REPORT DOCUMENTATION PAGE

Form Approved OMB NO. 0704-0188

1. REPORT DATE (DD-MM-YYYY)	2. REPORT TYPE	3. DATES COVERED (From - To)
31-08-2015	Conference Proceeding	-

4. TITLE AND SUBTITLE

The power of slightly more than one sample in randomized load balancing

5a. CONTRACT NUMBER
W911NF-12-1-0385

5b. GRANT NUMBER

5c. PROGRAM ELEMENT NUMBER
611103

6. AUTHORS
L. Ying, R. Srikant, X. Kang

5d. PROJECT NUMBER

5e. TASK NUMBER

5f. WORK UNIT NUMBER

7. PERFORMING ORGANIZATION NAMES AND ADDRESSES
Cornell University
Office of Sponsored Programs
373 Pine Tree Road
Ithaca, NY 14850 -2820

8. PERFORMING ORGANIZATION REPORT NUMBER

9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS (ES)

U.S. Army Research Office
P.O. Box 12211
Research Triangle Park, NC 27709-2211

10. SPONSOR/MONITOR'S ACRONYM(S)
ARO

11. SPONSOR/MONITOR'S REPORT NUMBER(S)
61783-MA-MUR.89

12. DISTRIBUTION AVAILIBILITY STATEMENT

Approved for public release; distribution is unlimited.

13. SUPPLEMENTARY NOTES

The views, opinions and/or findings contained in this report are those of the author(s) and should not contrued as an official Department of the Army position, policy or decision, unless so designated by other documentation.

14. ABSTRACT

In many computing and networking applications, arriving tasks have to be routed to one of many servers, with the goal of minimizing queueing delays. When the number of processors is very large, a popular routing algorithm works as follows: select two servers at random and route an arriving task to the least loaded of the two. It is well-known that this algorithm dramatically reduces queueing delays compared to an algorithm which routes to a single randomly selected server. In recent cloud computing applications, it has been observed that even sampling two queues per arriving task can be expensive and can even increase delays due to messaging overhead. So there is an

15. SUBJECT TERMS

Load Balancing, Mean-Field Analysis, Cloud Computing

16. SECURITY CLASSIFICATION OF:			17. LIMITATION OF ABSTRACT	15. NUMBER OF PAGES	19a. NAME OF RESPONSIBLE PERSON
a. REPORT	b. ABSTRACT	c. THIS PAGE			Sidney Resnick
UU	UU	UU	UU		19b. TELEPHONE NUMBER 607-255-1210

Report Title

The power of slightly more than one sample in randomized load balancing

ABSTRACT

In many computing and networking applications, arriving tasks have to be routed to one of many servers, with the goal of minimizing queueing delays. When the number of processors is very large, a popular routing algorithm works as follows: select two servers at random and route an arriving task to the least loaded of the two. It is well-known that this algorithm dramatically reduces queueing delays compared to an algorithm which routes to a single randomly selected server. In recent cloud computing applications, it has been observed that even sampling two queues per arriving task can be expensive and can even increase delays due to messaging overhead. So there is an interest in reducing the number of sampled queues per arriving task. In this paper, we show that the number of sampled queues can be dramatically reduced by using the fact that tasks arrive in batches (called jobs). In particular, we sample a subset of the queues such that the size of the subset is slightly larger than the batch size (thus, on average, we only sample slightly more than one queue per task). Once a random subset of the queues is sampled, we propose a new load balancing method called batch-filling to attempt to equalize the load among the sampled servers. We show that our algorithm maintains the same asymptotic performance as the so-called power-of-two-choices algorithm while using only half the number of samples.

Conference Name: Proc. of IEEE INFOCOM

Conference Date: April 26, 2015

The Power of Slightly More than One Sample in Randomized Load Balancing

Lei Ying, R. Srikant and Xiaohan Kang

***Abstract*—In many computing and networking applications, arriving tasks have to be routed to one of many servers, with the goal of minimizing queueing delays. When the number of processors is very large, a popular routing algorithm works as follows: select two servers at random and route an arriving task to the least loaded of the two. It is well-known that this algorithm dramatically reduces queueing delays compared to an algorithm which routes to a single randomly selected server. In recent cloud computing applications, it has been observed that even sampling two queues per arriving task can be expensive and can even increase delays due to messaging overhead. So there is an interest in reducing the number of sampled queues per arriving task. In this paper, we show that the number of sampled queues can be dramatically reduced by using the fact that tasks arrive in batches (called jobs). In particular, we sample a subset of the queues such that the size of the subset is slightly larger than the batch size (thus, on average, we only sample slightly more than one queue per task). Once a random subset of the queues is sampled, we propose a new load balancing method called *batch-filling* to attempt to equalize the load among the sampled servers. We show that our algorithm maintains the same asymptotic performance as the so-called power-of-two-choices algorithm while using only half the number of samples.**

I. Introduction

In many computing and networking applications, including routing, hashing, and load balancing (see [14]), a router (also called scheduler) has to route arriving tasks to one of many servers with the goal of minimizing queueing delays. Such applications have been increasingly relevant recently, due to the explosive growth of cloud computing where a large number of servers in a data center are used to process a large volume of tasks. Ideally, one would like the router to consider the queue lengths at all the servers and select the shortest of the queues since this is delay optimal, at least in certain traffic regimes (see [6] and references cited within). However, sampling all the queues can be expensive when the number of servers is very large. Motivated by such considerations, load balancing in the large-server limit was studied in [9], [11], [19]. The key result in those papers is that queueing delays can be dramatically reduced by sampling two servers for each task, instead of just one, and routing the task to the shorter of the two queues. We will call this basic algorithm the *power-of-two-choices* algorithm as in prior work. These results have been extended in various directions. In [3], [4], the results have been extended to the case of heavy-tailed distributions, in [17], [18], the effect of resource pooling has been considered, and the case of heterogeneous servers operating under the processor-sharing discipline has been treated in [12].

In this paper, we are motivated by cloud computing applications in which each arrival is a job consisting of many tasks, each of which can be executed in parallel in possibly different servers. In queueing theory parlance, this model differs from the models mentioned earlier due to the fact that task arrivals occur in batches, i.e., each job corresponding to a batch arrival of tasks. We note the terminology we use here: a job is a collection of tasks, and each task can be routed independently of each other. Such a model arises in the well-known Map/Reduce framework, for example, where each Map job consists of many Map tasks (here, we do not consider the Reduce phase of the job). More generally, any parallel processing computer system will have job arrivals which consist of many tasks which can be executed in parallel. The question of interest is whether the fact that there are batch arrivals can be exploited to significantly reduce the sample complexity. Here, by sample complexity, we mean the number of queues sampled per arriving task to make routing decision. Our motivation for this problem arises from a study of batch arrivals to computing clusters presented in [13], where the authors observe a phenomenon called messaging overhead, i.e., the overhead of providing task backlog feedback can slow down servers and increase the delays experienced by tasks/jobs. Further, [13] proposes an algorithm which achieves better performance than the power-of-two-choices algorithm when both of them use the same number of samples per arriving task. In this paper, we observe that this basic algorithm for batch arrivals suggested in [13] does not work well in all traffic conditions. Moreover, we present a new algorithm which exploits batch arrivals in a manner in which it provides much better sample complexity than the power-of-two-choices algorithm for the same delay performance. Further, when both algorithms are allowed the same sample complexity, our algorithm achieves better delay performance.

Our main contributions are as follows:

1) We present an algorithm which samples md queues where m is the batch size (i.e., number of tasks) of a job. Thus, d is the number of sampled queues per task. The tasks are routed to the queues using a novel algorithm called *water filling*.
2) We first study our algorithm and other previously proposed algorithms using a *mean-field analysis*. We show that, for any $d > 1$, we achieve better performance than the traditional power-of-two-choices algorithm in the large-

Lei Ying and Xiaohan Kang are with the School of Electrical, Computer and Energy Engineering, Arizona State University, Tempe, AZ 85287 (e-mail: lei.ying.2@asu.edu; xiaohan.kang@asu.edu)

R. Srikant is with the Department of Electrical and Computer Engineering at University of Illinois at Urbana-Champaign, Urbana, IL 61801 (e-mail: rsrikant@illinois.edu)

systems regime. Thus, the mean-field analysis shows that, in the large-systems regime, we can reduce the number of samples per arriving task dramatically: from $d = 2$ to any $d > 1$.

3) We then justify the mean-field analysis. In particular, we first show that the stochastic system dynamics converge to deterministic differential equations in the large-systems limit for any finite t. Our proof here is motivated by the proof of a celebrated result on density-dependent Markov processes called *Kurtz's theorem* [7], but our model is somewhat nonstandard and requires additional steps which are not needed in the original Kurtz's theorem. Further, using a novel Lyapunov function, we show that the system of differential equations converges to an equilibrium described by the mean-field analysis. Then by showing the interchange of the limits, we prove the stationary distribution of the queue size distribution converges to the solution of the differential equations.
4) Finally, we perform extensive simulations to justify that our analytical conclusions are indeed valid in large, but finite, systems. In particular, simulations show that our algorithm with just one sample per task on average, achieves the same job delay performance as the power-of-two-choices algorithm and dramatically reduces the delay compared to the algorithm proposed in [13].

II. Problem Statement and Main Results

We consider a computing cluster with n identical servers and a central scheduler as shown in Figure 1. Each server can process one task at a time. Tasks arrive at the scheduler in batches (also called jobs). Each batch consists of m tasks and the job arrival process is a Poisson process with rate $\frac{n}{m}\lambda$. We want the batch size to be not too small, so we assume that $m = \Theta(\log n)$ and m is increasing function of n. For simplicity, we consider a deterministic batch size here, but the results in the paper can be extended to random batch sizes as well in a straightforward manner, as will be discussed in the extended version of the paper. Furthermore, the results of this paper hold when the system has multiple distributed schedulers and the job arrivals on these schedulers are independent Poisson processes with aggregated rate $\frac{n}{m}\lambda$. This is because the sum of independent Poisson processes is Poisson. The scheduler dispatches the tasks to the servers when a job arrives. The service times of the tasks are exponentially distributed with mean 1, and are independent across tasks. When a task arrives at a server, it is processed immediately if the server is idle or waits in a FIFO (first-in, first-out) queue if the server is busy.

We first describe the traditional power-of-d-choices algorithm (which is a simple generalization of the power-of-two-choices mentioned in the previous section) and another previously-proposed idea called the batch sampling algorithm. Then, we present our idea which we call batch-filling, which combines batch sampling with our new load balancing technique called water-filling.

The-Power-of-d-Choices [10], [19]: *When a batch of m tasks arrive, the scheduler probes d servers uniformly at random* for each task. *The task is routed to the least loaded server.* ◇

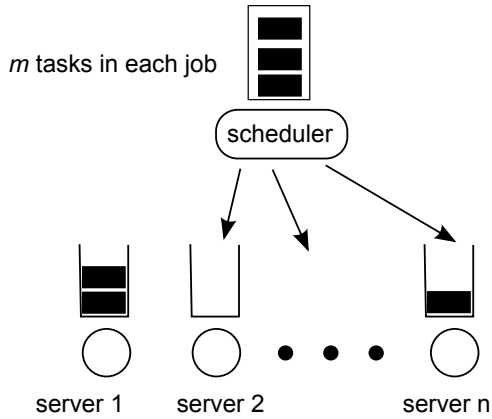

Fig. 1: A computing cluster with n servers and a central scheduler

Batch-Sampling [13]: *When a batch of m tasks arrive, the scheduler probes dm servers uniformly at random to acquire their queue lengths. The m tasks are added to the the least loaded m servers, one for each server.* ◇

In this paper, we propose a new load-balancing algorithm, named *batch-filling*: we sample queues as in the batch sampling algorithm but the way that tasks are routed to servers uses a different procedure which we call *water-filling*.

Batch-Filling: *When a batch of m tasks arrive, the scheduler probes dm servers uniformly at random to acquire their queue lengths. The m tasks are added to the dm servers using* water filling*, specifically, the tasks are dispatched one by one to the least loaded server, where the queue length of a server is updated after it receives a task.* ◇

Remark: *In batch-filling, the first task in a batch is routed to the least loaded server among the sampled servers, i.e., the one with the smallest number of tasks in its queue. The key difference compared to batch-sampling is that the server's queue size is updated after this (which means that this server may no longer be the least-loaded in the sampled servers), and then the next task in the batch is again routed to the least loaded server, and so on. As we will see later, this small change to the routing algorithm has dramatic consequences to the sample complexity of the algorithm. In all algorithms, at each step, ties are broken at random if there is more than one least-loaded server.*

In this paper, d is called *probe ratio,* which is assumed to be a constant independent of n. As in [10], [19], we will study the different algorithms in the *large-systems* limit, i.e., as $n \to \infty$, since a data center today may consist of tens of thousands of servers. The main theoretical results which will be established in the paper are summarized in Table I, and we discuss them below.

