Adv. Appl. Prob. **47,** 1015–1038 (2015)
Printed in Northern Ireland

RANDOMIZED LONGEST-QUEUE-FIRST SCHEDULING FOR LARGE-SCALE BUFFERED SYSTEMS

A. B. DIEKER [*] AND
T. SUK,[*] [**] *Georgia Institute of Technology*

Abstract

We develop diffusion approximations for parallel-queueing systems with the randomized *longest-queue-first* scheduling (LQF) algorithm by establishing new mean-field limit theorems as the number of buffers $n \to \infty$. We achieve this by allowing the number of sampled buffers $d = d(n)$ to depend on the number of buffers n, which yields an asymptotic 'decoupling' of the queue length processes. We show through simulation experiments that the resulting approximation is accurate even for moderate values of n and $d(n)$. To the best of the authors' knowledge, this is the first derivation of diffusion approximations for a queueing system in the large-buffer mean-field regime. Another noteworthy feature of our scaling idea is that the randomized LQF algorithm emulates the LQF algorithm, yet is computationally more attractive. The analysis of the system performance as a function of $d(n)$ is facilitated by the multi-scale nature in our limit theorems: the various processes we study have different space scalings. This allows us to show the trade-off between performance and complexity of the randomized LQF scheduling algorithm.

Keywords: Queueing; mean-field; fluid limit; diffusion limit

2010 Mathematics Subject Classification: Primary 90B22
Secondary 90B36

1. Introduction

Resource pooling is becoming increasingly common in modern applications of stochastic systems, such as in computer systems, wireless networks, workforce management, call centers, and health care delivery. At the same time, these applications give rise to systems which continue to grow in size. For instance, a traditional web server farm only has a few servers, while cloud data centers have thousands of processors. These two trends pose significant practical restrictions on admission, routeing, and scheduling decision rules or algorithms. Scalability and computability are becoming ever more important characteristics of decision rules, and, consequently, simple decision rules with good performance are of particular interest. An example is the so-called least connection rule implemented in many load balancers in computer clouds, which assigns a task to the server with the least number of active connections; cf. the join-the-shortest-queue routeing policy. From a design point of view, the search for desirable algorithmic features often presents trade-offs between system performance, information/communication, and required computational effort.

Over the past decades, mean-field models have become mainstream aids in the design and performance assessment of large-scale stochastic systems; see, for example, [2], [3], [10],

Received 16 April 2014; revision received 18 August 2014.

[*] Postal address: H. Milton Stewart School of Industrial and System Engineering, Georgia Institute of Technology, 755 Ferst Drive, NW, Atlanta, GA 30332-0205, USA.
[**] Email address: tonghoons@gatech.edu

https://doi.org/10.1239/aap/1449859798 Published online by Cambridge University Press

[12], and [15]. These models allow for summary system dynamics to be approximated using a mean-field scaling, which leads to deterministic 'fluid' approximations. Although these approximations are designed for large systems, they typically do not work well unless the scaling parameter n is excessively large. In view of this, it is of interest to find more refined approximations than fluid approximations. In this paper we derive diffusion approximations in a specific instance of a large-scale stochastic system: a queueing system with many buffers with a randomized longest-queue-first (LQF) scheduling algorithm. Under this scheduling algorithm, the server works on a task from the buffer with the longest queue length among several sampled buffers; it approximates the LQF scheduling policy, but it is computationally more attractive if the number of buffers is large.

In our model, each buffer is fed with an independent stream of tasks, which arrive according to a Poisson process. All n buffers are connected to a single centralized server. Under the randomized LQF policy, this server selects $d(n)$ buffers uniformly at random (with replacement) and processes a task from the longest queue among the selected buffers; it idles for a random amount of time if all buffers in the sample are empty. Tasks have random processing time requirements. The total processing capacity scales linearly with n and the processing time distribution is independent of n. We work in an underloaded regime, with enough processing capacity to eventually serve all arriving tasks. Note that this scheduling algorithm is agnostic in the sense that it does not use arrival rates. By establishing limit theorems, we develop approximations for the queue length processes in the system, and show that the approximations are accurate even for moderate n and $d(n)$. Also, we study the trade-off between performance and complexity of the algorithm.

Most existing work on the mean-field large-buffer asymptotic regime for queueing systems concentrates on the so-called supermarket model, which has received much attention over the past decades following the work of Vvedenskaya *et al.* [16]; see also [13] and follow-up work. The focus of this line of work centers on the question of how incoming tasks should be routed to buffers, i.e. the load balancing problem. For the randomized join-the-shortest-queue routeing policy, where tasks are routed to the buffer with the shortest queue length among d uniformly selected buffers, this line of work has exposed a dramatic improvement in performance when $d = 2$ versus $d = 1$. This phenomenon is known as the *power of two choices*. A recently proposed different approach for the load balancing problem is inspired by the cavity method [4]–[6]. This approach is a significant advance to current methodologies since it does not require exponentially distributed service times. However, applying this methodology to our setting presents significant challenges due to the scaling employed here. We do not consider this method here, it remains an open problem as to whether the cavity method can be applied to our setting.

The papers by Alanyali and Dashouk [1] and Tsitsiklis and Xu [14] are closely related to this paper. Both consider scheduling in the presence of a large number of buffers. In [1] the authors studied the randomized LQF policy with $d(n) = d$, and the main finding is that the empirical distribution of the queue lengths in the buffer is asymptotically geometric with parameter depending on d. It established an upper bound on the asymptotic order, but here we establish tightness and identify the limit. A certain time scaling that is not present in [1] is essential for the validity of our limit theorems. In [14] the authors analyzed a hybrid system with centralized and distributed processing capacity in a setting similar to ours. Their work exposes a dramatic improvement in performance in the presence of centralization compared to a fully distributed system.

https://doi.org/10.1239/aap/1449859798 Published online by Cambridge University Press

We establish a diffusion limit theory for a queueing system in the large-buffer mean-field regime. Diffusion approximations are well-known to arise in the context of mean-field models (see, for example, [11]) but off-the-shelf results typically cannot directly be applied due to intricate dependencies or technical intricacies. Thus, by and large, second-order diffusion approximations have been uncharted territory for many large-scale queueing systems.

Our analysis is facilitated by the idea to scale the number of sampled buffers $d(n)$ with the number of buffers n, which asymptotically 'decouples' the buffers and, consequently, removes certain dependencies among the buffer contents. The decoupling manifests itself through a limit theorem on multiple scales, where the various queue-length processes we study have different space scalings. We show empirically that this result leads to accurate approximations even when the number of buffers n is small, i.e. outside of the asymptotic regime that motivated the approximation.

For our system, since the scheduling algorithm depends on n, several standard arguments for large-scale systems break down due to the multi-scale nature of the various stochastic processes involved; thus, our work requires several technical novelties. Among these is an induction-based argument for establishing the existence of a fluid model. We also rely on an appropriate time scaling, which is specific to our case and has not been employed in other work.

Our fluid limit theory makes explicit the trade-off between performance and complexity for our algorithm. Intuitively, we expect better system performance for larger $d(n)$, since the likelihood of idling decreases; however, the computational effort also increases since we must sample (and compare) the queue length of more buffers. Our main insight into the interplay between performance (i.e. low queue lengths) and computational complexity of the scheduling algorithm within our model can be summarized as follows. We study the fraction of queues with at least k tasks, and show that it is of the order of $1/d(n)^k$ under the randomized LQF scheduling policy. This strengthens and generalizes the upper bound from [1]. Thus, the average queue length is of the order of $1/d(n)$ as n approaches ∞. This should be contrasted with $d(n)$, which is the order of the computational complexity of the scheduling algorithm.

The randomized LQF algorithm approximates the LQF algorithm, which is a fully centralized policy, so it is appropriate to make a comparison with the partially centralized scheduling algorithm from [14], where all n buffers are used with probability $p > 0$ (and one buffer is chosen uniformly at random otherwise). Our algorithm has better performance although it compares only $d(n) \ll n$ buffers per job as opposed to $pn + 1 - p$, which is the average number of buffers used in the partially centralized algorithm.

We introduce our model in Section 2. Our main results come in two pieces: limit theorems (Section 3) and approximations with validation (Section 4). Section 5 contains the proofs of our limit theorems. Finally, Appendix A has several standard results that we have included for quick reference.

2. Model and notation

The systems we are interested in consist of many parallel queues and a single server. Consider a system with n buffers, which temporarily store tasks to be served by the (central) server. The number of tasks in a buffer is called its queue length. Buffers temporarily hold tasks in anticipation of processing, and tasks arrive according to independent Poisson processes with rate $\lambda < 1$. The processing times of the tasks are independent and identically distributed (i.i.d.) with an exponential distribution with unit mean. All processing times are independent of the arrival processes. The server serves tasks at rate n.

https://doi.org/10.1239/aap/1449859798 Published online by Cambridge University Press

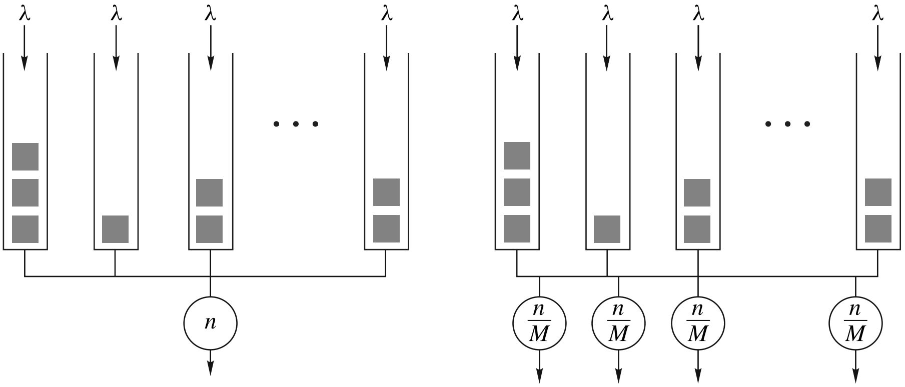

FIGURE 1: Our models with n buffers. One central server with service rate n (*left*) and M servers with service rates n/M (*right*).

The server schedules tasks as follows. It selects $d(n)$ buffers uniformly at random (with replacement) and processes a task in the buffer with the longest queue length among the selected buffers. Ties are broken by selecting a buffer uniformly at random among those with the longest queue length. If all selected buffers are empty, then the service opportunity is wasted and the server waits for an exponentially distributed amount of time with parameter n before resampling. Once a task has been processed, it immediately leaves the system. We do not consider scheduling within buffers, since we only study queue lengths. Throughout, we are interested in the case when $d(n)$ satisfies $d(n) = o(n)$ and $\lim_{n\to\infty} d(n) = \infty$.

