its fluid limit, when the number of servers goes to infinity and the arrival rate per server tends to a constant. The fluid limit is characterized as the unique solution to a countable system of coupled partial differential equations (PDE), which serve to approximate transient Quality of Service parameters such as the expected virtual waiting time and queue length distribution. In the special case when the service time distribution is exponential, our method recovers the well-known ordinary differential equation characterization of the fluid limit.

Furthermore, we develop a numerical scheme to solve the PDE, and demonstrate the efficacy of the PDE approximation by comparing it with Monte Carlo simulations. We also illustrate how the PDE can be used to gain insight into the performance of large networks in practical scenarios by analyzing relaxation times in a backlogged network. In particular, our numerical approximation of the PDE uncovers two interesting properties of relaxation times under the $SQ(2)$ algorithm. Firstly, when the service time distribution is Pareto with unit mean, the relaxation time decreases as the tail becomes heavier. This is *a priori* counterintuitive given that for the Pareto distribution, heavier tails have been shown to lead to worse tail behavior in equilibrium. Secondly, for unit mean light-tailed service distributions such as the Weibull and lognormal, the relaxation time decreases as the variance increases. This is in contrast to the behavior observed under random routing, where the relaxation time increases with increase in variance.

CCS CONCEPTS

• **Mathematics of computing** → Queueing theory; Probabilistic algorithms; Partial differential equations; Numerical analysis; • **Networks** → Network performance analysis;

Permission to make digital or hard copies of part or all of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for third-party components of this work must be honored. For all other uses, contact the owner/author(s).

SIGMETRICS '18 Abstracts, June 18–22, 2018, Irvine, CA, USA

© 2018 Copyright held by the owner/author(s).

ACM ISBN 978-1-4503-5846-0/18/06.

<https://doi.org/10.1145/3219617.3219672>

KEYWORDS

Load Balancing; Supermarket Model; Fluid Limit; Randomized Algorithms; ODE Method; PDE Method; Parallel-server Networks; Join-the-shortest-queue; Power of Two Choices; Cloud Computing

ACM Reference Format:

Reza Aghajani, Xingjie Li, and Kavita Ramanan. 2018. The PDE Method for the Analysis of Randomized Load Balancing Networks. In *SIGMETRICS '18 Abstracts: ACM SIGMETRICS International Conference on Measurement and Modeling of Computer Systems Abstracts, June 18–22, 2018, Irvine, CA, USA*. ACM, New York, NY, USA, 3 pages. <https://doi.org/10.1145/3219617.3219672>

1 INTRODUCTION

Load balancing is an effective method for improving the performance and reliability of networks by optimizing resource use. With growth in the use of server farms and computer clusters, large-scale load balancing networks appear in a variety of applications. The design and analysis of load balancing algorithms 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. The so-called $SQ(d)$ algorithm is a popular algorithm that achieves a better balance between network performance and communication overhead, without imposing additional memory requirements. This algorithm was analyzed by Vvydenskaya et al. in [27] and by Mitzenmacher in [24] in the simple setting 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. The limiting stationary distribution for the $SQ(d)$ model, as $N \rightarrow \infty$, was explicitly computed in [27] and shown to have a double exponential tail decay when $d \geq 2$, in contrast to the exponential decay seen 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 analysis is carried out using the so called “ODE Method”.

1.1 Prior Work

Various modifications of the $SQ(d)$ algorithm have been extensively studied when the service time distribution is exponential [15, 16, 18, 21–23, 25]. However, in most real-world applications, service times are typically not exponentially distributed [9, 10, 19, 20].

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 series of papers by Bramson et al. [6–8], using the so-called “cavity method” or mean-field approach. In particular, it was shown in [8] that for the sub-class of Pareto distributions with parameter β 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, unlike for quantiles of the stationary distribution, there is no interchange of limit result for the tails, i.e., the tail of the limiting stationary distribution need not coincide with the limit of the tails of the N -server stationary distributions. Moreover, the results in [6–8] assume that the cumulative arrivals are Poisson and the service time distribution has a decreasing hazard rate function.