- The expected per-task delay of batch-filling with any $d > 1$ is smaller than both batch-sampling with $d = 2$ and the-power-of-two-choices when $\lambda \to 1^-$. In other words, batch-filling outperforms the other two algorithms by sampling *slightly more than one server per task,* hence the title of the paper.
- The size of the longest-queue in the system under the-

	Batch-Filling	Batch-Sampling	Pod
Expected per-task delay	$-\frac{1}{\lambda}\frac{\log(1-\lambda)}{\log(1+\lambda d)}+O_\lambda(1)$	$-\frac{1}{\lambda}\frac{\log(1-\lambda)}{\log(\lambda d)}+O_\lambda(1)$	$-\frac{1}{\lambda}\frac{\log(1-\lambda)}{\log(\lambda d)}+O_\lambda(1)$
Maximum queue size in the system	$\left\lceil -\frac{\log(1-\lambda)}{\log(1+\lambda d)} \right\rceil$	$\left\lceil \frac{\log \frac{d-1}{d(1-\lambda)}}{\log(\lambda d)} \right\rceil$ if $\lambda d \neq 1$ $\left\lceil \frac{1}{1-\lambda} \right\rceil$ if $\lambda d = 1$	∞

TABLE I: This table summarizes the expected per-task delays and the maximum queue sizes of the three scheduling algorithms. The order notation $O_\lambda(\cdot)$ is defined when $1/(1-\lambda) \to \infty$, i.e., $\lambda \to 1^-$. Pod stands for the-power-of-d-choices. In batch-filling and batch-sampling, $d > 1$; and in the-power-of-d-choices, d is an integer and $d \geq 2$.

power-of-d-choices is unbounded for any $d \geq 2$ because the stationary queue length distribution has unbounded support. The sizes of the longest-queue under both batch-filling and batch-sampling are finite because the stationary distributions have bounded support. The longest queue under batch-filling with $d > 1$ is smaller than that of batch-sampling with $d = 2$ when $\lambda \to 1^-$. When d is close to 1, the size of longest queue under batch-filling is much smaller than that under batch-sampling (7 versus 26 when $d = 1.1$ and $\lambda = 0.99$).

- The small and bounded size of the queues under batch filling has important consequences. A job is said to be completed when all the tasks in the job are completed. Since the tail of the queue size is cut off, this has the effect of significantly reducing job completion delays, as we will see later in the simulations section.
- The above theoretical results suggest that the sample complexity (i.e., the number of samples per arriving task) can be significantly reduced under batch-filling. On the other hand, the computational complexity is slightly increased compared to batch-sampling since we require to have to compare the sizes of the smallest queues and the next smallest queues each time a task is routed. However, this increase in computational complexity is a cost to be paid at the router whereas increased sample complexity slows down the servers since they have to send queue length feedback which takes time away from their primary role of processing tasks. This is the reason why sample complexity is a more significant issue than the computational complexity in data centers (although we do not want the computational complexity to be very high either). The batch-sampling algorithm performs $O(dm \log m)$ computations per batch which corresponding to a sorting operation, while batch-filling algorithm performs an additional $2m$ operations since it has to keep track of the queue lengths of the smallest queues and the next smallest queue.

III. Mean-Field Analysis

In this section, we will use mean-field analysis to study the stationary distributions of the queue lengths under batch-filling and batch-sampling. The results will be further validated using a proof inspired by the proof of Kurtz's theorem in Section IV. Let $Q_k^{(n)}(t)$ denote the queue length of the kth server at time t in a system with n queues. It can be easily verified that $\mathbf{Q}^{(n)}(t)$ is an irreducible and nonexplosive Markov chain, and using the standard Foster-Lyapunov theorem (see, for example, [15]) it can be verified that the Markov chain is positive recurrent and hence, has a unique stationary distribution.

Theorem 1. *The Markov chain $\mathbf{Q}^{(n)}(t)$ is positive recurrent under batch-filling. Furthermore, there exists a constant $c > 0$, independent of n, such that*

$$E\left[\frac{1}{n}\sum_{k=1}^{n}\hat{Q}_k^{(n)}\right] < c$$

for any n, where $\hat{Q}_k^{(n)}$ denotes the queue length of server k in the steady state. $\diamond$

The proof of this theorem is presented in the appendix. Let $\pi_i^{(n)}$ denote the stationary distribution of queue k, i.e., the probability that the queue size is i at server k. Here, the index k is ignored because the stationary distributions are identically across servers. According to the theorem above, we have $\sum_i i\pi_i^{(n)} < c$, which further implies that $\pi_i^{(n)} \to 0$ as $i \to \infty$ and $\sum_{j=i}^{\infty}\pi_j^{(n)} \to 0$ as $i \to \infty$. We remark that one challenge in proving that the stochastic system dynamics converge to deterministic differential equations lies in that the system is an infinite-dimensional system. We will utilize the facts mentioned above to overcome this challenge in the proofs.

The mean-field analysis proceeds as follows. Assume the n queues are in the steady state, and further assume that the queue lengths are identically and independently distributed (i.i.d.) with distribution $\boldsymbol{\pi}$. This i.i.d. assumption in the mean-field analysis will be validated later in Section IV in the large-systems limit. Now consider the queue evolution of one server in the system. Each queue forms an independent Markov chain as shown in Figure 2, denoted by $Q^{(n)}(t)$, and the transition rates will be determined by the particular strategy used to route tasks to servers. We will derive the transition rates for each of the strategies described earlier, namely batch filling, batch sampling, and the power-of-d-choices, in the rest of this section.

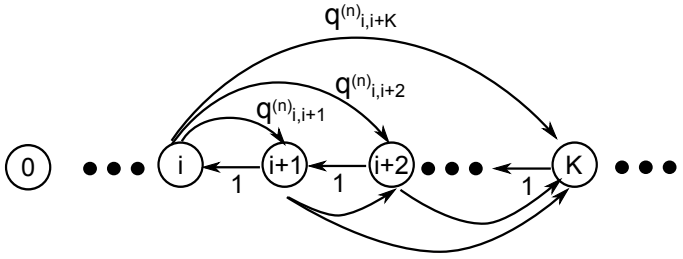

Fig. 2: The Markov chain representing the nth system in the mean field analysis

A. The stationary distribution under batch-filling

We first consider the batch-filling algorithm. The down-crossing transition rate from state i to $i-1$ is 1 for all $i \geq 1$, i.e.,

$$q_{i,i-1}^{(n)} = 1 \quad \forall\, i,$$

because the processing time of a task is exponentially distributed with mean 1. The up-crossing transition rate from state i to state j for $j > i$ is

$$\begin{aligned} q_{i,j}^{(n)} &= \frac{n}{m}\lambda \times \frac{dm}{n} \times \sum_{\boldsymbol{\phi}} \mathbb{P}(\boldsymbol{\phi}) \times \mathbb{P}(j|\boldsymbol{\phi}, i) \\ &= d\lambda \sum_{\boldsymbol{\phi}} \mathbb{P}(\boldsymbol{\phi})\, \mathbb{P}(j|\boldsymbol{\phi}, i). \end{aligned} \tag{1}$$

In the expression above,

- $\frac{n}{m}\lambda$ is the batch arrival rate;
- dm/n is the probability a server is probed when dm servers are sampled;
- $\boldsymbol{\phi}$ is a $(dm-1)$-vector that denotes the queue lengths of the other $dm-1$ sampled servers, so

$$\mathbb{P}(\boldsymbol{\phi}) = \prod_{k=1}^{dm-1} \pi_{\phi_k};$$

and

- $\mathbb{P}(j|\boldsymbol{\phi}, i)$ is the probability that a server's queue length becomes j when the server is sampled and is in state i, and the the states of the other $dm-1$ sampled servers are $\boldsymbol{\phi}$.

Without loss of generality, assume $\phi_k \leq \phi_l$ if $k \leq l$, i.e., $\boldsymbol{\phi}$ is ordered. Recall that batch-filling dispatches tasks using water filling among the sampled dm queues. Therefore, given i and $\boldsymbol{\phi}$, either $j = i$ if no task is assigned to the server, or j takes two possible values. Consider a simple example in Figure 3 where three tasks will be dispatched to four servers with queue lengths 1, 1, 4, and 4. Then the servers whose queue size is 4 will not receive any task, and the servers whose queue size is 1 will receive one or two tasks.

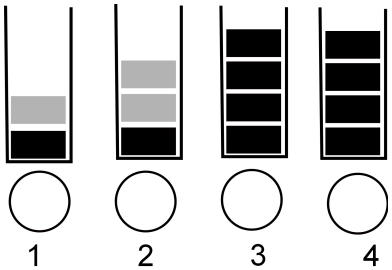

Fig. 3: An example of water filling

Assume ties are broken uniformly at random. The values of $\mathbb{P}(j|\boldsymbol{\phi}, i)$ are summarized below.

- If

$$\sum_{k=1}^{dm-1} (i - \phi_k)\mathbb{I}_{\phi_k \leq i-1} \geq m, \tag{2}$$

which means that the tasks will be assigned to servers whose original queue sizes are smaller than i, then

$$\mathbb{P}(j|\boldsymbol{\phi}, i) = \begin{cases} 1 & \text{if } j = i, \\ 0 & \text{if } j \neq i. \end{cases}$$

- If condition (2) does not hold, then the server with queue size i will receive some tasks, and

$$\mathbb{P}(j|\boldsymbol{\phi}, i) = \begin{cases} 1 - \alpha_{\boldsymbol{\phi},i} & \text{if } j = \bar{Q}_{\boldsymbol{\phi},i} - 1, \\ \alpha_{\boldsymbol{\phi},i} & \text{if } j = \bar{Q}_{\boldsymbol{\phi},i}, \end{cases}$$

where

$$\bar{Q}_{\boldsymbol{\phi},i} = \min\left\{ j : (j-i) + \sum_{k=1}^{dm-1} (j - \phi_k)\mathbb{I}_{\phi_k \leq j-1} \geq m \right\},$$

which is the maximum size a queue can be filled up to during the water filling, and $\alpha_{\boldsymbol{\phi},i}$ is given by

$$\frac{m - (\bar{Q}_{\boldsymbol{\phi},i} - 1 - i) - \sum_{k=1}^{dm-1} (\bar{Q}_{\boldsymbol{\phi},i} - 1 - \phi_k)\mathbb{I}_{\phi_k \leq \bar{Q}_{\boldsymbol{\phi},i}-1}}{1 + \sum_{k=1}^{dm-1} \mathbb{I}_{\phi_k \leq \bar{Q}_{\boldsymbol{\phi},i}-1}},$$

which is the probability that a server receives one more task after its queue size becomes $\bar{Q}_{\boldsymbol{\phi},i} - 1$ during water-filling.

While the transition rate $q_{i,j}^{(n)}$ in (1) is a complex expression for finite n, the following lemma shows that $q_{i,j}^{(n)}$ converges to some simple $q_{i,j}$ as $n \to \infty$. The proof of this lemma is presented in the appendix.

Lemma 2. *Under batch-filling, the transition rates given distribution $\boldsymbol{\pi}$, denoted by $q_{i,j}^{(n)}(\boldsymbol{\pi})$, converges; and specifically,*

$$\lim_{n\to\infty} q_{i,j}^{(n)}(\boldsymbol{\pi}) = q_{i,j}(\boldsymbol{\pi}),$$

where for $j \neq i$,

$$q_{i,j}(\boldsymbol{\pi}) = \begin{cases} 1 & \textit{if } j = i-1, \\ \lambda d(1 - \alpha_{\boldsymbol{\pi}}) & \textit{if } j = \bar{Q}_{\boldsymbol{\pi}} - 1 > i, \\ \lambda d \alpha_{\boldsymbol{\pi}} & \textit{if } j = \bar{Q}_{\boldsymbol{\pi}} > i, \\ 0 & \textit{otherwise}, \end{cases}$$

$$\bar{Q}_{\boldsymbol{\pi}} = \min\left\{ j \colon \sum_{l=0}^{j-1} (j-l)\pi_l \geq \frac{1}{d} \right\} \tag{3}$$

and

$$\alpha_{\boldsymbol{\pi}} = \frac{\frac{1}{d} - \sum_{j=0}^{\bar{Q}_{\boldsymbol{\pi}}-2} (\bar{Q}_{\boldsymbol{\pi}} - 1 - j)\pi_j}{\sum_{j=0}^{\bar{Q}_{\boldsymbol{\pi}}-1} \pi_j} \in (0, 1].$$

$\diamond$

According to the lemma above, the queue length dynamics of a single server, in the limit as the number of servers becomes infinity, can be represented by the Markov chain in Figure 4, where the up-crossing transitions are *into* only two states $\bar{Q}_{\boldsymbol{\pi}} - 1$ and $\bar{Q}_{\boldsymbol{\pi}}$ due to water filling. Based on Lemma 2, we can calculate the stationary distribution of the queue length of a single server in the large-system limit by finding $\hat{\boldsymbol{\pi}}$ that satisfies the global balance equation [15].