In this model description it is not essential that there is exactly one server. Indeed, the same dynamics arise if an arbitrary number M of servers process tasks at rate n/M, as long as each server uses the randomized LQF policy. This model arises in the content of cellular data communications [1]. An abstract representation of the model is shown in Figure 1.

Let $F_{n,k}(t)$ be the fraction of buffers with queue length greater than or equal to k at time t in the system with n buffers, so that $\{F_{n,k}(t)\}_{k\in\mathbb{N}}$ is a Markov process. Such mean-field quantities have been used in analyzing various scheduling and load balancing policies; see, for example, [1], [13], [14]. However, under the randomized LQF policy, we can expect from [1] that, whenever $\lim_{n\to\infty} d(n) = \infty$,

$$\lim_{t\to\infty}\lim_{n\to\infty} F_{n,k}(t) = 0 \quad \text{for all } k \geq 1,$$

i.e. in this sense the performance is asymptotically the same as that of the LQF policy, and these random variables are asymptotically degenerate.

3. Limit theorems

In this section we present limit theorems which are stated in terms of $F_{n,k}(\cdot)$ under appropriate scaling. Let $K \in \mathbb{N}$ be a fixed finite integer satisfying $\lim_{n\to\infty} n/d(n)^K = \infty$. Let $U_{n,k}(\cdot)$ be the following modification of $F_{n,k}(\cdot)$:

$$U_{n,k}(t) := d(n)^k\, F_{n,k}\left(\frac{t}{d(n)}\right) \quad \text{for all } k = 0, 1, \dots, K.$$

https://doi.org/10.1239/aap/1449859798 Published online by Cambridge University Press

Our first limit theorem is that $\{(U_{n,1}(t), \ldots, U_{n,K}(t))\}_{n\in\mathbb{N}}$ has a fluid limit as $n \to \infty$ and that this fluid limit satisfies the system of differential equations described in the following definition.

Definition 1. For $v_1, \ldots, v_K \in \mathbb{R}_+$, $(u_1(t), \ldots, u_K(t))$ is said to be a *LQF fluid limit system* with initial condition $(v_1, \ldots, v_K)$ if

(i) $u_k \colon [0, \infty) \to \mathbb{R}_+$ with $u_k(0) = v_k$ for all $k = 1, \ldots, K$;

(ii) $u_1'(t) = \mathrm{e}^{-u_1(t)} - 1 + \lambda$;

(iii) $u_k'(t) = \lambda\, u_{k-1}(t) - u_k(t)$ for all $k = 2, \ldots, K$.

By the usual existence and the uniqueness theorem of first-order ordinary differential equations (see, for example, [7]), there is a unique differentiable function $u_1 \colon [0, \infty) \to \mathbb{R}_+$ with $u_1(0) = v_1$ satisfying Definition 1(ii). For $k \geq 2$, when $u_{k-1}(t)$ and v_k are given, the differential equation of u_k is linear with inhomogeneous part $u_{k-1}(t)$ and, therefore, $u_k \colon [0, \infty) \to \mathbb{R}_+$ is unique. Thus, by induction, for any given initial condition, there is a unique LQF fluid limit system.

We remark that the following is an explicit expression of the solution if $v_1 < \ln(1/(1-\lambda))$ (the other case yields a similar expression):

$$u_1(t) = \ln\left(\frac{C_1\mathrm{e}^{(1-\lambda)t} - 1}{C_1(1-\lambda)\mathrm{e}^{(1-\lambda)t}}\right),$$

$$u_k(t) = \mathrm{e}^{-t}\, v_k + \lambda\int_0^t \mathrm{e}^{-(t-s)} u_{k-1}(s)\,\mathrm{d}s, \qquad k = 2, \ldots, K,$$

where $C_1 = 1/(1 - (1-\lambda)\mathrm{e}^{v_1})$. Moreover, a LQF fluid limit system has a unique critical point which is stable:

$$\left(\ln\left(\frac{1}{1-\lambda}\right), \lambda\ln\left(\frac{1}{1-\lambda}\right), \ldots, \lambda^{K-1}\ln\left(\frac{1}{1-\lambda}\right)\right).$$

The following proposition summarizes these arguments.

Proposition 1. *For any* $(v_1, \ldots, v_K) \in \mathbb{R}_+^K$, *there is a unique LQF fluid limit system* $(u_1(t), \ldots, u_K(t))$ *with* $u_k(0) = v_k$ *for all* $k = 1, \ldots, K$, *and*

$$(u_1(t), u_2(t), \ldots, u_K(t)) \to \left(\ln\left(\frac{1}{1-\lambda}\right), \lambda\ln\left(\frac{1}{1-\lambda}\right), \ldots, \lambda^{K-1}\ln\left(\frac{1}{1-\lambda}\right)\right)$$

as $t \to \infty$.

Our first limit theorem states that, with an appropriate initial condition, $(U_{n,1}(t), \ldots, U_{n,K}(t))$ converges to a fluid limit system as $n \to \infty$.

Theorem 1. (Fluid limit.) *Consider a sequence of systems indexed by* n. *Fix a number* $K \in \mathbb{N}$ *such that* $\lim_{n\to\infty} n/d(n)^K = \infty$. *Assume that* $U_{n,k}(0)$ *is deterministic for every* n *and* $k \leq K$, *and that there exist* $v_1, \ldots, v_K \in \mathbb{R}_+$ *such that*

$$\lim_{n\to\infty} U_{n,k}(0) = v_k, \quad k = 1, \ldots, K, \qquad \lim_{n\to\infty} d(n)^K (F_{n,K+1}(0) + F_{n,K+2}(0) + \cdots) = 0.$$

Then the sequence of stochastic processes $\{(U_{n,1}(t), \ldots, U_{n,K}(t))\}_{n\in\mathbb{N}}$ *converge almost surely to the LQF fluid limit system* $(u_1(t), \ldots, u_K(t))$ *with initial condition* $(v_1, \ldots, v_K)$, *uniformly on compact sets.*

https://doi.org/10.1239/aap/1449859798 Published online by Cambridge University Press

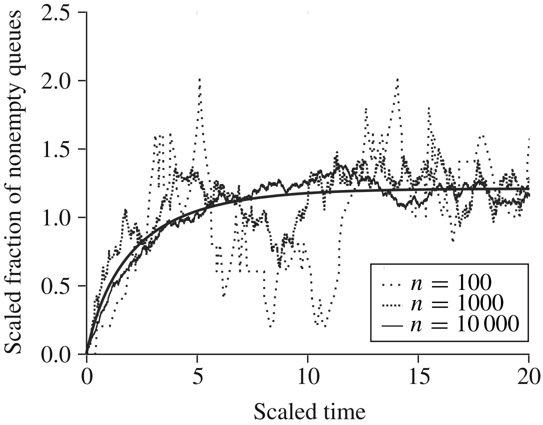

FIGURE 2: Sample paths of $U_{n,1}(t)$ for various n, with $d(n) = 10\log_{10}(n)$ and $\lambda = 0.7$. The solid curve is the solution of $u'(t) = \mathrm{e}^{-u(t)} - 1 + \lambda$.

The proof of the above theorem is based on mathematical induction, and we give a high-level overview of this proof at the beginning of Section 5.

This result makes the explicit trade-off between performance and complexity for randomized LQF algorithms. Theorem 1 shows that for $k = 1, \dots, K$, as $n \to \infty$,

$$F_{n,k}\left(\frac{t}{d(n)}\right) = \Theta\left(\frac{1}{d(n)^k}\right).$$

For $k = 1$, this agrees with the upper bound sketched in [1]. Then the average queue length is of the order of $1/d(n)$, inverse of the complexity. In the next section we investigate this by simulation.

In Figure 2 we show sample paths of $U_{n,1}(t)$ (the scaled fraction of nonempty queues) for various n, empirically confirming our first limit theorem. However, even for n as large as 10 000, the sample paths fluctuate around the fluid limit, especially for large t. This means that it is important to incorporate a second-order approximation.

Our second limit theorem concerns the diffusion limit of $U_{n,1}(t)$ as $n \to \infty$. Precisely, we show that the stochastic processes $U_{n,1}(t)$ converge in distribution to a diffusion process after appropriate scaling. We believe it is the first diffusion limit theorem for a queueing system in the large-buffer mean-field regime, and is based on an asymptotic 'decoupling' of the queue length processes. Note that $U_{n,1}(t)$ is not a Markov process, but the approximating process $Z(t)$ is a Markov process. In Appendix A we explain the exact meaning of this type of convergence, for which we use the symbol '$\xrightarrow{\mathrm{D}}$'.

Theorem 2. (Diffusion limit.) *Consider a sequence of systems indexed by n. Suppose that* $\lim_{n\to\infty} n/d(n) = \infty$ *and* $\lim_{n\to\infty} n/d(n)^2 = 0$. *Assume that* $U_{n,1}(0)$ *is deterministic for all n, and that there exists some* $v_1 \in \mathbb{R}_+$ *such that*

$$\lim_{n\to\infty} \sqrt{\frac{n}{d(n)}}(U_{n,1}(0) - v_1) = 0 \tag{1}$$

and

$$\lim_{n\to\infty} \sqrt{nd(n)}(F_{n,2}(0) + F_{n,3}(0) + \cdots) = 0. \tag{2}$$

https://doi.org/10.1239/aap/1449859798 Published online by Cambridge University Press

Then, we have

$$\sqrt{\frac{n}{d(n)}}(U_{n,1}(t) - u_1(t)) \xrightarrow{\mathrm{D}} Z(t) \quad \text{as } n \to \infty,$$

where $Z(t)$ is the solution of the following Itô integral equation:

$$Z(t) = \sqrt{\lambda}B^{(1)}(t) - \int_0^t \sqrt{1 - \mathrm{e}^{-u_1(s)}}\,\mathrm{d}B^{(2)}(s) - \int_0^t \mathrm{e}^{-u_1(s)}Z(s)\,\mathrm{d}s$$

for independent Wiener processes $B^{(1)}(t)$ and $B^{(2)}(t)$.

We anticipate that this theorem can be generalized as follows. The process $U_{n,k}(t)$ couples with $u_{k+1}(t)$ (the scaling limit of $U_{k+1}(t)$), but the fact that their scaling behavior is different ($\sqrt{n/d(n)^k}$ versus $\sqrt{n/d(n)^{k+1}}$) introduces complications for the proof technique used for Theorem 2.