1.2 Our Contributions

In this paper, we analyzing the $SQ(d)$ load balancing networks with general service time distributions and renewal arrival processes. We perform the analysis by 1) the introduction of a new function-valued state descriptor for the $SQ(d)$ network, and 2) characterization of the associated fluid model by a countable system of coupled partial differential equations (PDE), which we call the *hydrodynamic PDE*. We show that the hydrodynamic PDE is amenable to numerical approximation, and is sufficiently rich to approximate important performance measures such as the expected virtual waiting time and the tail probability of a typical queue length. Finally, we illustrate how our PDE method can be used to gain insight into the performance of the network. In particular, while it is known in some cases that heavier tails lead to worse equilibrium performance [8], we identify the somewhat surprising phenomenon that they can lead to better performance with respect to some transient QoS parameters.

2 MAIN RESULTS

2.1 Model Description

We study the performance of a network with N homogeneous parallel servers, each with its own infinite capacity queue, which processes a common stream of arriving jobs that are routed immediately on arrival according to the $SQ(d)$ randomized load balancing algorithm. Here, we only consider the case $d = 2$. The arrival process $E^{(N)}$ is assumed to be a renewal process with average rate λN for some $\lambda > 0$. Moreover, the sequence of service times is assumed to be iid with common cumulative distribution function G , finite mean equal to 1, continuous density g , and a bounded hazard rate function h (see Assumptions 1 and 2 in [1]).

2.2 Hydrodynamic PDE

The fluid model associated to the state representation of the network is characterized as the solution to the so-called hydrodynamic PDE. For every $\lambda > 0$ and appropriate initial conditions $Z_\ell^0 \in \mathbb{C}_{\leq 1}^1[0, \infty)$, $\ell \geq 1$, a set of functions $(Z_\ell; \ell \geq 1)$ is said to solve the *hydrodynamic*

PDE if it satisfies

$$\partial_t Z_\ell(t, r) - \partial_r Z_\ell(t, r) = -\bar{G}(r) \partial_r Z_{\ell+1}(t, 0) + F_\ell(t, r) \quad (1)$$

with initial condition $Z_\ell(0, r) = Z_\ell^0(r)$, where

$$F_1(t, r) = \lambda \bar{G}(r) (1 - Z_1(t, 0))^2, \quad (2)$$

and for $\ell \geq 2$,

$$F_\ell(t, r) = \lambda (Z_{\ell-1}(t, 0) + Z_\ell(t, 0)) (Z_{\ell-1}(t, r) - Z_\ell(t, r)). \quad (3)$$

The hydrodynamic PDE is shown to have a unique solution ([1, Theorem 4.1]).

2.3 Main Results

Our first result shows that the queue length probabilities can be approximated by the solution to the hydrodynamic PDE. Moreover, we show that a “propagation of chaos” result holds at any finite time t . Let $X^{(N), i}(t)$ denote the queue length at server i at time t .

THEOREM 2.1. ([1, Corollary 4.3]) *Under suitable assumptions on the initial condition of the network, for every $t \geq 0$ we have*

$$\lim_{N \rightarrow \infty} \mathbb{P} \left\{ X^{(N), 1}(t) \geq \ell \right\} = Z_\ell(t, 0), \quad (4)$$

where $(Z_\ell; \ell \geq 1)$ be the unique solution to the hydrodynamic PDE. Moreover, the lengths of any finite number of queues are asymptotically independent, that is, for every $t \geq 0$, $K \geq 1$ and $\ell_1, \dots, \ell_K \geq 1$,

$$\lim_{N \rightarrow \infty} \mathbb{P} \left\{ X^{(N), 1}(t) \geq \ell_1, \dots, X^{(N), K}(t) \geq \ell_K \right\} = \prod_{k=1}^K Z_{\ell_k}(t, 0). \quad (5)$$

Our next result shows that for large N , the hydrodynamic PDE also provides an approximation to the expected virtual waiting time $W^{(N)}(t)$ at time t in an N -server network.

THEOREM 2.2. ([1, Theorem 4.5]) *Under suitable assumptions on the initial condition of the network and under Poisson arrival process,*

$$\begin{aligned} \lim_{N \rightarrow \infty} \mathbb{E} \left[W^{(N)}(t) \right] &= \sum_{\ell \geq 2} Z_\ell(t, 0)^2 + \sum_{\ell \geq 1} [Z_\ell(t, 0) + Z_{\ell+1}(t, 0)] \\ &\quad \times \int_0^\infty [Z_\ell(t, r) - Z_{\ell+1}(t, r)] dr, \end{aligned} \quad (6)$$

where $(Z_\ell; \ell \geq 1)$ be the unique solution to the hydrodynamic PDE.

A numerical scheme to solve the hydrodynamic PDE as well as simulation results to validate Theorems 2.1 and 2.2 are provided in [1, Section 5].