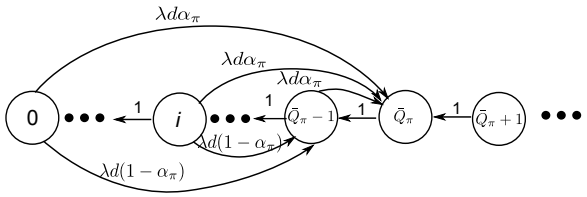

Fig. 4: The queue-length Markov chain of a single-server, in the large-system limit, under batch-filing

Theorem 3. *The stationary distribution of the queue length of a single server in the large-system limit under batch-filling is*

$$\hat{\pi}_i = \begin{cases} 1-\lambda & i=0, \\ (1-\lambda)\lambda d(1+\lambda d)^{i-1} & 1 \le i \le \bar{Q}_{BF}-1, \\ 1-(1-\lambda)(1+\lambda d)^{\bar{Q}_{BF}-1} & i=\bar{Q}_{BF}, \\ 0 & \textit{otherwise}. \end{cases} \tag{4}$$

where $\bar{Q}_{BF} = \left\lceil -\frac{\log(1-\lambda)}{\log(1+\lambda d)} \right\rceil$. *The expected queue length is*

$$-\frac{\log(1-\lambda)}{\log(1+\lambda d)} + O_\lambda(1).$$

Proof. We first show $\bar{Q}_{BF} = \bar{Q}_{\hat{\boldsymbol{\pi}}}$, where $\bar{Q}_{\boldsymbol{\pi}}$ is defined in (3). Note that $\bar{Q}_{BF} \ge 1$. If λ and d are such that $\bar{Q}_{BF} = 1$, then equivalently

$$-\frac{\log(1-\lambda)}{\log(1+\lambda d)} \le 1,$$

which implies that $\frac{1}{1-\lambda} \le 1+\lambda d$, or $\frac{1}{d} \le 1-\lambda$. Then $\hat{\boldsymbol{\pi}} = (1-\lambda, \lambda, 0, \dots)$ and

$$\bar{Q}_{\hat{\boldsymbol{\pi}}} = 1 = \bar{Q}_{BF}.$$

If λ and d are such that $\bar{Q}_{BF} > 1$, according to (3), to show $\bar{Q}_{\hat{\boldsymbol{\pi}}} = \bar{Q}_{BF}$ we only need to show

$$\sum_{l=0}^{\bar{Q}_{BF}-2} (\bar{Q}_{BF}-1-l)\hat{\pi}_l < \frac{1}{d} \le \sum_{l=0}^{\bar{Q}_{BF}-1} (\bar{Q}_{BF}-l)\hat{\pi}_l. \tag{5}$$

Let LHS and RHS denote the left-hand-side and the right-hand-side of (5). Then

$$\text{LHS} = \sum_{i=0}^{\bar{Q}_{BF}-2} \sum_{j=0}^{i} \hat{\pi}_j = (1-\lambda)\frac{(1+\lambda d)^{\bar{Q}_{BF}-1}-1}{\lambda d},$$

and

$$\text{RHS} = \sum_{i=0}^{\bar{Q}_{BF}-1} \sum_{j=0}^{i} \hat{\pi}_j = (1-\lambda)\frac{(1+\lambda d)^{\bar{Q}_{BF}}-1}{\lambda d}.$$

Then (5) is equivalent to

$$\bar{Q}_{BF} - 1 < -\frac{\log(1-\lambda)}{\log(1+\lambda d)} \le \bar{Q}_{BF},$$

which holds according to the definition of $\bar{Q}_{BF}$.

We next check the global balance equations. For $i = 0$,

$$\begin{aligned} &\hat{\pi}_0(q_{0,\bar{Q}_{BF}} + q_{0,\bar{Q}_{BF}-1}) - \hat{\pi}_1 q_{1,0} \\ =&(1-\lambda)\lambda d - (1-\lambda)\lambda d \\ =&0. \end{aligned}$$

For $1 \le i \le \bar{Q}_{BF} - 2$,

$$\begin{aligned} &\hat{\pi}_i(q_{i,i-1} + q_{i,\bar{Q}_{BF}} + q_{i,\bar{Q}_{BF}-1}) - \hat{\pi}_{i+1} q_{i+1,i} \\ =&(1-\lambda)\lambda d(1+\lambda d)^{i-1}(1+\lambda d) - (1-\lambda)\lambda d(1+\lambda d)^{i} \\ =&0. \end{aligned}$$

For $i = \bar{Q}_{BF} - 1$,

$$\begin{aligned} &\hat{\pi}_{\bar{Q}_{BF}-1}(q_{\bar{Q}_{BF}-1,\bar{Q}_{BF}-2} + q_{\bar{Q}_{BF}-1,\bar{Q}_{BF}}) \\ &- \sum_{i=0}^{\bar{Q}_{BF}-2} \hat{\pi}_i q_{i,\bar{Q}_{BF}-1} \\ &- \hat{\pi}_{\bar{Q}_{BF}} q_{\bar{Q}_{BF},\bar{Q}_{BF}-1} \\ =&(1-\lambda)\lambda d(1+\lambda d)^{\bar{Q}_{BF}-2}(1+\lambda d\alpha_{\hat{\boldsymbol{\pi}}}) \\ &- (1-\lambda)(1+\lambda d)^{\bar{Q}_{BF}-2}\lambda d(1-\alpha_{\hat{\boldsymbol{\pi}}}) \\ &- (1-(1-\lambda)(1+\lambda d)^{\bar{Q}_{BF}-1}) \\ =&(1-\lambda)(1+\lambda d)^{\bar{Q}_{BF}-1}(\lambda d\alpha_{\hat{\boldsymbol{\pi}}} + 1) - 1. \end{aligned}$$

From the definition of $\alpha_{\hat{\boldsymbol{\pi}}}$ we can verify that

$$\alpha_{\hat{\boldsymbol{\pi}}} = \frac{1}{\lambda d(1-\lambda)(1+\lambda d)^{\bar{Q}_{BF}-1}} - \frac{1}{\lambda d}.$$

So we have

$$\begin{aligned} &\hat{\pi}_{\bar{Q}_{BF}-1}(q_{\bar{Q}_{BF}-1,\bar{Q}_{BF}-2} + q_{\bar{Q}_{BF}-1,\bar{Q}_{BF}}) \\ &- \sum_{i=0}^{\bar{Q}_{BF}-2} \hat{\pi}_i q_{i,\bar{Q}_{BF}-1} \\ &- \hat{\pi}_{\bar{Q}_{BF}} q_{\bar{Q}_{BF},\bar{Q}_{BF}-1} \\ =&0. \end{aligned}$$

For $i = \bar{Q}_{BF}$,

$$\begin{aligned} &\hat{\pi}_{\bar{Q}_{BF}} q_{\bar{Q}_{BF},\bar{Q}_{BF}-1} - \sum_{i=0}^{\bar{Q}_{BF}-1} \hat{\pi}_i q_{i,\bar{Q}_{BF}} \\ =&(1-(1-\lambda)(1+\lambda d)^{\bar{Q}_{BF}-1}) - (1-\lambda)(1+\lambda d)^{\bar{Q}_{BF}-1}\lambda d\alpha_{\hat{\boldsymbol{\pi}}} \\ =&1-(1-\lambda)(1+\lambda d)^{\bar{Q}_{BF}-1}(1+\lambda d\alpha_{\hat{\boldsymbol{\pi}}}) \\ =&0. \end{aligned}$$

So the global balance equations holds.

Finally the expected queue length in stationary distribution

is

$$\begin{aligned}
&\hat{\pi}_1 + 2\hat{\pi}_2 + \cdots + \bar{Q}_{BF}\hat{\pi}_{\bar{Q}_{BF}} \\
=& \sum_{i=0}^{\bar{Q}_{BF}-1} \left(1 - \sum_{j=0}^{i} \hat{\pi}_j\right) \\
=& \sum_{i=0}^{\bar{Q}_{BF}-1} \left(1 - (1-\lambda)(1+\lambda d)^i\right) \\
=& \bar{Q}_{BF} - (1-\lambda)\frac{(1+\lambda d)^{\bar{Q}_{BF}} - 1}{\lambda d} \\
=& -\frac{\log(1-\lambda)}{\log(1+d)} + O_\lambda(1).
\end{aligned}$$

□

B. The stationary distribution under batch-sampling

Recall in batch-sampling, the m tasks are routed to the least-loaded m queues among the sampled dm queues. Consider a server with queue size i and assume it is probed. Then the server will receive a task with probability

$$\begin{aligned}
&\mathbb{E}\left[\min\left\{1, \left(\frac{m - \sum_{j=0}^{i-1}\sum_{k=1}^{dm-1}\mathbb{I}_{\phi_k=j}}{1+\sum_{k=1}^{dm-1}\mathbb{I}_{\phi_k=i}}\right)^+\right\}\right] \\
=&\mathbb{E}\left[\min\left\{1, \left(\frac{\frac{m}{dm-1} - \sum_{j=0}^{i-1}\frac{\sum_{k=1}^{dm-1}\mathbb{I}_{\phi_k=j}}{dm-1}}{\frac{1}{dm-1}+\frac{\sum_{k=1}^{dm-1}\mathbb{I}_{\phi_k=i}}{dm-1}}\right)^+\right\}\right] \\
\to&\mathbb{E}\left[\min\left\{1, \left(\frac{\frac{1}{d} - \sum_{j=0}^{i-1}\pi_j}{\pi_i}\right)^+\right\}\right].
\end{aligned}$$

Following a similar analysis as batch-filling, we can establish the following lemma. The details are omitted.

Lemma 4. *Under batch-sampling, the transition rates given distribution $\boldsymbol{\pi}$, denoted by $q_{i,j}^{(n)}(\boldsymbol{\pi})$ converges; and specifically,*

$$\lim_{n\to\infty} q_{i,j}^{(n)}(\boldsymbol{\pi}) = q_{i,j}(\boldsymbol{\pi}) = \begin{cases} 1 & \text{if } j = i-1, \\ \lambda d & \text{if } i+1 = j \le \bar{Q}_{\boldsymbol{\pi}} - 1, \\ \lambda\alpha_{\boldsymbol{\pi}} & \text{if } i+1 = j = \bar{Q}_{\boldsymbol{\pi}}, \\ 0 & \text{otherwise}, \end{cases}$$

where

$$\bar{Q}_{\boldsymbol{\pi}} = \min\left\{i\colon \sum_{l=0}^{i-1}\pi_j \ge \frac{1}{d}\right\}$$

and

$$\alpha_{\boldsymbol{\pi}} = \frac{\frac{1}{d} - \sum_{j=0}^{\bar{Q}_{\boldsymbol{\pi}}-2}\pi_j}{\pi_{\bar{Q}_{\boldsymbol{\pi}}-1}} \in (0,1].$$

◇

The Markov chain in the large-system limit is shown in Figure 5. Given $\boldsymbol{\pi}$, the Markov chain is a birth-death process up to state $\bar{Q}_{\boldsymbol{\pi}}$. The stationary distribution can again be calculated using the global balance equations. The results are presented in Theorem 5, and the details are omitted.

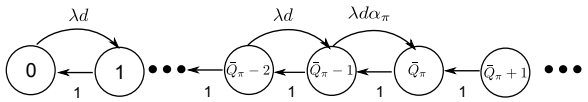

Fig. 5: The Markov chain in the large-system limit under batch-sampling

Theorem 5. *The stationary distribution of the queue length of a single server in the large-system limit under batch-sampling is*

$$\hat{\pi}_i = \begin{cases} 1-\lambda & i = 0, \\ (1-\lambda)\lambda^i d^i & 1 \le i \le \bar{Q}_{BS} - 1, \\ 1 - (1-\lambda)\frac{\lambda^i d^i - 1}{\lambda d - 1} & i = \bar{Q}_{BS}, \\ 0 & \text{otherwise}. \end{cases}$$

where

$$\bar{Q}_{BS} = \left\lceil \frac{\log\frac{d-1}{d(1-\lambda)}}{\log(\lambda d)} \right\rceil.$$

The expected queue length is

$$-\frac{\log(1-\lambda)}{\log(\lambda d)} + O_\lambda(1).$$

◇

C. The stationary distribution under the-power-of-d-choices

For a system with non-batch (single) arrivals, the stationary queue-length distribution of a single server in the large-system limit under the-power-of-d-choices has been established in [10], [19]. The power-of-d choices routing under our batch-arrival model also satisfies the same limiting queue-length distribution, which we provide below for comparison purposes.

Theorem 6. *The stationary distribution of the queue length of a server in the infinite system under the-power-of-d-choices is*

$$\hat{\pi}_i = \lambda^{\frac{d^i-1}{d-1}} - \lambda^{\frac{d^{i+1}-1}{d-1}}.$$

The expected queue length is

$$-\frac{\log(1-\lambda)}{\log(\lambda d)} + O_\lambda(1).$$

◇

IV. Differential Equations and Kurtz's Theorem

The results in the previous section were obtained using the mean-field analysis which assumes that the queues are i.i.d. across servers. We will justify the mean-field analysis in this section.