Conjecture 1. *Consider a sequence of systems indexed by n. Suppose that $\lim_{n\to\infty} n/d(n) = \infty$ and fix $k \le K$, where K is defined at the beginning of this section. Assume that $U_{n,k}(0)$ is deterministic for all n and $k \le K$, and that there exists $v_1, \dots, v_K \in \mathbb{R}_+$ and $v_1^*, \dots, v_K^* \in \mathbb{R}$ such that*

$$\lim_{n\to\infty} \sqrt{\frac{n}{d(n)^k}}(U_{n,k}(0) - v_k) = v_k^*.$$

Additionally, assume that

$$\lim_{n\to\infty} \sqrt{n\, d(n)^{K+1}}(F_{n,K+1}(0) + F_{n,K+2}(0) + \cdots) = 0.$$

Then, we have

$$\sqrt{\frac{n}{d(n)^k}}\left(U_{n,k}(t) - u_k(t) + \frac{1}{d(n)}u_{k+1}(t)\right) \xrightarrow{\mathrm{D}} Z_k(t) \quad \text{as } n \to \infty,$$

where we interpret $u_{K+1}(t)$ as 0, and $Z_1(t)$ is the solution of the following Itô integral equation:

$$Z_1(t) = v_1^* + \sqrt{\lambda}B_1^{(1)}(t) - \int_0^t \sqrt{1 - \mathrm{e}^{-u_1(s)}}\,\mathrm{d}B_1^{(2)}(s) - \int_0^t \mathrm{e}^{-u_1(s)}\,Z_1(s)\,\mathrm{d}s.$$

For $k = 2, \dots, K$, $Z_k(t)$ is the solution of the following Itô integral equation:

$$Z_k(t) = v_k^* + \int_0^t \sqrt{\lambda u_{k-1}(s)}\,\mathrm{d}B_k^{(1)}(s) - \int_0^t \sqrt{u_k(s)}\,\mathrm{d}B_k^{(2)}(s) - \int_0^t Z_k(s)\,\mathrm{d}s$$

for independent Wiener processes $B_k^{(1)}(t)$ and $B_k^{(2)}(t)$.

Next, we utilize our limit theorems outlined above to establish approximations of the processes in our system and show their accuracy by simulation.

4. Approximation and validation

In this section we propose diffusion approximations based on our limit theorems in the previous section, and we investigate the discrepancy between these approximations and the original prelimit system. In addition, we examine the trade-off between performance (average queue length) and complexity (the number of samples) through simulation.

Our limit theorems are stated in terms of a function $d(n)$, but here we investigate systems for which we sample a fixed number of buffers d. For simplicity, we only consider systems that are initially empty.

https://doi.org/10.1239/aap/1449859798 Published online by Cambridge University Press

4.1. Diffusion approximations

Our diffusion limit theorem suggests the following approximation for the distribution of the fraction of nonempty queues in a system with n buffers and d samples:

$$F_{n,1}(t) \approx \frac{1}{d} u_1(dt) + \frac{1}{\sqrt{nd}} Z(dt) \quad \text{(Diffusion approximation (DA))},$$

where $u_1(t)$ is the fluid limit of $U_{n,1}(t)$ from Theorem 1 and $Z(t)$ is the Gaussian process defined in Theorem 2. One of the assumptions in Theorem 2 is $\lim_{n\to\infty} n/d(n)^2 = 0$, which may not be plausible for systems with relatively small d compared to n; we confirm this later. Our conjecture in Section 3 suggests adjusting the DA as follows:

$$F_{n,1}(t) \approx \frac{1}{d} u_1(dt) - \frac{1}{d^2} u_2(dt) + \frac{1}{\sqrt{nd}} Z(dt) \quad \text{(Modified diffusion approximation (MDA))},$$

where $u_1(t)$ and $Z(t)$ are the same as the DA, and $u_2(t)$ is the fluid limit of $U_{n,2}(t)$ in Theorem 1.

Since Z is a centered Gaussian process, the distribution of $F_{n,1}(t)$ is approximately normal for fixed t. To be able to describe the variance, we need $\sigma^2(t) = \mathrm{var}[Z(t)]$. From standard stochastic differential equation results, $\sigma^2(t)$ satisfies the ordinary differential equation

$$\frac{\mathrm{d}}{\mathrm{d}t}\sigma^2(t) = -2\mathrm{e}^{-u_1(t)}\sigma^2(t) + \lambda + (1 - \mathrm{e}^{-u_1(t)}) \tag{3}$$

with initial condition $\sigma^2(0) = 0$.

To investigate the accuracy of our approximations, we collect simulation samples of the fraction of nonempty buffers $F_{n,1}(t)$ and compare the resulting histogram with our approximations. The normal distributions from our two approximations of $F_{n,1}(t)$ have the same variance, but their means are different.

First, we check the accuracy of the DA for moderate n and d. For $\lambda = 0.7$ and $n = 20$, we produce a histogram with 100 000 samples of $F_{20,1}(50)$ for $d = 4$ and $d = 12$ and compare this with the probability density function of the normal distribution from the DA; see Figure 3.

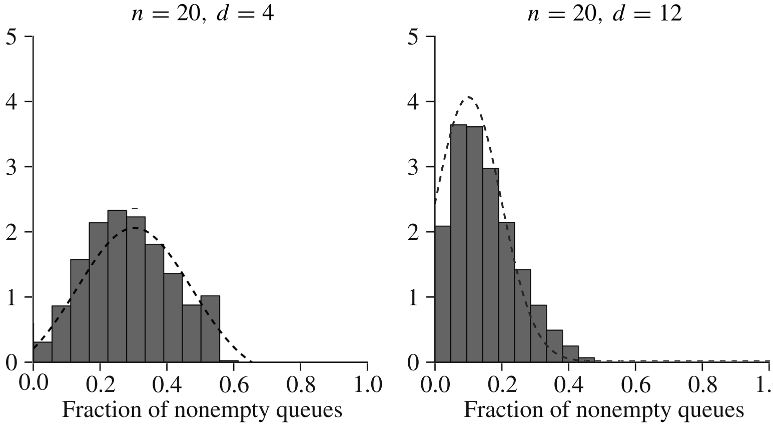

FIGURE 3: DA versus simulation of the distribution of $F_{n,1}(50)$ for moderate n and d. *Left:* $n = 20$, $d = 4$. *Right:* $n = 20$, $d = 12$.

https://doi.org/10.1239/aap/1449859798 Published online by Cambridge University Press

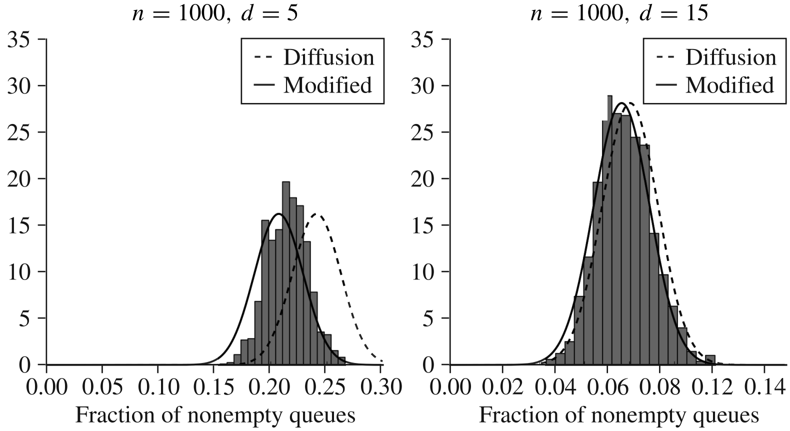

FIGURE 4: Our approximations versus simulation of the distribution of $F_{n,1}(50)$ for large $n = 1000$. *Left:* $d = 5$. *Right:* $d = 15$. Shown are from the DA (*dashed lines*) and results based on the MDA (*solid lines*).

Through these and other experiments we find that the DA is accurate even when n is moderate and works best in cases where d is small compared to n, which is the regime of our theoretical results. When d is large compared to n, then the distribution becomes more concentrated at 0.

Second, we verify our approximations for large n and small d. Applying algorithms with small computational complexity to large systems is most meaningful in practice, and this is the case in our model when the number of buffers n is large and the number of samples d is small. By simulation, we obtain histograms of 1000 samples of the fraction of nonempty queues at time 50 ($F_{n,1}(50)$) for $n = 1000$ and $\lambda = 0.7$ as shown in Figure 4. This result shows that the ordinary differential equation (3) gives a good approximation of the variance of $F_{n,1}(50)$. For the mean of $F_{n,1}(50)$, the MDA is more accurate than the DA when d is relatively small. As d grows, the DA better estimates the mean of $F_{n,1}(50)$. This shows that our theorems provide good approximations in practically attractive situations.

We next empirically study when our approximation works well, with the objective to find a criterion depending on n, d, and λ for the validity of our approximation. From the MDA, we find the following approximations for the mean and the standard deviation of $F_{n,1}(t)$ for reasonably large t:

$$\mu \simeq \left(\frac{1}{d} - \frac{\lambda}{d^2}\right) \ln\left(\frac{1}{1-\lambda}\right), \qquad \sigma \simeq \frac{1}{\sqrt{nd}}\sqrt{\frac{\lambda}{1-\lambda}},$$

where we use Proposition 1 and set $\mathrm{d}\sigma^2(t)/\,\mathrm{d}t = 0$ in (3).

We use the Kolmogorov–Smirnov distance between our approximation and the empirical distribution (from simulation) as a measure of accuracy of our approximation. We find that the quality of our approximation depends on n, d, and λ mostly through μ and σ; Figure 5 summarizes the data from our experiments by plotting the results in the (μ, σ) plane. The experiments show that the MDA works well if μ and σ satisfy $\sigma < \mu/3$ and $\sigma > 2(\mu - \frac{1}{4})/3$. We also tested the choice of t on the accuracy of our approximation and found that it does not have a significant effect.

Another observation we identified from these simulation experiments is that the variance is not negligible compared to the mean of the fraction of nonempty queues even when n is large.

https://doi.org/10.1239/aap/1449859798 Published online by Cambridge University Press

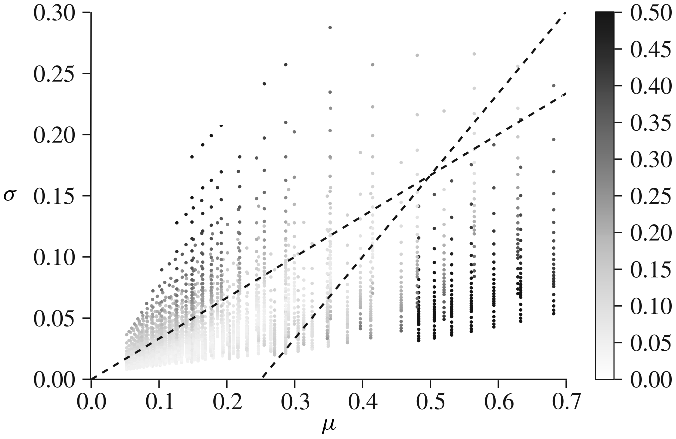

FIGURE 5: The Kolmogorov–Smirnov test statistic for various parameter values. We use 5000 simulation replications to estimate the distribution of $F_{n,1}(100)$ for $n = 100, 150, \dots, 1000, 1200, \dots, 2000$, $d = 2, 5, 7, 10, 12, \dots, 30$, and $\lambda = 0.80, 0.82, 0.84, \dots, 0.98, 0.99$.