3 INSIGHT INTO NETWORK BEHAVIOR

In practice, transient Quality of Service (QoS) parameters are of particular interest in many applications. We provide an example to illustrate how our PDE approximation can be used to shed insight into the effect of the $SQ(2)$ algorithm on transient phenomena of practical relevance.

Given a network that is congested with a backlog of jobs, the system administrator would like to estimate 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. We consider a network with a large initial backlog, and hence a large initial virtual waiting time, and study the “relaxation time”, which we define to be the

Figure 1: Expected virtual waiting time at time t vs. t for a network with initial backlog and unit mean Pareto service time distribution with $\beta = 1.25$ (red) and $\beta = 2.50$ (blue), under the SQ(2) routing algorithm using the PDE method.

time it takes for the network backlog to decrease to the extent that the expected virtual waiting time reaches half of its initial value. We investigate the effect of the variance of the service time distribution on the relaxation time. We consider a network that has been backlogged due to experiencing an unusually high arrival rate for a period of time. To study the relaxation time, in Figure 1, we plot the evolution in time of the expected virtual waiting time for two different Pareto service time distributions with unit mean: the first having parameter $\beta = 1.25$ (heavy-tailed), and the second having parameter $\beta = 2.50$ (light-tailed). To see the detailed list of parameters for this scenario, see [1, Section 6].

An interesting observation from Figure 1 is that the relaxation time under the SQ(2) algorithm in the heavy-tailed case is clearly smaller than in the light-tailed case. This phenomenon persists under different scenarios or mechanisms leading to a backlog, which correspond to different initial distributions for the queue lengths and ages. This may seem surprising in view of the steady-state result of Bramson et al. in [8], 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, which suggest that at equilibrium, the light-tailed Pareto distribution shows better performance.

A possible heuristic explanation for the superior performance of the relaxation time when the tail is heavier is as follows. When the service distribution is fixed to have mean equal to one, the median of the distribution decreases with a decrease in β (equivalently, increase in variance). As a result, the heavier the tail of the service time 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 process a large number of jobs faster to reduce their queue lengths, with only a few jobs with long service times leading to large queue lengths. See [1, Section 6] for a more detailed investigation of this observation.