Again, we will focus on batch-filling. The same results can be established for batch-sampling and the-power-of-d-choices by following similar steps. We first consider the following

non-linear system described by differential equations:

$$\frac{\mathrm{d}x_i}{\mathrm{d}t} = \begin{cases} -(1+\lambda d)x_i + x_{i+1} & i \le \bar{X}_{\mathbf{x}} - 2, \\ \lambda d(1-\alpha_{\mathbf{x}}) \sum_{j=0}^{i-1} x_j - (1+\lambda d\alpha_{\mathbf{x}})x_i + x_{i+1}, & i = \bar{X}_{\mathbf{x}} - 1 \\ \lambda d\alpha_{\mathbf{x}} \sum_{j=0}^{i} x_j - x_i + x_{i+1}, & i = \bar{X}_{\mathbf{x}} \\ -x_i + x_{i+1} & \text{otherwise,} \end{cases} \tag{6}$$

where

$$\bar{X}_{\mathbf{x}} = \min\left\{ j \colon \sum_{l=0}^{j-1} (j-l)x_l \ge \frac{1}{d} \right\}$$

and

$$\alpha_{\mathbf{x}} = \frac{\frac{1}{d} - \sum_{j=0}^{\bar{X}_{\mathbf{x}}-2} (\bar{X}_{\mathbf{x}} - 1 - j)x_j}{\sum_{j=0}^{\bar{X}_{\mathbf{x}}-1} x_j}.$$

These differential equations are derived from the Markov chain in Figure 4. View x_i as the fraction of queues with length i. Consider x_i for $i \le \bar{X}_{\mathbf{x}} - 2$. According to Figure 4, x_i decreases with rate $x_i \times (1+\lambda d)$ because the queue size of a server with size i becomes $i-1$ with rate 1 and becomes $\bar{X}_{\mathbf{x}} - 1$ or $\bar{X}_{\mathbf{x}}$ with total rate λd; and x_i increases with rate x_{i+1} because a queue with size $i+1$ becomes a queue with size i with rate 1. Note this is a non-linear system because $\alpha_{\mathbf{x}}$ and $\bar{X}_{\mathbf{x}}$ depend on the state $\mathbf{x}$.

We further define

$$s_i(t) = \sum_{j=i}^{\infty} x_j(t)$$

for $i \ge 0$, which is related to the fraction of the servers with queue size $\ge i$, and

$$\hat{s}_i = \sum_{j=i}^{\infty} \hat{\pi}_j$$

for $\hat{\boldsymbol{\pi}}$ defined in (4). Note that $s_0(t) = 1$ for any t. The differential equations of the non-linear system can be written in terms of $\mathbf{s}(t)$ as follows:

$$\frac{\mathrm{d}s_i}{\mathrm{d}t} = \begin{cases} \lambda d - (1+\lambda d)s_i + s_{i+1} & i \le \bar{X}_{\mathbf{s}} - 1, \\ \lambda - \lambda d \sum_{j=0}^{i-1} (1-s_j) - s_i + s_{i+1}, & i = \bar{X}_{\mathbf{s}} \\ -s_i + s_{i+1} & \text{otherwise,} \end{cases} \tag{7}$$

where

$$\bar{X}_{\mathbf{s}} = \max\left\{ i \colon \sum_{j=0}^{i-1} (1-s_j) \le \frac{1}{d} \right\}.$$

The following theorem establishes the equilibrium point and the stability of this non-linear system. The proof is presented in the appendix.

Theorem 7. *Assume the initial condition $\mathbf{s}(0)$ satisfies $1 = s_1(0) \ge s_2(0) \ge \cdots \ge 0$ and (ii) $|\mathbf{s}(0)| < \infty$. Starting from $\mathbf{s}(0)$, the system converges to the equilibrium point $\hat{\mathbf{s}}$ as $t \to \infty$, where $|\cdot|$ is the 1-norm.* ◇

Next define $\Pi_i^{(n)}(t)$ to be number of servers with queue size i in the nth system, and $\pi_i^{(n)}(t) = \frac{1}{n}\Pi_i^{(n)}(t)$ to be the fraction of servers with queue size i in the nth system. Here we deliberately reuse notation π because in the steady state, the fraction of servers with queue size i is equal to the probability that the queue size of a server is i. However, note that here $\boldsymbol{\pi}^{(n)}(t)$ is a random vector instead of a distribution. Define the vector $\boldsymbol{\Gamma}^{(n)}(t) \in \mathbb{N}^{\infty}$ such that its i^{th} component $\Gamma_i^{(n)}(t) = \sum_{j=i}^{\infty} \Pi_j^{(n)}(t)$ is the number of servers whose queue lengths are at least i, $\boldsymbol{\gamma}^{(n)}(t) = \frac{\boldsymbol{\Gamma}^{(n)}(t)}{n}$, and $\hat{\boldsymbol{\gamma}}$ such that $\hat{\gamma}_i = \sum_{j=i}^{\infty} \hat{\pi}_j$ for $\hat{\boldsymbol{\pi}}$ defined in (4).

The following theorem states that $\boldsymbol{\gamma}^{(n)}(t)$, which is stochastic, coincides with $\mathbf{s}(t)$ for any bounded time interval $[0,t]$ when $n \to \infty$. Here we define $\bar{\mathcal{U}}$ to be the space of all sequences $\boldsymbol{\gamma}$ such that

$$1 = \gamma_0 \ge \gamma_1 \ge \cdots \ge 0 \tag{8}$$

with the 1-norm. The proof is presented in the appendix.

Theorem 8. *Suppose that $\boldsymbol{\gamma}^{(n)}(0) \to \mathbf{s}(0)$ in probability, where $\mathbf{s}(0)$ is a deterministic initial condition such that $\mathbf{s}(0) \ge 0$ and $|\mathbf{s}(0)| < \infty$. Then the following holds*

$$\lim_{n\to\infty} \sup_{0\le u\le t} |\boldsymbol{\gamma}^{(n)}(u) - \mathbf{s}(u)| = 0 \quad \textit{in probability.}$$ ◇

This result is motivated by Kurtz's theorem [7]. However, we remark that $\Pi_i^{(n)}(t)$ is not a classical density dependent Markov chain because $q_{i,j}^{(n)}$ cannot be written in the form of $n\beta_l$ for some β_l independent of n, and $\boldsymbol{\gamma}^{(n)}$ is an infinite-dimensional vector. Therefore, the proof of Kurtz's theorem does not directly apply. Our proof is a non-trivial extension of Kurtz's theorem.

We also remark that $|\mathbf{s}(0)| = \sum_i i x_i(0) < \infty$ is related to the average queue size at a server, so the condition simply requires the average queue length per server is bounded initially.

Theorem 7 and Theorem 8 establish the following result:

$$\boldsymbol{\gamma}^{(n)}(t) \xrightarrow{n\to\infty} \mathbf{s}(t) \xrightarrow{t\to\infty} \hat{\boldsymbol{\gamma}}, \tag{9}$$

which further implies that

$$\boldsymbol{\pi}^{(n)}(t) \xrightarrow{n\to\infty} \mathbf{x}(t) \xrightarrow{t\to\infty} \hat{\boldsymbol{\pi}}. \tag{10}$$

A direct consequence of (10) is that if $\hat{\boldsymbol{\pi}}^{(n)}$ converges to some $\tilde{\boldsymbol{\pi}}$ or a subsequence of $\hat{\boldsymbol{\pi}}^{(n)}$ converges to some $\tilde{\boldsymbol{\pi}}$, then $\tilde{\boldsymbol{\pi}} = \hat{\boldsymbol{\pi}}$. The convergence of stationary distributions will be discussed in the next section.

V. Convergence of the Stationary Distributions

We first present a theorem on the interchange of limits. The theorem is similar to Theorem 5.1 in [1]. However, [1] assumes the state space of each system is finite but in our system, the state space of each queue is the set of nonnegative

integers. While the proofs are similar, we present it here for the completeness of the paper.

Theorem 9. *Consider a sequence of random processes $\mathbf{X}^{(n)}$ indexed by a scaling parameter n, where $\mathbf{X}^{(n)}$ is a vector that denotes value of the process at time t, and a dynamic system $\dot{\mathbf{X}}(t) = \mathbf{F}(\mathbf{X})$. Assume $\mathbf{X}^{(n)}$ and $\hat{\mathbf{X}}$ satisfy the following assumptions:*

- *(A1) Suppose that for any n,*
$$\mathbf{X}^{(n)}(t) \xrightarrow{w} \hat{\mathbf{X}}^{(n)}, \tag{11}$$
where $\hat{\mathbf{X}}^{(n)}$ is the stationary distribution of the random process and $\xrightarrow{x}$ denotes the weak convergence.
- *(A2) Suppose for each finite t,*
$$\mathbf{X}^{(n)}(t) \xrightarrow{w} \mathbf{X}(t), \tag{12}$$
when
$$\lim_{n\to\infty} \mathbf{X}^{(n)}(0) = \mathbf{X}(0)$$
where both $\mathbf{X}^{(n)}(0)$ and $\mathbf{X}(0)$ are deterministic initial conditions, and $\mathbf{X}(0) \in \mathcal{X}$, where $\mathcal{X}$ is a set of initial conditions.
- *(A3) Starting from each initial condition $\mathbf{X}(0) \in \mathcal{X}$, assume that*
$$\lim_{t\to\infty} \mathbf{X}(t) = \hat{\mathbf{X}}. \tag{13}$$
- *(A4) Any subsequence of $\hat{\mathbf{X}}^{(n)}$ has a subsubsequence that weakly converges. The limit of any convergent subsequence, denoted by $\bar{\mathbf{X}}$, satisfies $\mathbb{P}\left(\bar{\mathbf{X}} \in \mathcal{X}\right) = 1$ and its support is separable.*

Then $\hat{\mathbf{X}}^{(n)} \xrightarrow{w} \hat{\mathbf{X}}$. ◇

This result establishes an *interchange of limits* because from (A1) and (A2), we have

$$\lim_{t\to\infty}\lim_{n\to\infty} \mathbf{X}^{(n)}(t) = \lim_{t\to\infty} \mathbf{X}(t) = \hat{\mathbf{X}}.$$

The theorem says that with additional assumptions, we further have

$$\lim_{n\to\infty}\lim_{t\to\infty} \mathbf{X}^{(n)}(t) = \hat{\mathbf{X}}.$$

The proof is presented in the appendix.

By utilizing the result above, we show the convergence of the stationary distribution in the following theorem.

Theorem 10.
$$\hat{\boldsymbol{\gamma}}^{(n)} \xrightarrow{w} \boldsymbol{\gamma}.$$

Proof. Define

$$\mathcal{X} = \{\boldsymbol{\gamma} : 1 = \gamma_1 \geq \gamma_2 \geq \ldots \geq 0, \sum_i \gamma_i < \infty\},$$

which is separable because it is a subspace of $l^1 = \{\boldsymbol{\gamma} : \sum_i |\gamma_i| < \infty\}$, which is a separable metric space.