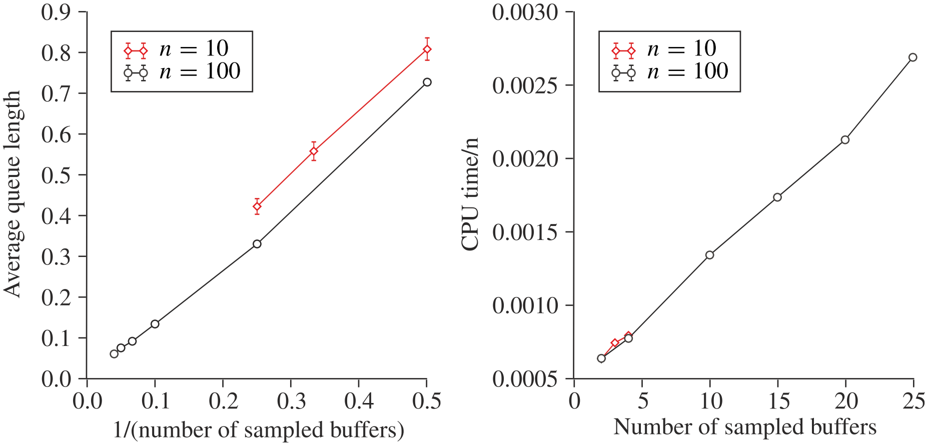

FIGURE 6: Performance versus complexity for $n = 10$, $d = 2, 3, 4$ and for $n = 100$, $d = 2, 4, 10, 15, 20, 25$. *Left:* average queue length versus sample size d. *Right:* CPU time per buffer versus sample size d.

Existing literature exclusively focuses on the performance of algorithms in the mean-field large-buffer regime with the fluid limit, but our experiments highlight that the second-order approximation is also important. Our work is the first investigation in this direction.

4.2. Performance versus complexity

To see the trade-off between performance and complexity, we measure the complexity and performance through CPU-time and average queue length, respectively. For a system with n buffers where the server samples d buffers, the CPU-time consumed during a fixed time is $O(dn)$ and our fluid limit theorem concludes that the average queue length is proportional to $1/d$.

https://doi.org/10.1239/aap/1449859798 Published online by Cambridge University Press

For a fixed number n of buffers in the system, we simulate systems with varying number of sampled buffers d. We run our simulation up to time $t = 50$ with $\lambda = 0.7$ and measure the CPU-time consumption and the average queue length at $t = 50$ for 1000 samples of each case. The results of our experiments are represented graphically in Figure 6.

In Figure 6 we show that CPU-time per buffer (computational complexity) is indeed proportional to the number of sampled buffers d, and that the average queue length (performance) is inversely proportional to the sample size d. Therefore, the simulation study confirms our theoretical results on the quantitative trade-off between performance and complexity.

5. Proofs of the limit theorems

This section provides the proofs of the two theorems in Section 3. Before going into detail, we first introduce the key ideas in the proofs.

We now discuss the starting point of the proofs of our limit theorems, particularly focusing on Theorem 1. Several additional technical tools are needed to fill in the details and we work these out in Sections 5.1–5.3.

Instead of working directly with the random variables $U_{n,k}$, as in [14] we rely on the auxiliary random variables

$$V_{n,k}(t) = \sum_{j=k}^{\infty} F_{n,j}(t) \quad \text{for all } k \geq 0.$$

For $k \geq 1$, $V_{n,k}(\cdot)$ increases by $1/n$ when there is an arrival in queues with length greater than or equal to $k - 1$ and it decreases by $1/n$ if the server processes a task in a queue with length greater than or equal to k. Thus, we have

$$\begin{aligned} V_{n,k}(t) = V_{n,k}(0) &+ \frac{1}{n} A_{n,k}\biggl(\lambda n \int_0^t F_{n,k-1}(s)\,\mathrm{d}s\biggr) \\ &- \frac{1}{n} S_{n,k}\biggl(n \int_0^t [1 - (1 - F_{n,k}(s))^{d(n)}]\,\mathrm{d}s\biggr), \end{aligned} \tag{4}$$

where $A_{n,k}(\cdot)$ and $S_{n,k}(\cdot)$ are independent Poisson processes with rate 1.

Upon multiplying (4) by $d(n)^k$ and rescaling time by a factor $d(n)$, we obtain, after substituting U in place of F,

$$\begin{aligned} d(n)^k V_{n,k}\biggl(\frac{t}{d(n)}\biggr) = d(n)^k V_{n,k}(0) &+ \frac{d(n)^k}{n} A_{n,k}\biggl(\lambda \frac{n}{d(n)^k} \int_0^t U_{n,k-1}(s)\,\mathrm{d}s\biggr) \\ &- \frac{d(n)^k}{n} S_{n,k}\biggl(\frac{n}{d(n)} \int_0^t \biggl[1 - \biggl(1 - \frac{U_{n,k}(s)}{d(n)^k}\biggr)^{d(n)}\biggr]\,\mathrm{d}s\biggr). \end{aligned}$$

Upon replacing $A_{n,k}$ and $S_{n,k}$ by their law-of-large-numbers approximations (the identity function), we obtain

$$\begin{aligned} d(n)^k V_{n,k}\biggl(\frac{t}{d(n)}\biggr) \approx d(n)^k V_{n,k}(0) &+ \lambda \int_0^t U_{n,k-1}(s)\,\mathrm{d}s \\ &- d(n)^{k-1} \int_0^t \biggl[1 - \biggl(1 - \frac{U_{n,k}(s)}{d(n)^k}\biggr)^{d(n)}\biggr]\,\mathrm{d}s, \end{aligned}$$

https://doi.org/10.1239/aap/1449859798 Published online by Cambridge University Press

and a similar 'second-order' representation can be obtained when $A_{n,k}$ and $S_{n,k}$ are replaced by their central limit theorem approximations. For these approximations to be justified, we need $d(n)^k = o(n)$. Continuing with the fluid approximation, since $U_0(t) = 1$, we obtain, for $k = 1$,

$$d(n)^k V_{n,1}\left(\frac{t}{d(n)}\right) \approx d(n)^k V_{n,1}(0) + \lambda t - \int_0^t [1 - \mathrm{e}^{-U_{n,1}(s)}]\,\mathrm{d}s,$$

while for $k \geq 2$, we obtain

$$d(n)^k V_{n,k}\left(\frac{t}{d(n)}\right) \approx d(n)^k V_{n,k}(0) + \lambda \int_0^t U_{n,k-1}(s)\,\mathrm{d}s - \int_0^t U_{n,k}(s)\,\mathrm{d}s.$$

Next we use the following relation between $V_{n,k}(t)$ and $U_{n,k}(t)$:

$$U_{n,k}(t) = d(n)^k V_{n,k}\left(\frac{t}{d(n)}\right) - d(n)^k V_{n,k+1}\left(\frac{t}{d(n)}\right). \tag{5}$$

The second term on the right-hand side of (5) vanishes on the fluid scale, but it has to be taken into account on the diffusion scale.

The above outline is formalized through a mathematical induction argument. The next section is devoted to the induction base for the fluid limit theorem, $k = 1$. In Section 5.2 we consider the induction hypotheses for the fluid limit theorem. In Section 5.3 we address the proof of the diffusion limit theorem.

5.1. Fluid limit: dynamics of the first term

In this section we prove the base of the induction by showing the existence of the fluid limit of $U_{n,1}(t)$ and finding the dynamics of the limit. The strategy of the proof is the following.

1. The proof evolves around the evolution of $d(n)\ V_{n,1}(t/d(n))$ and $d(n)V_{n,2}(t/d(n))$. By definition, we have

$$U_{n,1}(t) = d(n)F_{n,1}\left(\frac{t}{d(n)}\right) = d(n)V_{n,1}\left(\frac{t}{d(n)}\right) - d(n)V_{n,2}\left(\frac{t}{d(n)}\right). \tag{6}$$

2. We prove in Lemma 1 that $d(n)V_{n,2}(t/d(n))$ converges (in an appropriate sense) to the zero function. We then prove in Lemma 2 that $d(n)V_{n,1}(t/d(n))$ has a fluid limit. A key tool in the latter is Lemma 11 from Appendix A, which requires showing that $d(n)V_{n,1}(t/d(n))$ is Lipschitz in some asymptotic sense.

3. We deduce from (6) that the fluid limits of $U_{n,1}(t)$ and $d(n)V_{n,1}(t/d(n))$ are the same. Using (4) and the approach outlined in the previous section, we then formulate the differential equation satisfied by the fluid limit.

First, we prove that $d(n)V_{n,2}(t/d(n))$ converges to 0 uniformly on compact sets for appropriate initial conditions. In particular, it has a fluid limit.

Lemma 1. *Consider a sequence of systems indexed by* n. *Assume that* $\lim_{n\to\infty} d(n)\ V_{n,2}(0) = 0$ *and that* $\lim_{n\to\infty} F_{n,1}(0) = 0$. *Then, we have*

$$\lim_{n\to\infty} d(n)V_{n,2}\left(\frac{t}{d(n)}\right) = 0,$$

uniformly on compact sets, almost surely.

https://doi.org/10.1239/aap/1449859798 Published online by Cambridge University Press

Proof. Let $W_n(\cdot)$ be the process which increases by 1 whenever there is an arrival, a service completion, or the end of a wasted service in the nth system. Note that $W_n(\cdot)$ is a Poisson process with rate $(1+\lambda)n$. For any $t > 0$, the total number of increases of $F_{n,1}(\cdot)$ in $(0, t]$ is less than or equal to $W_n(t)$. Since $F_{n,1}(\cdot)$ increases by $1/n$ at a time, we obtain, for $t > 0$,

$$0 \le F_{n,1}\left(\frac{t}{d(n)}\right) \le F_{n,1}(0) + \frac{1}{n}W_n\left(\frac{t}{d(n)}\right).$$

By our assumption on $F_{n,1}(0)$ and Lemma 8, $F_{n,1}(t/d(n))$ thus converges almost surely to 0 as $n \to \infty$, uniformly on compact sets. From (4), we also deduce that

$$\begin{aligned} d(n)V_{n,2}\left(\frac{t}{d(n)}\right) &\le d(n)V_{n,2}(0) + \frac{d(n)}{n}A_{n,2}\left(\lambda n \int_0^{t/d(n)} F_{n,1}(s)\,\mathrm{d}s\right) \\ &= d(n)V_{n,2}(0) + \frac{d(n)}{n}A_{n,2}\left(\frac{\lambda n}{d(n)} \int_0^{t} F_{n,1}\left(\frac{s}{d(n)}\right)\mathrm{d}s\right). \end{aligned}$$

Upon applying Lemma 5, Lemma 8, and Lemma 10, the second term converges almost surely to 0 as $n \to \infty$, uniformly on compact sets. The claim thus follows from the assumption on $V_{n,2}(0)$.