REFERENCES

- [1] Reza Aghajani, Xingjie Li, and Kavita Ramanan. 2017. The PDE Method for the Analysis of Randomized Load Balancing Networks. *Proc. ACM Meas. Anal. Comput. Syst.* 1, 2, Article 38 (Dec. 2017), 28 pages. <https://doi.org/10.1145/3154497>
- [2] R. Aghajani and K. Ramanan. 2017. Hydrodynamic Limit of a Randomized Load Balancing Network. (2017). [arXiv:1707.02005 \[math.PR\]](https://arxiv.org/abs/1707.02005).
- [3] S. Asmussen. 2003. *Applied Probability and Queues* (2nd edition ed.). Springer-Verlag.
- [4] Yossi Azar, Andrei Z. Broder, Anna R. Karlin, and Eli Upfal. 1999. Balanced Allocations. *SIAM J. Comput.* 29, 1 (Sept. 1999), 180–200. <https://doi.org/10.1137/S0097539795288490>
- [5] P. Billingsley. 1968. *Convergence of Probability Measures*. John Wiley, New York.
- [6] Maury Bramson, Yi Lu, and Balaji Prabhakar. 2010. Randomized Load Balancing with General Service Time Distributions. *SIGMETRICS Perform. Eval. Rev.* 38, 1 (June 2010), 275–286. <https://doi.org/10.1145/1811099.1811071>
- [7] Maury Bramson, Yi Lu, and Balaji Prabhakar. 2012. Asymptotic independence of queues under randomized load balancing. *Queueing Systems* 71, 3 (2012), 247–292. <https://doi.org/10.1007/s11134-012-9311-0>
- [8] Maury Bramson, Yi Lu, and Balaji Prabhakar. 2013. Decay of tails at equilibrium for FIFO join the shortest queue networks. *The Annals of Applied Probability* 23, 5 (10 2013), 1841–1878. <https://doi.org/10.1214/12-AAP888>
- [9] L. Brown, N. Gans, A. Mandelbaum, A. Sakov, H. Shen, S. Zeltyn, and L. Zhao. 2005. Statistical Analysis of a Telephone Call Center. *J. Amer. Statist. Assoc.* 100, 469 (2005), 36–50.
- [10] Shengbo Chen, Yin Sun, U.C. Kozat, Longbo Huang, P. Sinha, Guanfang Liang, Xin Liu, and N.B. Shroff. 2014. When queueing meets coding: Optimal-latency data retrieving scheme in storage clouds. In *INFOCOM, 2014 Proceedings IEEE*. 1042–1050. <https://doi.org/10.1109/INFOCOM.2014.6848034>
- [11] Richard Courant, Eugene Isaacson, and Mina Rees. 1952. On the solution of nonlinear hyperbolic differential equations by finite differences. *Communications on Pure and Applied Mathematics* 5, 3 (1952), 243–255. <https://doi.org/10.1002/cpa.3160050303>
- [12] J.G. Dai, A.B. Dieker, and Xuefeng Gao. 2014. Validity of heavy-traffic steady-state approximations in many-server queues with abandonment. *Queueing Systems* 78, 1 (2014), 1–29. <https://doi.org/10.1007/s11134-014-9394-x>
- [13] S.N. Ethier and T.G. Kurtz. 1986. *Markov processes: characterization and convergence*. Wiley.
- [14] L.C. Evans. 1998. *Partial Differential Equations*. American Math. Soc.
- [15] V.F. Farias, C.C. Moallemi, and B. Prabhakar. 2005. Load balancing with migration penalties. In *Information Theory, 2005. ISIT 2005. Proceedings. International Symposium on*. 558–562. <https://doi.org/10.1109/ISIT.2005.1523397>
- [16] Carl Graham. 2000. Chaoticity on path space for a queueing network with selection of the shortest queue among several. *Journal of Applied Probability* 37, 1 (03 2000), 198–211. <https://doi.org/10.1239/jap/1014842277>
- [17] C. Grossmann, H.G. Roos, and M. Stynes. 2007. *Numerical Treatment of Partial Differential Equations*. Springer Berlin Heidelberg. <https://books.google.com/books?id=Nu7-ZZAYPZYC>
- [18] K. Kardassakis. 2014. Load Balancing in Stochastic Networks: Algorithms, Analysis, and Game Theory. Undergraduate Honors Thesis, Brown University.
- [19] Peter Kolesar. 1984. Stalking the Endangered CAT: A Queueing Analysis of Congestion at Automatic Teller Machines. *Interfaces* 14, 6 (1984), 16–26.
- [20] Guanfang Liang and U.C. Kozat. 2014. TOFEC: Achieving optimal throughput-delay trade-off of cloud storage using erasure codes. In *INFOCOM, 2014 Proceedings IEEE*. 826–834. <https://doi.org/10.1109/INFOCOM.2014.6848010>
- [21] Malwina J. Luczak and Colin McDiarmid. 2006. On the maximum queue length in the supermarket model. *The Annals of Probability* 34, 2 (03 2006), 493–527. <https://doi.org/10.1214/00911790500000710>
- [22] Malwina J. Luczak and James Norris. 2005. Strong approximation for the supermarket model. *The Annals of Applied Probability* 15, 3 (08 2005), 2038–2061. <https://doi.org/10.1214/105051605000000368>
- [23] M. Mitzenmacher. 2000. Analyses of Load Stealing Models Based on Families of Differential Equations. *Theory of Computing Systems* 34, 1 (2000), 77–98. <https://doi.org/10.1007/s002240010010>
- [24] Michael Mitzenmacher. 2001. The Power of Two Choices in Randomized Load Balancing. *IEEE Trans. Parallel Distrib. Syst.* 12, 10 (Oct. 2001), 1094–1104. <https://doi.org/10.1109/71.963420>
- [25] D. Mukherjee, S. Borst, J. van Leeuwen, and P. Whiting. 2016. Universality of Power-of-d Load Balancing in Many-Server Systems. (2016). [arXiv:1612.00723 \[math.PR\]](https://arxiv.org/abs/1612.00723).
- [26] L.P. Seelen. 1986. An algorithm for Ph/Ph/c queues. *European Journal of Operational Research* 23, 1 (1986), 118 – 127. [https://doi.org/10.1016/0377-2217\(86\)90222-5](https://doi.org/10.1016/0377-2217(86)90222-5)
- [27] N. D. Vvedenskaya, R. L. Dobrushin, and F. I. Karpelevich. 1996. A queueing system with a choice of the shorter of two queues—an asymptotic approach. *Problemy Peredachi Informatsii* 32, 1 (1996), 20–34.