- (A1) holds due to Theorem 1.
- Note $\lim_{n\to\infty} \boldsymbol{\gamma}^{(n)}(0) = \mathbf{s}(0)$ for deterministic initial conditions $\boldsymbol{\gamma}^{(n)}(0)$ and $\mathbf{s}(0)$ implies that $\boldsymbol{\gamma}^{(n)}(0) \to \mathbf{s}(0)$ in probability. Therefore, according to Theorem 8, given deterministic initial conditions $\boldsymbol{\gamma}^{(n)}(0)$ and $\mathbf{s}(0)$ such that $\lim_{n\to\infty} \boldsymbol{\gamma}^{(n)}(0) = \mathbf{s}(0)$, we have
$$\lim_{n\to\infty} \sup_{0\leq u\leq t} |\boldsymbol{\gamma}^{(n)}(u) - \mathbf{s}(u)| = 0 \quad \text{in probability,}$$
which implies weak convergence.
- (A3) is established in Theorem 7.
- To validate (A4), we consider the space $\tilde{\mathcal{U}}$ which is the set of $\boldsymbol{\gamma}$ satisfying (8) and with the following norm used in [19]
$$\|\boldsymbol{\gamma} - \boldsymbol{\gamma}'\| = \sup_{i\geq 0} \frac{|\gamma_i - \gamma_i'|}{i}.$$
Under this norm, space $\tilde{\mathcal{U}}$ is compact. Define $\hat{\gamma}_i^{(n)} = \sum_{j=i}^{\infty} \hat{\pi}_j^{(n)}$ where $\hat{\boldsymbol{\pi}}^{(n)}$ denotes the stationary distribution of $\boldsymbol{\pi}^{(n)}(t)$. By Prokhorov's theorem [2], since $\tilde{\mathcal{U}}$ is compact, there exists a subsubsequence for any subsequence of $\hat{\boldsymbol{\gamma}}^{(n)}$ that weakly converges to a random vector $\tilde{\boldsymbol{\gamma}}$ under $\|\cdot\|$, which is denoted by $\hat{\boldsymbol{\gamma}}^{(n_k)}$. By the Skorohod representation theorem, there exists a sequence of random vectors with the same distributions that converge almost surely. By slight abuse of the notation, we assume $\hat{\boldsymbol{\gamma}}^{(n_k)}$ converges to $\tilde{\boldsymbol{\gamma}}$ almost surely. Since $0 \leq \gamma_i \leq 1$, by the dominated convergence theorem, we have
$$\lim_{k\to\infty} \mathbb{E}[|\hat{\gamma}_i^{(n_k)} - \tilde{\gamma}_i|] = 0 \quad \forall i. \tag{14}$$
Define
$$f_k(\boldsymbol{\gamma}) = \sum_{i=1}^{k} \gamma_i.$$
It is easy to verify that $f_k(\cdot)$ is a continuous and bounded function under the 1-norm. According to the definition of weak convergence, we have
$$\begin{aligned} &\mathbb{E}\left[\textstyle\sum_i \tilde{\gamma}_i\right] = \lim_{k\to\infty} \mathbb{E}\left[f_k(\tilde{\boldsymbol{\gamma}})\right] \\ = \quad &\lim_{k\to\infty} \lim_{n_k\to\infty} \mathbb{E}\left[f_k(\hat{\boldsymbol{\gamma}}^{(n_k)})\right] \leq c, \end{aligned} \tag{15}$$
where the last inequality is due to Theorem 1, which implies that
$$\mathbb{P}\left(\tilde{\boldsymbol{\gamma}} \in \mathcal{X}\right) = 1.$$
The uniform convergence of the series
$$\sum_{i=k}^{\infty} \mathbb{E}\left[\left|\hat{\gamma}_i^{(n_k)} - \tilde{\gamma}_i\right|\right] \tag{16}$$
is established in Appendix F. By Tonelli's theorem,
$$\lim_{k\to\infty} \mathbb{E}\left[|\hat{\boldsymbol{\gamma}}^{(n_k)} - \tilde{\boldsymbol{\gamma}}|\right] = 0, \tag{17}$$
which implies $\hat{\boldsymbol{\gamma}}^{(n_k)}$ converges weakly to $\tilde{\boldsymbol{\gamma}}$ in 1-norm. Therefore (A4) holds.

□

Based on the theorem above, we further have the following results according to using the same analysis for getting (14) and (17).

Corollary 11.
$$\lim_{n\to\infty} \mathbb{E}[\hat{\gamma}_i^{(n)}] = \hat{\gamma}_i \quad \forall i,$$

$$\lim_{n\to\infty} \mathbb{E}\left[\sum_i \hat{\gamma}_i^{(n)}\right] = \sum_i \hat{\gamma}_i, \tag{18}$$

and

$$\lim_{n\to\infty} \mathbb{E}\left[|\hat{\boldsymbol{\gamma}}^{(n)} - \hat{\boldsymbol{\gamma}}|\right] = 0. \tag{19}$$

In the next corollary, we show that any k queues are independently and identically distributed with distribution $\hat{\boldsymbol{\pi}}$ in the large-system limit, where k is a constant independent of n. Then the system is said to be $\hat{\boldsymbol{\pi}}$-chaotic [16]. We prove the result by showing that the unique stationary distribution of k queues that satisfies the detailed balance equations in the large-system limit has a product form.

Corollary 12. *Consider a set of k servers, and without loss of generality, assume the severs are $1, 2, \cdots, k$. Let $\pi^{(n)}(Q_1, Q_2, \cdots, Q_k)$ denote the station distribution of the queue lengths of these k servers. In the large-system limit, we have*

$$\lim_{n\to\infty} \pi^{(n)}(Q_1, Q_2, \cdots, Q_k) = \prod_{i=1}^{k} \hat{\pi}_{Q_i},$$

i.e., the k queues are independently and identically distributed with distribution $\hat{\boldsymbol{\pi}}$. $\diamond$

VI. Simulations

In this section, we use simulations to evaluate the performance of the three load balancing algorithms in large, but finite-server, systems.

A. Deterministic Batch Size

We first considered systems with $n = 10{,}000$ servers, batch size $m = 100$. We evaluated the per-task and per-job delays of the three algorithms with different probe ratios d. Figures 6 and 7 show the per-task delays and per-job delays, respectively, when $\lambda = 0.7$. Figures 8 and 9 show the per-task delays and per-job delays, respectively, when $\lambda = 0.9$.

From these figures, we have the following observations.

- In terms of per-task delays, batch-filling matches the power-of-two-choices with $d = 1.3$ when $\lambda = 0.7$ and with $d = 1.2$ when $\lambda = 0.9$. Batch-sampling, on the other hand, requires $d = 1.6$ when $\lambda = 0.7$ and $d = 1.7$ when $\lambda = 0.9$ to achieve the same per-task delay as the power-of-two-choices. Furthermore, even with $d = 1$, the per-task delay of batch-filling is only slightly larger than that of the power-of-two-choices; but batch-sampling has much larger per-task delay when $d = 1$ (10 versus 3 when $\lambda = 0.9$). Note that the per-job delay of batch-sampling with $d = 1$ has been omitted in the figure for readability of the figure.
- Batch-filling performs even better in terms of per-job delays. As we can see from Figures 7 and 9, *batch-filling matches the power-of-two-choices even with $d = 1$!* We believe this is because the maximum queue size of batch-filling is smaller than that of the power-of-two-choices when $d = 1$ even though the average queue size is larger. Batch-sampling requires larger probe ratios to match the per-job delays of the power-of-two-choices. This is because the maximum queue size of batch-sampling is larger than that of batch-filling as shown in Table I.

B. Random Batch Size

In this set of simulations, we evaluated the performance of algorithms under random batch sizes. We assume the batch size M is random variable such that with probability 0.5, M is geometrically distributed with mean 75; and with probability 0.5, M is geometrically distributed with mean 125. The other settings are the same as those used with fixed batch sizes. The results for $\lambda = 0.7$ are shown in Figures 10 and Figure 11; and the results for $\lambda = 0.9$ are shown in Figures 12 and 13. We note that the conclusions of our previous simulations do not change with these modifications.

VII. Conclusions and extension

In this paper, we proposed a new load-balancing algorithm, named batch-filling, which uses water-filling to attempt to equalize the load among the sampled servers. The algorithm provides a much lower sample complexity than the power-of-two-choices algorithm for the same delay performance. Specifically, it only needs to sample slightly more than one queue per task to match the per-job delay of the power-of-two-choices algorithm.

We remark that the theoretical results of this paper can be extended to random batch sizes. Let $M^{(n)}(t)$ denote the batch size at time t in the nth system. Assume $M^{(n)}(t)$ are i.i.d. across time t. The main results of this paper hold given the sequence of random variables $\frac{M^{(n)}}{\mathbb{E}[M^{(n)}]}$ converge in distribution, are uniformly integrable, and $M^{(n)}(t) = \Theta(\log n)$. In particular, Theorem 1 can be established by using the same idea that the Lyapunov drift of water-filling is dominated by random routing. Lemma 2 also holds because $\frac{M^{(n)}}{\mathbb{E}[M^{(n)}]}$ converge in probability. The differential equations remain the same under random batch size, so Theorem 7 is still valid. Finally, it is easy to verify that $D_i/(dm)$ converges in mean as $m \to \infty$, where $m = \mathbb{E}[M^{(n)}]$.

Acknowledgments

Research was funded in part by ARO MURI W911NF-12-1-0385 and NSF Grants ECCS-1255425 and ECCS-1202065.

References

[1] V. Anantharam and M. Benchekroun. A technique for computing sojourn times in large networks of interacting queues. *Probability in the engineering and informational sciences*, 7(04):441–464, 1993.

[2] P. Billingsley. Weak convergence of measures: Applications in probability. *Philadelphia: SIAM*, 1971.

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

[4] 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, 2013.

[5] M. Draief and L. Massoulie. *Epidemics and rumours in complex networks*. Cambridge University Press, 2010.

[6] A. Eryilmaz and R. Srikant. Asymptotically tight steady-state queue length bounds implied by drift conditions. *Queueing Systems*, 72(3-4):311–359, 2012.

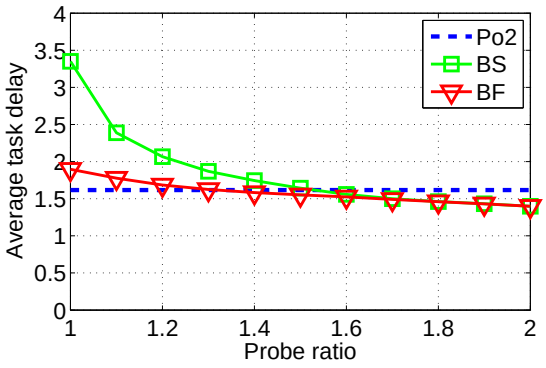

Fig. 6: The average task delays for power-of-two-choices (Po2), batch-sampling (BS) and batch-filling (BF) with $\lambda = 0.7$ and deterministic batch sizes.

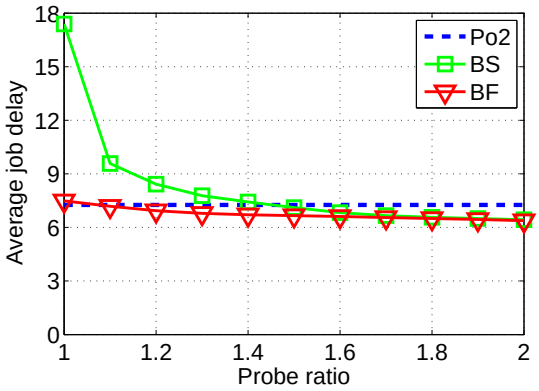

Fig. 7: The average job delays for power-of-two-choices (Po2), batch-sampling (BS) and batch-filling (BF) with $\lambda = 0.7$ and deterministic batch sizes.

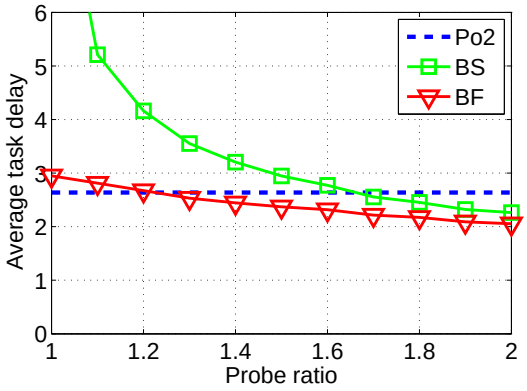

Fig. 8: The average task delays for power-of-two-choices (Po2), batch-sampling (BS) and batch-filling (BF) with $\lambda = 0.9$ and deterministic batch sizes.

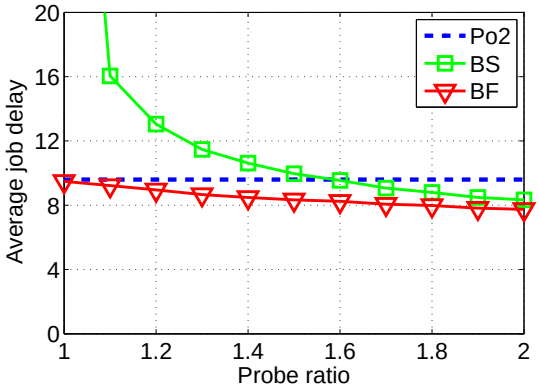

Fig. 9: The average job delays for power-of-two-choices (Po2), batch-sampling (BS) and batch-filling (BF) with $\lambda = 0.9$ and deterministic batch sizes.

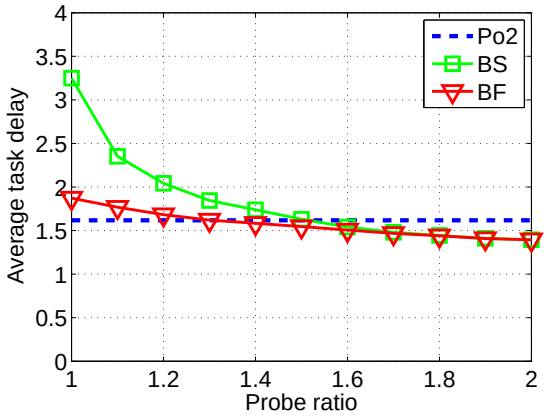

Fig. 10: The average task delays for power-of-two-choices (Po2), batch-sampling (BS) and batch-filling (BF) with $\lambda = 0.7$ with random batch sizes.