In the next lemma we prove that, almost surely, $d(n)V_{n,1}(t/d(n))$ satisfies the assumptions of Lemma 11, i.e. that it is Lipschitz in some asymptotic sense. This is a key ingredient in establishing the existence of the fluid limit of $d(n)V_{n,1}(t/d(n))$.

Lemma 2. *Consider a sequence of systems indexed by n. Assume that there is some $v \in \mathbb{R}_+$ such that*

$$\lim_{n\to\infty} d(n)V_{n,1}(0) = v.$$

Then any subsequence of $\{d(n)V_{n,1}(t/d(n))\}_{n\in\mathbb{N}}$ has a subsequence that converges to a Lipschitz function uniformly on compact sets, almost surely.

Proof. Fix $T > 0$, and recall the construction of the Poisson process $W_n(\cdot)$ with rate $(1+\lambda)n$ from the proof of Lemma 1. For $a, b \in [0, T]$ with $a < b$, the total number of increases or decreases of $V_{n,1}(t)$ in $(a, b]$ is less than or equal to $|W_n(a) - W_n(b)|$. Since $d(n)V_{n,1}(\cdot)$ increases or decreases by $d(n)/n$ at a time, there exists some $\gamma_n = \gamma_n(T)$ such that $\lim_{n\to\infty} \gamma_n = 0$ almost surely and

$$\begin{aligned} \left| d(n)V_{n,1}\left(\frac{a}{d(n)}\right) - d(n)V_{n,1}\left(\frac{b}{d(n)}\right)\right| &\le 2\left|\frac{d(n)}{n}W_n\left(\frac{a}{d(n)}\right) - \frac{d(n)}{n}W_n\left(\frac{b}{d(n)}\right)\right| \\ &\le 2(1+\lambda)|a-b| + \gamma_n. \end{aligned}$$

By Lemma 11, any subsequence of $\{d(n)V_{n,k}(t/d(n))\}_{n\in\mathbb{N}}$ has a subsequence that converges to a $2(1+\lambda)$-Lipschitz function uniformly on $[0, T]$ almost surely.

With (6) and the preceding lemmas, we can prove that any subsequence of $\{U_{n,1}(t)\}_{n\in\mathbb{N}}$ has a convergent subsequence which converges to a Lipschitz function $u(t)$. In the next proposition we prove that the limit is independent of the subsequence, so that convergence of $\{U_{n,1}(t)\}_{n\in\mathbb{N}}$ to $u(t)$ on compact sets follows.

Proposition 2. *Consider a sequence of systems indexed by n. Suppose that, for some $v \in \mathbb{R}_+$,*

$$\lim_{n\to\infty} d(n)V_{n,1}(0) = v, \qquad \lim_{n\to\infty} d(n)V_{n,2}(0) = 0 \quad \textit{almost surely.}$$

https://doi.org/10.1239/aap/1449859798 Published online by Cambridge University Press

Then there exists a Lipschitz function $u\colon [0,\infty) \to \mathbb{R}_+$ *such that, almost surely,*

$$\lim_{n\to\infty} U_{n,1}(t) = u(t),$$

uniformly on compact sets and u is the unique solution to the differential equation

$$u'(t) = \mathrm{e}^{-u(t)} - (1-\lambda)$$

with initial value $u(0) = v$. *Also, almost surely,*

$$\lim_{n\to\infty} d(n) V_{n,2}\biggl(\frac{t}{d(n)}\biggr) = 0,$$

uniformly on compact sets.

Proof. By the existence of the limit of $d(n)\ V_{n,1}(0)$, we have $\lim_{n\to\infty} F_{n,1}(0) = 0$. Consider the sequence of bivariate random processes $\{(d(n)V_{n,1}(t/d(n)), U_{n,1}(t))\}_{n\in\mathbb{N}}$. From (6) and the preceding two lemmas, any subsequence has a subsequence which converges uniformly on compact sets, almost surely. Suppose that the convergent subsequence converges to $(u(t), u(t))$ for some Lipschitz function $u\colon [0,\infty) \to \mathbb{R}$.

From (4), we obtain

$$\begin{aligned} d(n)V_{n,1}\biggl(\frac{t}{d(n)}\biggr) &= d(n)V_{n,1}(0) + \frac{d(n)}{n} A_{n,1}\biggl(\lambda n \int_0^{t/d(n)} \mathbf{1}\,\mathrm{d}s\biggr) \\ &\quad - \frac{d(n)}{n} S_{n,1}\biggl(n \int_0^{t/d(n)} [1-(1-F_{n,1}(s))^{d(n)}]\,\mathrm{d}s\biggr) \\ &= d(n)V_{n,1}(0) + \frac{d(n)}{n} A_{n,1}\biggl(\lambda \frac{n}{d(n)} t\biggr) \\ &\quad - \frac{d(n)}{n} S_{n,1}\biggl(\frac{n}{d(n)} \int_0^t \biggl[1 - \biggl(1 - \frac{U_{n,1}(t)}{d(n)}\biggr)^{d(n)}\biggr]\,\mathrm{d}s\biggr). \end{aligned}$$

Thus, letting n go to ∞ along the convergent subsequence, we find that, almost surely, the second term converges to λt uniformly on compact sets by Lemma 8. Moreover, by Lemma 6, Lemma 7, Lemma 8, and Lemma 10, the last term converges almost surely to $\int_0^t (1 - \mathrm{e}^{-u(s)})\,\mathrm{d}s$,, uniformly on compact sets. Therefore, $u(t)$ satisfies the integral equation

$$u(t) = v + \lambda t + \int_0^t (1 - \mathrm{e}^{-u(s)})\,\mathrm{d}s.$$

Since u is absolutely continuous, u is differentiable almost everywhere. If $u(t)$ is differentiable at t, we obtain

$$u'(t) = \mathrm{e}^{-u(t)} - (1-\lambda). \tag{7}$$

By standard existence and uniqueness theorems for ordinary differential equations, there is a unique solution $u\colon [0,\infty) \to \mathbb{R}_+$ satisfying the above differential equation (7) with initial condition $u(0) = v$. Thus, every subsequence of $\{U_{n,1}(t)\}_{n\in\mathbb{N}}$ has a subsequence which converges to the same limit $u(t)$. Therefore, $\{U_{n,1}(t)\}_{n\in\mathbb{N}}$ converges to $u(t)$ uniformly on compact sets, almost surely.

https://doi.org/10.1239/aap/1449859798 Published online by Cambridge University Press

5.2. Fluid limit: dynamics of higher terms

In this section we state and prove the induction step. Let $k \geq 1$ and assume throughout that $\lim_{n\to\infty} n/d(n)^{k+1} = \infty$. We work under the induction hypothesis that there exists a Lipschitz continuous function $u_k \colon [0,\infty) \to \mathbb{R}_+$ such that

$$\lim_{n\to\infty} U_{n,k}(t) = u_k(t), \tag{8}$$

uniformly on compact sets, almost surely, and

$$\lim_{n\to\infty} d(n)^k V_{n,k+1}\left(\frac{t}{d(n)}\right) = 0, \tag{9}$$

uniformly on compact sets, almost surely. Starting from this hypothesis, we prove the existence of the fluid limit of $U_{n,k+1}(t)$ and characterize it through a differential equation.

The proof roughly follows the same outline as for the dynamics of the first term in Section 5.1, i.e. we first establish the existence of the fluid limits and then use (4) to establish the differential equations they satisfy. The details, however, are different; for instance, we must avoid a circular argument for establishing an asymptotic Lipschitz property of $d(n)^{k+1} V_{n,k+1}(t/d(n))$ (Lemma 4), an issue that did not arise in Section 5.1.

Lemma 3. *Consider a sequence of systems indexed by n, for which (8) and (9) hold. Assume that*

$$\lim_{n\to\infty} d(n)^{k+1} V_{n,k+2}(0) = 0,$$

almost surely. Then, we have

$$\lim_{n\to\infty} d(n)^{k+1} V_{n,k+2}\left(\frac{t}{d(n)}\right) = 0,$$

uniformly on compact sets, almost surely.

Proof. By (4), we have

$$\begin{aligned} &d(n)^{k+1} V_{n,k+2}\left(\frac{t}{d(n)}\right) \\ &\quad \leq d(n)^{k+1} V_{n,k+2}(0) + \frac{d(n)^{k+1}}{n} A_{n,k+2}\left(\lambda n \int_0^{t/d(n)} F_{n,k+1}(s)\,\mathrm{d}s\right) \\ &\quad = d(n)^{k+1} V_{n,k+2}(0) + \frac{d(n)^{k+1}}{n} A_{n,k+2}\left(\lambda \frac{n}{d(n)^{k+1}} \int_0^{t} d(n)^k F_{n,k+1}\left(\frac{s}{d(n)}\right) \mathrm{d}s\right). \end{aligned}$$

Hypothesis (9) implies that $\lim_{n\to\infty} d(n)^k F_{n,k+1}(t/d(n)) = 0$ almost surely, uniformly on compact sets. Thus, by Lemma 5, Lemma 8, and Lemma 10, we obtain, almost surely,

$$\lim_{n\to\infty} d(n)^{k+1} V_{n,k+2}\left(\frac{t}{d(n)}\right) = 0,$$

uniformly on compact sets.

To show the existence of the fluid limit of $d(n)^{k+1} V_{n,k+1}(t/d(n))$, we need to prove that it is Lipschitz in some asymptotic sense, cf. Lemma 11. For the $k = 0$ case, we used a scaled version of a Poisson process $W_n(t)$ to prove this for $d(n)V_{n,1}(t/d(n))$. However, when $k \geq 1$, a similar modification of $W_n(t)$ does not work for $d(n)^{k+1}V_{n,k+1}(t/d(n))$ since $d(n)^{k+1}W_n(t/d(n))$ diverges for $k > 0$. We resolve this difficulty by partitioning an expression for $d(n)^{k+1}V_{n,k+1}(t/d(n))$ into three parts—an initial part, an arrival part, and a departure part; see (4). Assuming the existence of a limit for the initial part, we then show that the other two parts admit fluid limits.

As we shall see, the arrival part depends on $U_{n,k}(t)$ and the induction hypothesis guarantees its convergence. Thus, the existence of the fluid limit of the arrival part follows immediately. We cannot directly apply the induction hypothesis for the departure part because it turns out to involve $U_{n,k+1}(t)$, the very quantity we are trying to establish a fluid limit for. To circumvent this issue, we show that $U_{n,k+1}(t)$ is locally bounded and this allows us to show that the departure part is Lipschitz continuous in the sense of Lemma 11.

Lemma 4. *Consider a sequence of systems indexed by n, for which (8) and (9) hold. Suppose that there exists some $v \in \mathbb{R}_+$ such that $\lim_{n\to\infty} d(n)^{k+1}V_{n,k+1}(0) = v$, almost surely. Then any subsequence of $\{d(n)^{k+1}V_{n,k+1}(t/d(n))\}_{n\in\mathbb{N}}$ has a subsequence which converges almost surely to a Lipschitz continuous function uniformly on compact sets.*

Proof. Fix $T > 0$. Decompose $d(n)^{k+1}V_{n,k+1}(t/d(n))$ into three parts as follows:

$$d(n)^{k+1}V_{n,k+1}\left(\frac{t}{d(n)}\right) = d(n)^{k+1}V_{n,k+1}(0) + I_n(t) - D_n(t),$$

where $I_n(t)$ and $D_n(t)$ are the total increase and decrease amount of $d(n)^{k+1}V_{n,k+1}(t/d(n))$ by time t, respectively.

The almost sure limit of $I_n(t)$ is readily found. Indeed, from (4), we have

$$\begin{aligned} I_n(t) &= \frac{d(n)^{k+1}}{n} A_{n,k+1}\left(\lambda n \int_0^{t/d(n)} F_{n,k}(s)\,\mathrm{d}s\right) \\ &= \frac{d(n)^{k+1}}{n} A_{n,k+1}\left(\frac{n}{d(n)^{k+1}} \int_0^t U_{n,k}(s)\,\mathrm{d}s\right), \end{aligned}$$

which converges almost surely to $\int_0^t u_k(s)\,\mathrm{d}s$ uniformly on $[0, T]$ by Lemma 5 and Lemma 10.

Proving the almost sure limit of $D_n(t)$ is more intricate. From (4), we obtain

$$\begin{aligned} D_n(t) &= \frac{d(n)^{k+1}}{n} S_{n,k+1}\left(n \int_0^{t/d(n)} (1-(1-F_{n,k+1}(s))^{d(n)})\,\mathrm{d}s\right) \\ &= \frac{d(n)^{k+1}}{n} S_{n,k+1}\left(\frac{n}{d(n)} \int_0^t \left(1-\left(1-F_{n,k+1}\left(\frac{s}{d(n)}\right)\right)^{d(n)}\right)\mathrm{d}s\right) \\ &= \frac{d(n)^{k+1}}{n} S_{n,k+1}\left(\frac{n}{d(n)^{k+1}} \int_0^t d(n)^k\left[1-\left(1-\frac{U_{n,k+1}(s)}{d(n)^{k+1}}\right)^{d(n)}\right]\mathrm{d}s\right). \end{aligned} \qquad (10)$$

The first step for analyzing this expression is to bound the integrand. Write $M = \sup_{t\in[0,T]} \int_0^t u_k(s)\,\mathrm{d}s$ and let $\varepsilon > 0$. Then, for all $t \in [0, T]$ and large enough n, we have

$$U_{n,k+1}(t) \leq d(n)^{k+1}V_{n,k+1}\left(\frac{t}{d(n)}\right) \leq d(n)^{k+1}V_{n,k+1}(0) + I_n(t) \leq v + M + \varepsilon.$$

https://doi.org/10.1239/aap/1449859798 Published online by Cambridge University Press

Thus, for all large enough n, we have, almost surely,

$$d(n)^k\left[1-\left(1-\frac{U_{n,k+1}(t)}{d(n)^{k+1}}\right)^{d(n)}\right] \leq d(n)^k\left[1-\left(1-\frac{v+M+\varepsilon}{d(n)^{k+1}}\right)^{d(n)}\right]$$
$$\leq v+M+2\varepsilon \quad \text{for all } t \in [0,T].$$

Lemma 8 implies that, almost surely,

$$\lim_{n\to\infty} \sup_{a,b\in[0,(v+M+2\varepsilon)T]} \left| \frac{d(n)^{k+1}}{n} S_{n,k+1}\left(\frac{n}{d(n)^{k+1}}b\right) - \frac{d(n)^{k+1}}{n} S_{n,k+1}\left(\frac{n}{d(n)^{k+1}}a\right) - (b-a)\right| = 0,$$

which by (10) shows that $\lim_{n\to\infty}\gamma_n = 0$ almost surely, where

$$\gamma_n = \sup_{0\leq s<t\leq T}\left| D_n(t)-D_n(s) - \int_s^t d(n)^k\left[1-\left(1-\frac{U_{n,k+1}(u)}{d(n)^{k+1}}\right)^{d(n)}\right]\mathrm{d}u\right|.$$

We next note that, for $a, b \in [0,T]$,

$$|D_n(a)-D_n(b)| \leq (v+M+2\varepsilon)|a-b|+\gamma_n.$$

Thus, by Lemma 11, any subsequence of $\{D_{n,k}(\cdot)\}$ has a subsequence that converges to a Lipschitz continuous function. Therefore, any subsequence of $\{d(n)^{k+1}V_{n,k+1}(t/d(n))\}_{n\in\mathbb{N}}$ has a subsequence converging to a Lipschitz continuous function uniformly on $[0,T]$, almost surely.

By the preceding two lemmas, any subsequence of $\{U_{n,k+1}(t)\}_{n\in\mathbb{N}}$ has a subsequence which converges almost surely to a Lipschitz function uniformly on compact sets. We prove the induction step through the same argument used in the induction base.

Proposition 3. *Consider a sequence of systems indexed by n, for which the induction hypothesis (8) and (9) hold. Assume that there exists some $v \in \mathbb{R}_+$ such that $\lim_{n\to\infty} d(n)^{k+1}V_{n,k+1}(0) = v$, almost surely, and $\lim_{n\to\infty} d(n)^{k+1}V_{n,k+2}(0) = 0$. Then the sequence $\{U_{n,k+1}(t)\}_{n\in\mathbb{N}}$ converges almost surely to the unique Lipschitz function $u_{k+1}\colon [0,\infty) \to \mathbb{R}_+$ satisfying*

$$u'_{k+1}(t) = \lambda u_k(t) - u_{k+1}(t),$$

with $u(0) = v$, uniformly on compact sets. Moreover, we have

$$\lim_{n\to\infty} d(n)^{k+1}F_{n,k+2}\left(\frac{t}{d(n)}\right) = 0,$$

uniformly on compact sets.

Proof. Consider the sequence of coupled random processes

$$\left\{\left(d(n)^{k+1}V_{n,k+1}\left(\frac{t}{d(n)}\right), U_{n,k+1}(t)\right)\right\}_{n\in\mathbb{N}}.$$

https://doi.org/10.1239/aap/1449859798 Published online by Cambridge University Press

By the preceding lemmas, any subsequence has a subsequence which converges uniformly on compact sets, almost surely. Moreover, the convergent subsequence converges to $(u_{k+1}(t), u_{k+1}(t))$ for some Lipschitz function $u_{k+1}(t)$.

From (4), we obtain

$$\begin{aligned}
&d(n)^{k+1} V_{n,k+1}\left(\frac{t}{d(n)}\right)\\
&\quad = d(n)^{k+1} V_{n,k+1}(0) + \frac{d(n)^{k+1}}{n} A_{n,k+1}\left(\lambda n \int_0^{t/d(n)} F_{n,k}(s)\,\mathrm{d}s\right)\\
&\qquad - \frac{d(n)^{k+1}}{n} S_{n,k+1}\left(n \int_0^{t/d(n)} (1-(1-F_{n,k+1}(s))^{d(n)})\,\mathrm{d}s\right)\\
&\quad = d(n)^{k+1} V_{n,k+1}(0) + \frac{d(n)^{k+1}}{n} A_{n,k+1}\left(\frac{\lambda n}{d(n)^{k+1}} \int_0^t U_{n,k}(s)\,\mathrm{d}s\right)\\
&\qquad - \frac{d(n)^{k+1}}{n} S_{n,k+1}\left(\frac{n}{d(n)^{k+1}} \int_0^t d(n)^k\left(1-\left(1-\frac{U_{n,k+1}(s)}{d(n)^{k+1}}\right)^{d(n)}\right)\mathrm{d}s\right).
\end{aligned}$$

From Lemma 6, Lemma 7, Lemma 8, and Lemma 10, and by taking the limit as $n \to \infty$ along the convergent subsequence, we conclude that $u_{k+1}(t)$ satisfies

$$u_{k+1}(t) = v + \lambda \int_0^t u_k(s)\,\mathrm{d}s - \int_0^t u_{k+1}(s)\,\mathrm{d}s.$$

Since $u_{k+1}(t)$ is absolutely continuous, $u_{k+1}(t)$ is differentiable almost everywhere. If $u_{k+1}(t)$ is differentiable at t, we obtain

$$u'_{k+1}(t) = \lambda u_k(t) - u_{k+1}(t). \tag{11}$$

Since (11) is linear with inhomogeneous term $\lambda u_k(t)$, it uniquely determines $u_{k+1}(t)$. Thus, every sequence of $\{U_{n,k+1}(t)\}_{n\in\mathbb{N}}$ has a subsequence that converges to the same limit $u_{k+1}(t)$. Therefore, $U_{n,k+1}(t)$ converges to $u_{k+1}(t)$ uniformly on compact sets, almost surely.

The last statement of the proposition follows from Lemma 3.

Using Proposition 2 and Proposition 3, we are now ready to prove our fluid limit theorem.

Proof of Theorem 1. From the assumptions of Theorem 1, we have

$$\lim_{n\to\infty} U_{n,1}(0) = v_1$$

and

$$\lim_{n\to\infty} d(n)V_{n,2}(0) = \lim_{n\to\infty}\left(\frac{U_{n,2}(0)}{d(n)} + \cdots + \frac{U_{n,K}(0)}{d(n)^{K-1}} + \frac{d(n)^K(F_{n,K+1}(0)+\cdots)}{d(n)^{K-1}}\right) = 0.$$

Therefore, Proposition 2 yields the fluid limit for $U_{n,1}(t)$, which is (8) for $k=1$. Lemma 1 yields (9) for $k=1$.

We next assume that (8) and (9) hold. The assumptions in Proposition 3 hold because of the assumptions from Theorem 1, as can be seen with a similar argument as above. Thus, Proposition 3 and Lemma 3 show that (8) and (9) hold, respectively, with k replaced by $k+1$. This induction argument establishes Theorem 1.

https://doi.org/10.1239/aap/1449859798 Published online by Cambridge University Press

5.3. Diffusion limit

In this section we prove our second limit theorem, Theorem 2, a diffusion limit of $U_{n,1}(t)$. To this end, we introduce a new sequence of stochastic processes with the same fluid limit $u_1(t)$ as $\{U_{n,1}(t)\}_{n\in\mathbb{N}}$. For this new sequence, we can apply a result from Kurtz [11] to obtain its second-order approximation. We then compare the new processes with $\{U_{n,1}(t)\}_{n\in\mathbb{N}}$ and show that the difference vanishes.