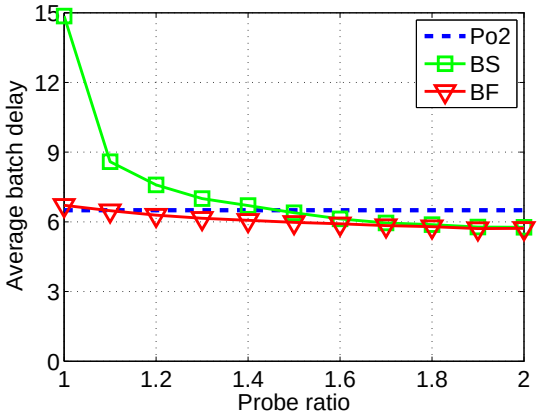

Fig. 11: The average job delays for power-of-two-choices (Po2), batch-sampling (BS) and batch-filling (BF) with $\lambda = 0.7$ with random batch sizes.

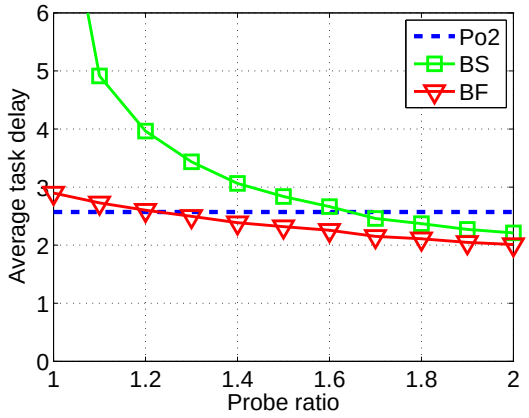

Fig. 12: The average task delays for power-of-two-choices (Po2), batch-sampling (BS) and batch-filling (BF) with $\lambda = 0.9$ with random batch sizes.

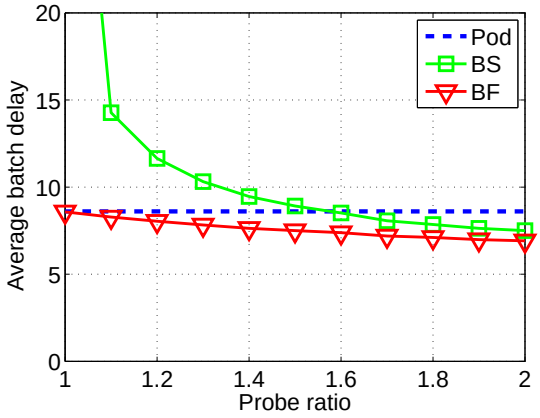

Fig. 13: The average job delays for power-of-two-choices (Po2), batch-sampling (BS) and batch-filling (BF) with $\lambda = 0.9$ with random batch sizes.

[7] S. N. Ethier and T. G. Kurtz. *Markov processes: characterization and convergence*, volume 282. John Wiley & Sons, 2009.

[8] H. K. Khalil. *Nonlinear systems*. Prentice hall Upper Saddle River, 2001.

[9] M. Mitzenmacher. Load balancing and density dependent jump Markov processes. In *Foundations of Computer Science, 1996. Proceedings., 37th Annual Symposium on*, pages 213–222. IEEE, 1996.

[10] M. Mitzenmacher. *The Power of Two Choices in Randomized Load Balancing*. PhD thesis, University of California at Berkeley, 1996.

[11] M. Mitzenmacher. The power of two choices in randomized load balancing. *Parallel and Distributed Systems, IEEE Transactions on*, 12(10):1094–1104, 2001.

[12] A. Mukhopadhyay and R. R. Mazumdar. Analysis of load balancing in large heterogeneous processor sharing systems. *arXiv preprint arXiv:1311.5806*, 2013.

[13] K. Ousterhout, P. Wendell, M. Zaharia, and I. Stoica. Sparrow: distributed, low latency scheduling. In *Proceedings of the Twenty-Fourth ACM Symposium on Operating Systems Principles*, pages 69–84. ACM, 2013.

[14] A. Richa, M. Mitzenmacher, and R. Sitaraman. The power of two random choices: A survey of techniques and results. *Combinatorial Optimization*, 9:255–304, 2001.

[15] R. Srikant and L. Ying. *Communication Networks: An Optimization, Control and Stochastic Networks Perspective*. Cambridge University Press, 2014.

[16] A.-S. Sznitman. Topics in propagation of chaos. In *Ecole d'Eté de Probabilités de Saint-Flour XIX1989*, pages 165–251. Springer, 1991.

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

[18] J. N. Tsitsiklis and K. Xu. Queueing system topologies with limited flexibility. In *Proceedings of the ACM SIGMETRICS/international conference on Measurement and modeling of computer systems*, pages 167–178. ACM, 2013.

[19] N. D. Vvedenskaya, R. L. Dobrushin, and F. I. Karpelevich. Queueing system with selection of the shortest of two queues: An asymptotic approach. *Problemy Peredachi Informatsii*, 32(1):20–34, 1996.

APPENDIX A
PROOF OF THEOREM 1

We ignore the superscript (n) of $Q_k^{(n)}(t)$ as we will focus on the nth system. Define the Lyapunov function to be

$$V(\mathbf{Q}(t)) = \sum_{k=1}^{n} Q_k^2(t).$$

Let $\mathbf{x}, \mathbf{y} \in \mathbb{N}^n$ denote the state of the Markov chains, and $q_{\mathbf{x},\mathbf{y}}$ denote the transition rate from state $\mathbf{x}$ to state $\mathbf{y}$. According to the Foster-Lyapunov theorem for continuous-time Markov chain (see, for example, Theorem 9.1.8 in [15]), we consider

$$\sum_{\mathbf{y}\neq\mathbf{x}} q_{\mathbf{x},\mathbf{y}} \left(V(\mathbf{y}) - V(\mathbf{x})\right). \tag{20}$$

Define $1 \times n$ vector $\mathbf{e}_k$ such that $\mathbf{e}_k[k] = 1$ and $\mathbf{e}_k[l] = 0$ for any $l \neq k$. Then

$$q_{\mathbf{x},\mathbf{x}-\mathbf{e}_k}\left(V((\mathbf{x}-\mathbf{e}_k)^+) - V(\mathbf{x})\right) \leq -2x_k + 1,$$

which corresponds to a departure at server k. Next define $\Psi_{\mathbf{x}}$ to be the set of possible states of the Markov chain when a batch arrival occurs when the system is in state $\mathbf{x}$, then

$$\begin{aligned}\sum_{\mathbf{y}\in\Psi_{\mathbf{x}}} q_{\mathbf{x},\mathbf{y}} \left(V(\mathbf{y}) - V(\mathbf{x})\right) &\leq_{(a)} \frac{\lambda n}{m}\left(2\frac{m}{n}\sum_k x_k + m\right)\\ &= 2\lambda\sum_k x_k + \lambda n,\end{aligned}$$

The inequality (a) can be established by comparing batch-filling with the load-balancing policy that places the m tasks to a set of randomly selected m servers, one for each server. Note that water-filling is the optimal solution to the following problem:

$$\begin{aligned}&\min_{\mathbf{a}} \textstyle\sum_{k=1}^{dm} (a_k + Q_k)^2\\ \text{subject to:}\quad &\textstyle\sum_{k=1}^{n} a_k = m\\ &a_k \in \mathbb{N} \quad \forall\ k.\end{aligned}$$

Therefore $\sum_{\mathbf{y}\in\Psi_{\mathbf{x}}} q_{\mathbf{x},\mathbf{y}} V(\mathbf{y})$ is minimized under water-filling, conditioned on the same set of dm sampled queues, and inequality (a) holds.

Therefore, we have

$$\sum_{\mathbf{y}\neq\mathbf{x}} q_{\mathbf{x},\mathbf{y}} \left(V(\mathbf{y}) - V(\mathbf{x})\right) \leq -\left(2 - 2\lambda\right)\sum_k x_k + n + \lambda n.$$

Therefore, the Markov chain is positive recurrent according to the Foster-Lyapunov theorem. Now assume the system is in

the steady state, then we have

$$\begin{aligned}0 =& \mathbb{E}\left[\sum_{\mathbf{y}\neq\mathbf{x}} q_{\mathbf{x},\mathbf{y}}\left(V(\mathbf{y})-V(\mathbf{x})\right)\right]\\ \leq& -(2-2\lambda)\,\mathbb{E}\left[\sum_k x_k\right]+n+\lambda n,\end{aligned}$$

which implies that

$$\mathbb{E}\left[\frac{1}{n}\sum_k x_k\right]\leq\frac{1+\lambda}{(2-2\lambda)}\leq\frac{1}{1-\lambda}.$$

Therefore, the theorem holds by choosing $c=1/(1-\lambda)$.

Appendix B
Proof of Lemma 2

Without loss of generality, assume server 1 has queue size i and has been probed. Now given any $j\geq 0$, define

$$X_j=\sum_{k=1}^{dm-1}\mathbb{I}_{\phi_k=j},$$

which is the number of probed servers with queue size j without including server 1, and is the summation of $dm-1$ i.i.d. Bernoulli random variables with mean π_j. We further define $\mu_j=E[X_j]=(dm-1)\pi_j$.

Consider any i such that $i\geq\bar{Q}_{\boldsymbol{\pi}}$. The probability that server 1 receives a task in water filling is upper bounded by

$$\begin{aligned}&\mathbb{E}\left[\left(\frac{m-\sum_{j=0}^{i-1}(i-j)X_j}{1+\sum_{j=0}^{i}X_j}\right)^+\right]\\ \leq&\mathbb{E}\left[\left(\frac{m-\sum_{j=0}^{\bar{Q}_{\boldsymbol{\pi}}-1}(\bar{Q}_{\boldsymbol{\pi}}-j)X_j}{1+\sum_{j=0}^{\bar{Q}_{\boldsymbol{\pi}}}X_j}\right)^+\right]\\ =&\mathbb{E}\left[\left(\frac{\frac{m}{dm-1}-\sum_{j=0}^{\bar{Q}_{\boldsymbol{\pi}}-1}(\bar{Q}_{\boldsymbol{\pi}}-j)\frac{X_j}{dm-1}}{\frac{1}{dm-1}+\sum_{j=0}^{\bar{Q}_{\boldsymbol{\pi}}}\frac{X_j}{dm-1}}\right)^+\right]\end{aligned}\tag{21}$$

which converges to

$$\left(\frac{\frac{1}{d}-\sum_{j=0}^{\bar{Q}_{\boldsymbol{\pi}}-1}(\bar{Q}_{\boldsymbol{\pi}}-j)\pi_j}{\sum_{j=0}^{\bar{Q}_{\boldsymbol{\pi}}}\pi_j}\right)^+\tag{22}$$

as $m\to\infty$ because $X_j/(dm-1)$ converges to π_j in distribution and the term inside the expectation is bounded and continuous in terms of $X_j/(dm-1)$. According to the definition of $\bar{Q}_{\boldsymbol{\pi}}$ (3), we know that

$$\frac{1}{d}-\sum_{j=0}^{\bar{Q}_{\boldsymbol{\pi}}-1}(\bar{Q}_{\boldsymbol{\pi}}-j)\pi_j\leq 0,$$

so (21) $\to 0$ and,

$$q_{i,j}=0\quad i\geq\bar{Q}_{\boldsymbol{\pi}}\text{ and }j\notin\{i,i-1\}.\tag{23}$$

Now we assume $i<\bar{Q}_{\boldsymbol{\pi}}$. In this case, the queue size of server 1 becomes $\geq Q$ ($Q>i$) after water filling with probability

$$\mathbb{E}\left[\min\left\{1,\left(\frac{m-(Q-1-i)-\sum_{j=0}^{Q-2}(Q-1-j)X_j}{1+\sum_{j=0}^{Q-1}X_j}\right)^+\right\}\right].$$

Similar to the analysis above, it can be shown that

$$\frac{m-(Q-1-i)-\sum_{j=0}^{Q-2}(Q-1-j)X_j}{1+\sum_{j=0}^{Q-1}X_j}$$

converges to

$$\frac{\frac{1}{d}-\sum_{j=0}^{Q-2}(Q-1-j)\pi_j}{\sum_{j=0}^{Q-1}\pi_j}.$$

For $Q\geq\bar{Q}_{\boldsymbol{\pi}}+1$, according to the definition of $\bar{Q}_{\boldsymbol{\pi}}$, we have

$$\frac{1}{d}-\sum_{j=0}^{\bar{Q}_{\boldsymbol{\pi}}-1}(\bar{Q}_{\boldsymbol{\pi}}-j)\pi_j\leq 0.$$

For $Q=\bar{Q}_{\boldsymbol{\pi}}$,

$$\frac{\frac{1}{d}-\sum_{j=0}^{\bar{Q}_{\boldsymbol{\pi}}-2}(\bar{Q}_{\boldsymbol{\pi}}-1-j)\pi_j}{\sum_{j=0}^{\bar{Q}_{\boldsymbol{\pi}}-1}\pi_j}=\alpha_{\boldsymbol{\pi}}.$$

For $Q\leq\bar{Q}_{\boldsymbol{\pi}}-1$,

$$\begin{aligned}&\frac{\frac{1}{d}-\sum_{j=0}^{Q-2}(Q-1-j)\pi_j}{\sum_{j=0}^{Q-1}\pi_j}\\ \geq&\frac{\frac{1}{d}-\sum_{j=0}^{\bar{Q}_{\boldsymbol{\pi}}-3}(\bar{Q}_{\boldsymbol{\pi}}-2-j)\pi_j}{\sum_{j=0}^{\bar{Q}_{\boldsymbol{\pi}}-2}\pi_j}\\ \geq&\frac{\sum_{j=0}^{\bar{Q}_{\boldsymbol{\pi}}-2}(\bar{Q}_{\boldsymbol{\pi}}-1-j)\pi_j-\sum_{j=0}^{\bar{Q}_{\boldsymbol{\pi}}-3}(\bar{Q}_{\boldsymbol{\pi}}-2-j)\pi_j}{\sum_{j=0}^{\bar{Q}_{\boldsymbol{\pi}}-2}\pi_j}\\ =&1.\end{aligned}\tag{24}$$

Therefore, for any $i<\bar{Q}_{\boldsymbol{\pi}}$ and $i\neq j$, we have

$$q_{i,j}=\begin{cases}\lambda d\alpha_{\boldsymbol{\pi}}, & \text{if } j=\bar{Q}_{\boldsymbol{\pi}}\\ \lambda d(1-\alpha_{\boldsymbol{\pi}}), & \text{if } j=\bar{Q}_{\boldsymbol{\pi}}-1\\ 0, & \text{otherwise.}\end{cases}\tag{25}$$

Hence, the lemma holds.