Proof of Theorem 2. From (4), we have

$$U_{n,1}(t) = -d(n)V_{n,2}\bigg(\frac{t}{d(n)}\bigg) + V_{n,1}(0) + \frac{d(n)}{n}A_{n,1}\bigg(\frac{\lambda n}{d(n)}t\bigg) \\ - \frac{d(n)}{n}S_{n,1}\bigg(\frac{n}{d(n)}\int_0^t \bigg[1-\bigg(1-\frac{U_{n,1}(s)}{d(n)}\bigg)^{d(n)}\bigg]\,\mathrm{d}s\bigg). \tag{12}$$

Let $\lim_{n\to\infty} n/d(n) = \infty$ and $\lim_{n\to\infty} n/d(n)^2 = 0$ and assume that $U_{n,k}(0)$ for all n and k, and $v_1 \in \mathbb{R}_+$ satisfies (1) and (2) in Theorem 2.

Define a sequence of stochastic processes $\{\widehat{U}_n(t)\}$ as the unique solution to

$$\widehat{U}_n(t) = v_1 + \frac{d(n)}{n}A_{n,1}\bigg(\frac{n}{d(n)}\int_0^t f_{n,1}(\widehat{U}_n(s))\,\mathrm{d}s\bigg) \\ - \frac{d(n)}{n}S_{n,1}\bigg(\frac{n}{d(n)}\int_0^t f_{n,-1}(\widehat{U}_n(s))\,\mathrm{d}s\bigg), \tag{13}$$

where $f_{n,1} = \lambda$ and

$$f_{n,-1}(x) = \begin{cases} 1-\bigg(1-\dfrac{x}{d(n)}\bigg)^{d(n)} & \text{if } 0 \le x \le d(n), \\ 1-\mathrm{e}^{-x}+\mathrm{e}^{-d(n)} & \text{otherwise.} \end{cases}$$

The process $\widehat{U}_n(t)$ is coupled with $U_{n,1}(t)$. We next argue that $\widehat{U}_n(t)$ has a fluid and diffusion approximation prescribed by the theory developed by Kurtz [11, Lemma 12, Appendix A]. Note that the index in [11] is $N = n/d(n)$ and n can often also be expressed in terms of N. This cannot always be done, but we suppress the arguments needed to deal with such cases.

Let $f_1(x) = \lambda$ and $f_{-1}(x) = 1-\mathrm{e}^{-x}$. After noting that the maximum of $m(\mathrm{e}^{-x} - (1 - x/m)^m)$ over $0 \le x \le m$ converges to 2 as $m \to \infty$, we have, for large enough n,

$$|f_{n,-1}(x) - f_{-1}(x)| \le \frac{3}{d(n)} \le 3\frac{d(n)}{n}.$$

Thus, all conditions from Lemma 12 are satisfied and $\widehat{U}_n(t)$ converges almost surely to $u_1(t)$ uniformly on compact sets, and we have the second-order approximation of $\widehat{U}_n(t)$ such that

$$\sqrt{\frac{n}{d(n)}}(\widehat{U}_n(t) - u_1(t)) \xrightarrow{\mathrm{D}} Z(t), \tag{14}$$

where $Z(t)$ satisfies

$$Z(t) = \sqrt{\lambda}B^{(1)}(t) - \int_0^t \sqrt{1-\mathrm{e}^{-u_1(s)}}\,\mathrm{d}B^{(2)}(s) - \int_0^t \mathrm{e}^{-u_1(s)}Z(s)\,\mathrm{d}s$$

for independent Wiener processes $B^{(1)}(t)$ and $B^{(2)}(t)$. We note that the results in [11] yield strong approximations; here we only use weaker results of convergence in distribution.

We next compare $U_{n,1}(t)$ with $\widehat{U}_n(t)$ and show that $\sqrt{n/d(n)}|U_{n,1}(t) - \widehat{U}_n(t)| \xrightarrow{\mathrm{D}} 0$. Fix some $T > 0$. From (12) and (13), we have, since $0 \le U_{n,1}(t) \le d(n)$ and $f_{n,-1}(t)$ is 1-Lipschitz continuous,

$$\begin{aligned}\sqrt{\frac{n}{d(n)}}|U_{n,1}(t) - \widehat{U}_n(t)| &\le \sqrt{\frac{n}{d(n)}}(V_{n,1}(0) - v_1) + \sqrt{nd(n)}V_{n,2}\bigg(\frac{t}{d(n)}\bigg) \\ &\quad + \bigg|\widetilde{S}_n\bigg(\int_0^t f_{n,-1}(U_{n,1}(s))\,\mathrm{d}s\bigg) - \widetilde{S}_n\bigg(\int_0^t f_{n,-1}(\widehat{U}_n(s))\,\mathrm{d}s\bigg)\bigg| \\ &\quad + \sqrt{\frac{n}{d(n)}}\int_0^t |f_{n,-1}(U_{n,1}(s)) - f_{n,-1}(\widehat{U}_n(s))|\,\mathrm{d}s \\ &\le \varepsilon_n(t) + \int_0^t \sqrt{\frac{n}{d(n)}}|U_{n,1}(s) - \widehat{U}_n(s)|\,\mathrm{d}s,\end{aligned}$$

where

$$\widetilde{S}_n(t) = \sqrt{\frac{n}{d(n)}}\bigg(\frac{d(n)}{n}S_{n,1}\bigg(\frac{n}{d(n)}t\bigg) - t\bigg)$$

and

$$\begin{aligned}\varepsilon_n(t) = {}&\sqrt{\frac{n}{d(n)}}(V_{n,1}(0) - v_1) + \sqrt{nd(n)}V_{n,2}\bigg(\frac{t}{d(n)}\bigg) \\ &+ \bigg|\widetilde{S}_n\bigg(\int_0^t f_{n,-1}(U_{n,1}(t))\,\mathrm{d}s\bigg) - \widetilde{S}_n\bigg(\int_0^t f_{n,-1}(\widehat{U}_{n,1}(t))\,\mathrm{d}s\bigg)\bigg|.\end{aligned}$$

By Gronwall's inequality, we obtain, for $t \in [0, T]$,

$$\sqrt{\frac{n}{d(n)}}|U_{n,1}(t) - \widehat{U}_n(t)| \le \varepsilon_n(t) + \mathrm{e}^t \int_0^t \varepsilon_n(t)\,\mathrm{d}s \le L \sup_{t\in[0,T]} \varepsilon_n(t),$$

where $L = 1 + T\mathrm{e}^T$.

We proceed by showing that $\varepsilon_n(t) \xrightarrow{\mathrm{D}} 0$. From (4), we find that

$$\sqrt{nd(n)}V_{n,2}\bigg(\frac{t}{d(n)}\bigg) \le \sqrt{nd(n)}V_{n,2}(0) + \sqrt{nd(n)}S_{n,2}\bigg(\frac{n}{d(n)^2}\int_0^t U_{n,1}(s)\,\mathrm{d}s\bigg),$$

which converges to 0 almost surely as $n \to \infty$ uniformly on compact sets, by (2), Lemma 5, and Lemma 10 with $\lim_{n\to\infty} n/d(n)^2 = 0$. Also, from Lemma 9 and Lemma 10, we deduce that

$$\begin{aligned}&\bigg(\widetilde{S}_n\bigg(\int_0^t f_{n,-1}(U_{n,1}(s))\,\mathrm{d}s\bigg), \widetilde{S}_n\bigg(\int_0^t f_{n,-1}(\widehat{U}_n(s))\,\mathrm{d}s\bigg)\bigg) \\ &\qquad \xrightarrow{\mathrm{D}} \bigg(B\bigg(\int_0^t [1 - \mathrm{e}^{-u_1(s)}]\,\mathrm{d}s\bigg), B\bigg(\int_0^t [1 - \mathrm{e}^{-u_1(s)}]\,\mathrm{d}s\bigg)\bigg) \quad \text{as } n \to \infty,\end{aligned}$$

where B is a standard Wiener process. By the continuous mapping theorem, we conclude that, as $n \to \infty$,

$$\varepsilon_n(t) \xrightarrow{\mathrm{D}} 0$$

https://doi.org/10.1239/aap/1449859798 Published online by Cambridge University Press

and, therefore,

$$\sqrt{\frac{n}{d(n)}}\,(U_{n,1}(t)-\widehat{U}_n(t)) \xrightarrow{\mathrm{D}} 0.$$

From (14), we conclude that, as $n \to \infty$,

$$\sqrt{\frac{n}{d(n)}}\,(U_{n,1}(t)-u_1(t)) \xrightarrow{\mathrm{D}} Z(t),$$

as claimed.

Appendix A.

In this appendix we review elements of convergence theory of functions and stochastic processes.

For fixed $T > 0$, $D^k[0, T]$ is the space of functions from $[0, T]$ to $\mathbb{R}^k$ that are right-continuous with left-limits equipped with the norm

$$\|f\|_T := \sup_{0\le t\le T} \|f(t)\|_\infty$$

and the associated topology of uniform convergence. We define $D^k[0, \infty)$ similarly, and we equip it with the product metric (of convergence on compact sets) and its associated topology.

We interpret a stochastic process X in this context as a measurable mapping from a probability space $(\Omega, \mathcal{F}, \mathbb{P})$ to $D^k[0, \infty)$. For a sequence $\{X_n\}_{n\in\mathbb{N}}$ of stochastic processes and a stochastic process X, we say that $\{X_n\}_{n\in\mathbb{N}}$ converges almost surely to X uniformly on compact sets if

$$\mathbb{P}\Big(\lim_{n\to\infty} \|X_n - X\|_T = 0\Big) = 1 \quad \text{for all } T > 0.$$

For a stochastic process X, we can define a probability measure $\mathbb{P}_X$ on $D^k[0, T)$ for any $T > 0$. We say that a sequence $\{X_n\}_{n\in\mathbb{N}}$ of stochastic processes converges in distribution to a stochastic process X if, for all $T > 0$,

$$\lim_{n\to\infty} \int_{D^k[0,T]} f \,\mathrm{d}\mathbb{P}_{X_n} = \int_{D^k[0,T]} f \,\mathrm{d}\mathbb{P}_X$$

for every bounded and continuous real-valued function f on $D^k[0, T]$. We abbreviate this to

$$X_n \xrightarrow{\mathrm{D}} X \quad \text{as } n \to \infty.$$

The following lemmas contain results about convergence of functions that are needed to prove our theorems. The first three lemmas are basic results about uniform convergence on compact sets. The proof of the third lemma can be found in [9].

Lemma 5. *Let $\{f_n\}_{n\in\mathbb{N}}$ be a sequence of real-valued functions defined on $[0, \infty)$ and assume that it converges to a function $f\colon [0, \infty) \to \mathbb{R}$ uniformly on compact sets. Assume that the functions $F_n\colon [0, \infty) \to \mathbb{R}$ with $F_n(t) = \int_0^t f_n(s)\,\mathrm{d}s$ and $F\colon [0, \infty) \to \mathbb{R}$ with $F(t) = \int_0^t f(s)\,\mathrm{d}s$ are well-defined. Then, as $n \to \infty$, $\{F_n\}_{n\in\mathbb{N}}$ converges to F uniformly on compact sets.*

Lemma 6. *Let $\{f_n\}_{n\in\mathbb{N}}$ and $\{g_n\}_{n\in\mathbb{N}}$ be two sequences of real-valued functions defined on $[0, \infty)$. Assume that g_n is nonnegative. If, as $n \to \infty$, $\{f_n\}_{n\in\mathbb{N}}$ and $\{g_n\}_{n\in\mathbb{N}}$ converge uniformly on compact sets to real-valued functions f and g, respectively, and f and g are continuous, then, as $n \to \infty$, the sequence $\{f_n(g_n)\}_{n\in\mathbb{N}}$ converges to $f(g)$ uniformly on compact sets.*

https://doi.org/10.1239/aap/1449859798 Published online by Cambridge University Press

Proof. Fix $T > 0$ and $\varepsilon > 0$. Since g is continuous on $[0, T]$, there exists $M > 0$ such that $|g(t)| \leq M$ for all $t \in [0, T]$. Since f is continuous on $[0, M+1]$, there exists $0 < \delta < 1$ such that, for $s, t \in [0, M+1]$, $|t-s| < \delta$ implies that $|f(t)-f(s)| \leq \varepsilon/2$. Let $L = \max\{T, M+1\}$.

From the fact that $f_n \to f$ and $g_n \to g$ as $n \to \infty$ uniformly on compact sets, there exists some $N \in \mathbb{N}$ such that $n \geq N$ implies that $|f_n(t) - f(t)| \leq \min\{\varepsilon/2, \delta\}$ and $|g_n(t) - g(t)| \leq \min\{\varepsilon/2, \delta\}$ for all $t \in [0, L]$. Then, for all $t \in [0, T]$ and $n \geq N$, we have

$$|g_n(t)| \leq |g_n(t) - g(t)| + |g(t)| \leq 1 + M.$$

Thus, if $n \geq N$, we have

$$|f_n(g_n(t)) - f(g(t))| \leq |f_n(g_n(t)) - f(g_n(t))| + |f(g_n(t)) - f(g(t))| < \varepsilon$$

for all $t \in [0, T]$. Therefore, $f_n(g_n)$ converges to $f(g)$ as $n \to \infty$ uniformly on compact sets.

Lemma 7. *Let $\{f_n\}_{n\in\mathbb{N}}$ be a sequence of nondecreasing real-valued functions on $[0, \infty)$ and let f be a continuous function on $[0, \infty)$. Assume that $\lim_{n\to\infty} f_n(t) = f(t)$ for all rational numbers $t \in [0, \infty)$. Then $\{f_n\}_{n\in\mathbb{N}}$ converges to f as $n \to \infty$ uniformly on compact sets.*

The next lemmas are the functional law of large numbers and the functional central limit theorem for Poisson processes; see, for example, [8].

Lemma 8. (Functional law of large numbers.) *Let A be a Poisson process with rate λ. Then, as $n \to \infty$, we have, almost surely,*

$$\frac{1}{n}A(nt) \to \lambda t,$$

uniformly on compact sets. Also, if $f(n) = o(n)$ and $\lim_{n\to\infty} f(n) = \infty$, we have, almost surely,

$$\frac{1}{n}A\left(\frac{n}{f(n)}t\right) \to 0 \quad \text{as } n \to \infty,$$

uniformly on compact sets.

Lemma 9. (Functional central limit theorem.) *Let A be a Poisson process with rate 1. Then, as $n \to \infty$,*

$$\sqrt{n}\left(\frac{1}{n}A(nt) - t\right) \xrightarrow{\text{D}} B(t),$$

where $B(t)$ is the standard Wiener process.

The following lemma is often called the random time-change theorem; see, for example, [8].

Lemma 10. (Random time-change theorem.) *Let $\{f_n\}_{n\in\mathbb{N}}$ and $\{g_n\}_{n\in\mathbb{N}}$ be two sequences in $D^k[0, \infty)$. Assume that each component of g_n is nondecreasing with $g_n(0) = 0$. If, as $n \to \infty$, (f_n, g_n) converges uniformly on compact sets to (f, g) and f and g are continuous, then*

$$\lim_{n\to\infty} f_n(g_n) \to f(g),$$

uniformly on compact sets, where the ith component of $f(g)$ is the composition of the ith component of f and the ith component of g.

https://doi.org/10.1239/aap/1449859798 Published online by Cambridge University Press

The next lemma can be used to show the existence of a fluid limit of a sequence of stochastic processes. Intuitively, if the fluctuations of a sequence of functions are asymptotically bounded by the fluctuations of a Lipschitz function, then any subsequence has a convergent subsequence which converges to a Lipschitz function. This lemma immediately follows from arguments in [14, Appendix A].

Lemma 11. *Fix $T > 0$. Let $\{f_n\}_{n\in\mathbb{N}}$ be a sequence in $D[0, T]$. Assume that $|f_n(0)| \le M$ and*

$$|f_n(a) - f_n(b)| \le L|a - b| + \gamma_n \quad \textit{for all } a, b \in [0, T],$$

for constants M, L, and a sequence $\gamma_n \downarrow 0$. Then any subsequence of $\{f_n\}_{n\in\mathbb{N}}$ has a subsequence that converges to an L-Lipschitz function f uniformly on $[0, T]$ with $|f(0)| \le M$.

The next lemma is used to prove Theorem 2. Kurtz [11] derived diffusion approximations for a variety of continuous Markov chains and the following lemma is a special case of the results in [11]. We use it to obtain the diffusion limit of $\{\widehat{U}_n(t)\}_{n\in\mathbb{N}}$ in the proof of Theorem 2.

Lemma 12. *Consider a sequence of real-valued Markov processes $\{U_N(t)\}_{N\in\mathbb{N}}$ which satisfies*

$$U_N(t) = u_0 + \frac{1}{N}A_N\bigg(N\int_0^t f_{N,1}(U_N(s))\,\mathrm{d}s\bigg) - \frac{1}{N}S_N\bigg(N\int_0^t f_{N,-1}(U_N(s))\,\mathrm{d}s\bigg),$$

where $A_N(\cdot)$ and $S_N(\cdot)$ are independent Poisson processes with rate 1, and $f_{N,i}$ are positive valued continuous functions for $i = \pm 1$. Suppose that there exists a constant $M > 0$ and functions f_1 and f_{-1} such that

$$f_{N,i}(x) \le M, \qquad |f_{N,i}(x) - f_i(x)| \le \frac{M}{N}, \qquad |\sqrt{f_i(x)} - \sqrt{f_i(y)}|^2 \le M|x-y|^2 \quad \textit{for } i = \pm 1.$$

Let $F(x) = f_1(x) - f_{-1}(x)$ and also assume that $|F'(x)| \le M$ and $|F''(x)| \le M$. Then, we have

$$\sqrt{N}(U_N(t) - u(t)) \xrightarrow{\mathrm{D}} V(t),$$

where $u(t)$ is a function satisfying

$$u(t) = u_0 + \int_0^t f_1(u(s))\,\mathrm{d}s - \int_0^t f_{-1}(u(s))\,\mathrm{d}s$$

and $V(t)$ is a stochastic process given by

$$V(t) = \int_0^t \sqrt{f_1(u(s))}\,\mathrm{d}B^{(1)}(s) - \int_0^t \sqrt{f_{-1}(u(s))}\,\mathrm{d}B^{(2)}(s) + \int_0^t F'(u(s))V(s)\,\mathrm{d}s,$$

where $B^{(1)}(t)$ and $B^{(2)}(t)$ are independent Wiener processes.

Acknowledgements

This research was supported in part by The National Science Foundation (grant no. CMMI-1252878). We thank Ilyas Iyoob for fruitful discussions.

https://doi.org/10.1239/aap/1449859798 Published online by Cambridge University Press

References

[1] Alanyali, M. and Dashouk, M. (2011). Occupancy distributions of homogeneous queueing systems under opportunistic scheduling. *IEEE Trans. Inf. Theory* **57,** 256–266.

[2] Bakhshi, R., Cloth, L., Fokkink, W. and Haverkort, B. (2011). Mean-field analysis for the evaluation of gossip protocols. In *Quantitative Evaluation of Systems*, IEEE, New York, pp. 247–256.

[3] Benaïm, M. and Le Boudec, J.-Y. (2008). A class of mean field interaction models for computer and communication systems. *Performance Evaluation* **65,** 823–838.

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

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

[6] Bramson, M., Lu, Y. and Prabhakar, B. (2013). Decay of tails at equilibrium for FIFO join the shortest queue networks. *Ann. Appl. Prob.* **23,** 1841–1878.

[7] Braun, M. (1993). *Differential Equations and Their Applications. An Introduction to Applied Mathematics*, 4th edn. Springer, New York.

[8] Chen, H. and Yao, D. D. (2001). *Fundamentals of Queueing Networks: Performance, Asymptotics, and Optimization*. Springer, New York.

[9] Dai, J. G. (1995). On positive Harris recurrence of multiclass queueing networks: a unified approach via fluid limit models. *Ann. Appl. Prob.* **5,** 49–77.

[10] Gast, N. and Gaujal, B. (2010). A mean field model of work stealing in large-scale systems. *ACM SIGMETRICS Performance Eval. Rev.* **38,** 13–24.

[11] Kurtz, T. G. (1977/78). Strong approximation theorems for density dependent Markov chains. *Stoch. Process. Appl.* **6,** 223–240.

[12] Le Boudec, J.-Y., McDonald, D. and Mundinger, J. (2007). A generic mean field convergence result for systems of interacting objects. In *Proc. Fourth International Conference on the Quantitative Evaluation of Systems*, pp. 3–18.

[13] Mitzenmacher, M. D. (1996). The power of two choices in randomized load balancing. Doctoral thesis. University of California, Berkeley.

[14] Tsitsiklis, J. N. and Xu, K. (2012). On the power of (even a little) resource pooling. *Stoch. Systems* **2,** 1–66.

[15] Van Houdt, B. (2013). Performance of garbage collection algorithms for flash-based solid state drives with hot/cold data. *Performance Evaluation* **70,** 692–703.

[16] Vvedenskaya, N. D., Dobrushin, R. L. and Karpelevich, F. I. (1996). A queueing system with a choice of the shorter of two queues—an asymptotic approach. *Problems Inf. Transmission* **32,** 15–27.

https://doi.org/10.1239/aap/1449859798 Published online by Cambridge University Press