Appendix C
Proof of Theorem 7

Motivated by the proof in [10], we consider the following Lyapunov function

$$V(t)=\sum_{i=1}^{\infty}|s_i(t)-\hat{s}_i|.$$

Define $\epsilon_i=s_i-\hat{s}_i$, so the Lyapunov function can be written as

$$V(t)=\sum_{i=1}^{\infty}|\epsilon_i(t)|.$$

We will analyze the upper right-hand derivative

$$\frac{\mathrm{d}V(t)}{\mathrm{d}t} = \lim_{t' \to t^+} \frac{V(t') - V(t)}{t' - t}$$

in three different cases.

- In the first case, consider $\mathbf{s}$ such that $\bar{X}_{\mathbf{s}} = \bar{Q}_{BF}$. In this case, the differential equations can be written in terms of $\boldsymbol{\epsilon}$ in the following form:

$$\frac{\mathrm{d}\epsilon_i}{\mathrm{d}t} = \begin{cases} -(1+\lambda d)\epsilon_i + \epsilon_{i+1} & i \le \bar{Q}_{BF} - 1, \\ \lambda d \sum_{j=0}^{i-1} \epsilon_j - \epsilon_i + \epsilon_{i+1}, & i = \bar{Q}_{BF} \\ -\epsilon_i + \epsilon_{i+1} & \text{otherwise.} \end{cases} \tag{26}$$

  Now for $i \le \bar{Q}_{BF} - 1$,

$$\frac{\mathrm{d}|\epsilon_i|}{\mathrm{d}t} \begin{cases} = -(1+\lambda d)\epsilon_i + \epsilon_{i+1} & \text{if } \epsilon_i > 0, \\ = (1+\lambda d)\epsilon_i - \epsilon_{i+1}, & \text{if } \epsilon_i < 0, \\ = |\epsilon_{i+1}|, & \text{if } \epsilon_i = 0. \end{cases}$$

  which implies that

$$\frac{\mathrm{d}|\epsilon_i|}{\mathrm{d}t} \le -(1+\lambda d)|\epsilon_i| + |\epsilon_{i+1}| \quad i \le \bar{Q}_{BF} - 1.$$

  Similarly, we can obtain that

$$\frac{\mathrm{d}|\epsilon_i|}{\mathrm{d}t} \le \begin{cases} -|\epsilon_i| + \lambda d \sum_{j=1}^{i-1} |\epsilon_j| + |\epsilon_{i+1}| & \text{if } i = \bar{Q}_{BF}, \\ -|\epsilon_i| + |\epsilon_{i+1}| & \text{if } i > \bar{Q}_{BF}. \end{cases}$$

  Combining the results above and the fact that $s_i(t) \to 0$ as $i \to \infty$ for any t, we conclude in this case,

$$\frac{\mathrm{d}V(t)}{\mathrm{d}t} = \sum_{i=1}^{\infty} \frac{\mathrm{d}|\epsilon_i|}{\mathrm{d}t} \le -|\epsilon_1|.$$

- In the second case, consider $\mathbf{s}$ such that $\bar{X}_{\mathbf{s}} > \bar{Q}_{BF}$. Then, similar to the analysis of the first case, we have

$$\frac{\mathrm{d}\epsilon_i}{\mathrm{d}t} \le -(1+\lambda d)|\epsilon_i| + |\epsilon_{i+1}| \quad \forall\ i \le \bar{Q}_{BF} - 1. \tag{27}$$

  We next consider two subcases.

  – In the first subcase, $s_{\bar{Q}_{BF}} \ge \hat{s}_{\bar{Q}_{BF}}$. Note that $\hat{s}_i = 0$ for any $i > \bar{Q}_{BF}$, so we have

$$\begin{aligned} \sum_{i=\bar{Q}_{BF}}^{\infty} \frac{\mathrm{d}|\epsilon_i|}{\mathrm{d}t} &= \sum_{i=\bar{Q}_{BF}}^{\infty} \frac{\mathrm{d}\epsilon_i}{\mathrm{d}t} = \sum_{i=\bar{Q}_{BF}}^{\infty} \frac{\mathrm{d}s_i}{\mathrm{d}t} \\ &= \lambda - \lambda d \sum_{j=0}^{\bar{Q}_{BF}-1} (1 - s_j) - s_{\bar{Q}_{BF}} \\ &= \lambda d \sum_{j=0}^{\bar{Q}_{BF}-1} \epsilon_j - \epsilon_{\bar{Q}_{BF}} \\ &\le -|\epsilon_{\bar{Q}_{BF}}| + \lambda d \sum_{j=0}^{\bar{Q}_{BF}-1} |\epsilon_j|. \end{aligned} \tag{28}$$

  Combining (27) and (28), we obtain

$$\frac{\mathrm{d}V(t)}{\mathrm{d}t} \le -|\epsilon_1| - |\epsilon_{\bar{Q}_{BF}}| \le 0.$$

  – In the second subcase, $s_{\bar{Q}_{BF}} < \hat{s}_{\bar{Q}_{BF}}$. In this case

$$\begin{aligned} &\sum_{i=\bar{Q}_{BF}+1}^{\infty} \frac{\mathrm{d}|\epsilon_i|}{\mathrm{d}t} \\ &= \sum_{i=\bar{Q}_{BF}+1}^{\infty} \frac{\mathrm{d}\epsilon_i}{\mathrm{d}t} = \sum_{i=\bar{Q}_{BF}+1}^{\infty} \frac{\mathrm{d}s_i}{\mathrm{d}t} \\ &= \lambda - \lambda d \sum_{j=0}^{\bar{Q}_{BF}} (1 - s_j) - s_{\bar{Q}_{BF}+1} \\ &= \lambda - \lambda d \sum_{j=0}^{\bar{Q}_{BF}} (1 - \hat{s}_j) \\ &\quad + \lambda d \sum_{j=0}^{\bar{Q}_{BF}} \epsilon_j - \epsilon_{\bar{Q}_{BF}+1} \\ &\le \lambda d \sum_{j=0}^{\bar{Q}_{BF}} |\epsilon_j| - |\epsilon_{\bar{Q}_{BF}+1}|, \end{aligned} \tag{29}$$

  where the last inequality holds due to the definition of $\bar{Q}_{BF}$ and the fact that $\epsilon_{\bar{Q}_{BF}+1}(t) = s_{\bar{Q}_{BF}+1}(t) \ge 0$ for any t.
  Next, given $s_{\bar{Q}_{BF}} < \hat{s}_{\bar{Q}_{BF}}$, we have

$$\begin{aligned} &\frac{\mathrm{d}|\epsilon_{\bar{Q}_{BF}}|}{\mathrm{d}t} \\ =& -\frac{\mathrm{d}s_{\bar{Q}_{BF}}}{\mathrm{d}t} \\ =& -\lambda d + (1+\lambda d) s_{\bar{Q}_{BF}} - s_{\bar{Q}_{BF}+1} \\ =& -\lambda d + (1+\lambda d)\hat{s}_{\bar{Q}_{BF}} \\ &+ (1+\lambda d)\epsilon_{\bar{Q}_{BF}} - \epsilon_{\bar{Q}_{BF}+1} \\ \le& -(1+\lambda d)|\epsilon_{\bar{Q}_{BF}}| + |\epsilon_{\bar{Q}_{BF}+1}|, \end{aligned} \tag{30}$$

  where the last inequality holds because $\epsilon_{\bar{Q}_{BF}} < 0$, and

$$\begin{aligned} &-\lambda d + (1+\lambda d)\hat{s}_{\bar{Q}_{BF}} \\ =& -\lambda d + (1+\lambda d)\left(1 - (1-\lambda)(1+\lambda d)^{\bar{Q}_{BF}-1}\right) \\ =& 1 - (1-\lambda)(1+\lambda d)^{\bar{Q}_{BF}} \\ \le& 1 - (1-\lambda)\frac{1}{1-\lambda} = 0. \end{aligned}$$

  Combining inequalities (27), (29) and (30), we obtain

$$\frac{\mathrm{d}V(t)}{\mathrm{d}t} \le -|\epsilon_1| \le 0.$$

- In the third case, consider $\mathbf{s}$ such that $\bar{X}_{\mathbf{s}} < \bar{Q}_{BF}$. In this case, we first have

$$\sum_{i=\bar{Q}_{BF}+1}^{\infty} \frac{\mathrm{d}|\epsilon_i|}{\mathrm{d}t} = \sum_{i=\bar{Q}_{BF}+1}^{\infty} \frac{\mathrm{d}s_i}{\mathrm{d}t} = -|\epsilon_{\bar{Q}_{BF}+1}|, \tag{31}$$

  and

$$\frac{\mathrm{d}\epsilon_i}{\mathrm{d}t} \le -(1+\lambda d)|\epsilon_i| + |\epsilon_{i+1}| \quad \forall\ i < \bar{X}_{\mathbf{s}}. \tag{32}$$

  We next further consider the following subcases.

– Assume $s_{\bar{X}_\mathbf{s}} < \hat{s}_{\bar{X}_\mathbf{s}}$, so

$$\frac{\mathrm{d}|\epsilon_{\bar{X}_\mathbf{s}}|}{\mathrm{d}t} = -\lambda + \lambda d \sum_{j=0}^{\bar{X}_\mathbf{s}-1}(1-s_j) + s_{\bar{X}_\mathbf{s}} - s_{\bar{X}_\mathbf{s}+1}$$

Note that $\hat{s}_i - \hat{s}_{i+1} = \lambda d(1-\hat{s}_i)$ for any $i < \bar{Q}_{BF}$, so

$$\begin{aligned}\frac{\mathrm{d}|\epsilon_{\bar{X}_\mathbf{s}}|}{\mathrm{d}t} =& -\lambda + \lambda d \sum_{j=0}^{\bar{X}_\mathbf{s}}(1-\hat{s}_j) \\ & - \lambda d \sum_{j=0}^{\bar{X}_\mathbf{s}-1}\epsilon_j + \epsilon_{\bar{X}_\mathbf{s}} - \epsilon_{\bar{X}_\mathbf{s}+1} \\ \leq& -\lambda + \lambda d \sum_{j=0}^{\bar{X}_\mathbf{s}}(1-\hat{s}_j) \\ & + \lambda d \sum_{j=0}^{\bar{X}_\mathbf{s}-1}|\epsilon_j| - |\epsilon_{\bar{X}_\mathbf{s}}| + |\epsilon_{\bar{X}_\mathbf{s}+1}|. \end{aligned} \tag{33}$$

Next for $\bar{X}_\mathbf{s} < i < \bar{Q}_{BF}$, we have

$$\begin{aligned}\frac{\mathrm{d}\epsilon_i}{\mathrm{d}t} =& -s_i + s_{i+1} \\ =& -\hat{s}_i + \hat{s}_{i+1} - \epsilon_i + \epsilon_{i+1} \\ =& \lambda d - \lambda d\hat{s}_i - \epsilon_i + \epsilon_{i+1},\end{aligned}$$

which implies that

$$\frac{\mathrm{d}|\epsilon_i|}{\mathrm{d}t} \leq \lambda d(1-\hat{s}_i) - |\epsilon_i| + |\epsilon_{i+1}| \quad \forall\ \bar{X}_\mathbf{s} < i < \bar{Q}_{BF}. \tag{34}$$

For $i = \bar{Q}_{BF}$, we have

$$\begin{aligned}\frac{\mathrm{d}\epsilon_{\bar{Q}_{BF}}}{\mathrm{d}t} =& -s_{\bar{Q}_{BF}} + s_{\bar{Q}_{BF}+1} \\ =& -\hat{s}_{\bar{Q}_{BF}} + \hat{s}_{\bar{Q}_{BF}+1} - \epsilon_{\bar{Q}_{BF}} + \epsilon_{\bar{Q}_{BF}+1} \\ =& \lambda - \lambda d \sum_{j=0}^{\bar{Q}_{BF}-1}(1-\hat{s}_j) - \epsilon_{\bar{Q}_{BF}} + \epsilon_{\bar{Q}_{BF}+1},\end{aligned}$$

which implies that

$$\frac{\mathrm{d}|\epsilon_{\bar{Q}_{BF}}|}{\mathrm{d}t} \leq \lambda - \lambda d \sum_{j=0}^{\bar{Q}_{BF}-1}(1-\hat{s}_j) - |\epsilon_{\bar{Q}_{BF}}| + |\epsilon_{\bar{Q}_{BF}+1}|. \tag{35}$$

Summing inequalities (31) - (35), we

$$\frac{\mathrm{d}V(t)}{\mathrm{d}t} \leq -|\epsilon_1| \leq 0.$$

– Assume $s_{\bar{X}_\mathbf{s}} \geq \hat{s}_{\bar{X}_\mathbf{s}}$, then

$$\begin{aligned}\frac{\mathrm{d}|\epsilon_{\bar{X}_\mathbf{s}}|}{\mathrm{d}t} =& \lambda - \lambda d \sum_{j=0}^{\bar{X}_\mathbf{s}-1}(1-s_j) - s_{\bar{X}_\mathbf{s}} + s_{\bar{X}_\mathbf{s}+1} \\ =& \lambda - \lambda d \sum_{j=0}^{\bar{X}_\mathbf{s}}(1-s_j) + \lambda d(1-s_{\bar{X}_\mathbf{s}}) - s_{\bar{X}_\mathbf{s}} + s_{\bar{X}_\mathbf{s}+1} \\ \leq_{(a)}& \lambda d - (1+\lambda d)s_{\bar{X}_\mathbf{s}} + s_{\bar{X}_\mathbf{s}+1} \\ \leq& -(1+\lambda d)|\epsilon_{\bar{X}_\mathbf{s}}| + |\epsilon_{\bar{X}_\mathbf{s}+1}|, \end{aligned} \tag{36}$$

where inequality (a) holds due to the definition of $\bar{X}_\mathbf{s}$, and the last inequality holds because

$$\lambda d - (1+\lambda d)\hat{s}_{\bar{X}_\mathbf{s}} + \hat{s}_{\bar{X}_\mathbf{s}+1} = 0$$

when $\bar{X}_\mathbf{s} < \bar{Q}_{BF}$.
The summation of (34) and (35) yields that

$$\begin{aligned}\sum_{i=\bar{X}_\mathbf{s}+1}^{\bar{Q}_{BF}} \frac{\mathrm{d}|\epsilon_i|}{\mathrm{d}t} \leq& \lambda - \lambda d \sum_{j=0}^{\bar{X}_\mathbf{s}}(1-\hat{s}_j) - |\epsilon_{\bar{X}_\mathbf{s}+1}| + |\epsilon_{\bar{Q}_{BF}+1}| \\ \leq& \lambda - \lambda d \sum_{j=0}^{\bar{X}_\mathbf{s}}(1-s_j) \\ & - \lambda d \sum_{j=0}^{\bar{X}_\mathbf{s}}\epsilon_j - |\epsilon_{\bar{X}_\mathbf{s}+1}| + |\epsilon_{\bar{Q}_{BF}+1}| \\ \leq& \lambda d \sum_{j=0}^{\bar{X}_\mathbf{s}}|\epsilon_j| - |\epsilon_{\bar{X}_\mathbf{s}+1}| + |\epsilon_{\bar{Q}_{BF}+1}|, \end{aligned} \tag{37}$$

where the last inequality holds due to the definition of $\bar{X}_\mathbf{s}$. The summation of (31), (32), (36) and (37) yields

$$\frac{\mathrm{d}V(t)}{\mathrm{d}t} \leq -|\epsilon_1| \leq 0.$$

In a summary, we have shown that

$$\frac{\mathrm{d}V(t)}{\mathrm{d}t} \begin{cases} \leq 0, & \text{if } \mathbf{s}(t) \neq \hat{\mathbf{s}} \\ = 0, & \text{otherise.} \end{cases} \tag{38}$$

Next define $i^* = \min\{i : \epsilon_i < 0\}$. If such an i^* exists, since $\hat{s}_i = 0$ for any $i > \bar{Q}_{BF}$, $i^* \leq \bar{Q}_{BF}$. Furthermore, if $\bar{X}_s < \bar{Q}_{BF}$, then $i^* \leq \bar{X}_s$. It is easy to verify that when i^* exists,

$$\frac{\mathrm{d}|\epsilon_{i^*-1}|}{\mathrm{d}t} = \frac{\mathrm{d}\epsilon_{i^*-1}}{\mathrm{d}t} = -(1+\lambda d)|\epsilon_{i^*-1}| - |\epsilon_{i^*}|.$$

Since the following bound has been used throughout in the proof

$$\frac{\mathrm{d}|\epsilon_{i^*-1}|}{\mathrm{d}t} \leq -(1+\lambda d)|\epsilon_{i^*-1}| + |\epsilon_{i^*}|,$$

when i^* exits, we can further obtain

$$\frac{\mathrm{d}V(t)}{\mathrm{d}t} \leq -|\epsilon_1| - |\epsilon_{i^*}|, \tag{39}$$

which implies

$$\frac{\mathrm{d}V(t)}{\mathrm{d}t} \begin{cases} = 0, & \text{if } \mathbf{s}(t) = \hat{\mathbf{s}}. \\ < 0, & \text{if } s_i(t) < \hat{s}_i \text{ for some } i \\ < 0, & \text{if } s_1(t) > \hat{s}_1 \\ \leq 0, & \text{if } s_i(t) \geq \hat{s}_i\ \forall i \text{ and } s_1(t) = \hat{s}_1. \end{cases} \tag{40}$$

The result above shows that $|\mathbf{s}(t) - \hat{\mathbf{s}}|$ is non-increasing.

For any x such that $|\mathbf{x}| < \infty$, we define

$$S_x = \{y\colon |y_n| \leq |x_n| \text{ for all } n\}.$$

Then we can see that S_x is compact since we can approximate the tail with $\epsilon/2$ and the first finitely-many elements are in an equivalent Euclidean space and hence the the finite-dimensional part is totally bounded with the remaining $\epsilon/2$ as well.

Since $|\mathbf{s}(t)-\hat{\mathbf{s}}|$ is non-increasing, given a fixed $r>0$ and initial condition $\mathbf{s}(0)\in\bar{B}(\hat{\mathbf{s}},r)$, where

$$\bar{B}(\hat{s},r)=\{s\in\mathcal{X}\colon \|s-\bar{s}\|\le r\},$$

we have

$$|\mathbf{s}(t)|\le r+|\hat{\mathbf{s}}| \quad \forall\ t.$$

Since $s_1(t)\ge s_2(t)\ge\ldots\ge$, there exists $N(r)$ such that for any $i\ge N(r)$ and any $t\ge 0$,

$$s_i(t)\le\frac{1}{d},$$

$$\dot{s}_{i+1}(t)\le 0.$$

Now consider any initial state $\mathbf{s}(0)\in\mathcal{X}$. Let $r=\|\mathbf{s}(0)-\hat{\mathbf{s}}\|$ and

$$s_i'=\begin{cases}1 & \text{if } i\le N(r),\\ s_i(0) & \text{if } i>N(r).\end{cases}$$

Then $s'\in\mathcal{X}$. Let $\Omega=\bar{B}(\hat{\mathbf{s}},r)\cap S_{\mathbf{s}'}$. Since both $\bar{B}(\hat{\mathbf{s}},r)$ and $S_{s'}$ are closed and $S_{\mathbf{s}'}$ is compact, we have that Ω is compact. Also note that for any initial state $\mathbf{s}(0)\in\Omega$ we have $s(t)\in\Omega$ as well, so Ω is positive invariant and compact.

Furthermore, given $\mathbf{s}(t)$ such that $s_1(t)=\hat{s}_1$ and $s_i(t)\ge\hat{s}_i(i\ge 2)$, it can be easily shown that $s_1(t+\delta t)>\hat{s}_1$ for a sufficiently small δt unless $\mathbf{s}(t)=\hat{\mathbf{s}}$. The result can be proved by following the idea of LaSalle's invariance principle [8].

Appendix D
Proof of Theorem 8

Recall the definition of $\mathbf{\Pi}^{(n)}(t)\in\mathbb{N}^\infty$ where the i^{th} component $\Pi_i^{(n)}(t)$ is the number of servers whose queue lengths are equal to i. Since $\mathbf{\Pi}^{(n)}(t)$ can be uniquely determined by $\mathbf{\Gamma}^{(n)}(t)$ and vice versa, and $\mathbf{\Pi}^{(n)}(t)$ is a Markov chain, $\mathbf{\Gamma}^{(n)}(t)$ is a Markov chain and we have

$$\mathbf{\Gamma}^{(n)}(t)=\mathbf{\Gamma}^{(n)}(0)+\sum_{\mathbf{L}\in\mathbb{N}^\infty}\mathbf{L}N_{\mathbf{L}}\left(\int_0^t R_{\mathbf{L}}^{(n)}(\mathbf{\Gamma}^{(n)}(u))\,\mathrm{d}u\right), \tag{41}$$

where $N_{\mathbf{L}}(x)$ are independent standard Poisson processes and $R_{\mathbf{L}}^{(n)}(\mathbf{\Gamma})$ is the transition rate of the Markov chain from state $\mathbf{\Gamma}$ to state $\mathbf{\Gamma}+\mathbf{L}$. For example, given

$$\mathbf{L}=(0,-1,0,\cdots,)',$$

which corresponds to the event that there is a departure from a server with queue size 1,

$$R_{\mathbf{L}}^{(n)}(\mathbf{\Gamma}^{(n)})=\Gamma_1^{(n)}-\Gamma_2^{(n)}$$

because there are $\Gamma_1^{(n)}-\Gamma_2^{(n)}$ servers with queue size 1. Dividing by n on both sides of equation (41), we get

$$\boldsymbol{\gamma}^{(n)}(t) \;=\; \boldsymbol{\gamma}^{(n)}(0)+\sum_{\mathbf{L}\in\mathbb{N}^\infty}\frac{\mathbf{L}}{n}N_{\mathbf{L}}\left(\int_0^t R_{\mathbf{L}}^{(n)}(n\boldsymbol{\gamma}^{(n)}(u))\,\mathrm{d}u\right).$$

Now define $B_n(t)$ to be the total number of batch arrivals within time interval $[0,t]$ in the nth system. Then $B_n(t)=N(\frac{n}{m}\lambda t)$, i.e., a Poisson random variable with mean $\frac{n}{m}\lambda t$. Define event $\mathcal{B}_{n,\alpha}$ to be

$$\mathcal{B}_{n,\alpha} \;=\; \left\{B_n(t)\le(1+\alpha)\frac{n}{m}\lambda t \right.$$
$$\left. \text{and } \sum_i\gamma_i^{(n)}(0)\le(1+\alpha)\sum_i s_i(0)\right\}.$$

Applying the Chernoff bound, we obtain

$$\mathbb{P}\left(B_n(t)\le(1+\alpha)\frac{n}{m}\lambda t\right)\ge 1-e^{-\frac{n}{m}\lambda t h(\alpha)},$$

where $h(\alpha)=(1+\alpha)\log(1+\alpha)-\alpha$. Also

$$\lim_{n\to\infty}\mathbb{P}\left(\sum_i\gamma_i^{(n)}(0)\le(1+\alpha)\sum_i s_i(0)\right)=1$$

because $\boldsymbol{\gamma}^{(n)}(0)$ converges to $\mathbf{s}(0)$ in probability according to the assumption of the theorem. Thus, we have

$$\lim_{n\to\infty}\mathbb{P}\left(\mathcal{B}_{n,\alpha}\right)=1.$$

Note that $n\sum_i\gamma_i^{(n)}(u)$ is the total number of tasks in the system at time u. When $\mathcal{B}_{n,\alpha}$ occurs,

$$\max_{0\le u\le t}\sum_i\gamma_i^{(n)}(u)\le(1+\alpha)\left(\lambda t+\sum_i s_i(0)\right).$$

Define $C_\alpha=(1+\alpha)(\lambda t+\sum_i s_i(0))$. When the inequality above holds, we have

$$\gamma_i^{(n)}(u)=\sum_{j=i}^{\infty}\pi_j^{(n)}(u)\le\frac{C_\alpha}{i}\quad\forall\ 0\le u\le t,\quad\forall i, \tag{42}$$

which further implies that for $k=\left\lceil\frac{C_\alpha}{\frac{1}{2}\left(1-\frac{1}{d}\right)}\right\rceil$, we have

$$\gamma_i^{(n)}(u)\le\frac{1}{2}\left(1-\frac{1}{d}\right)\quad\forall\ 0\le u\le t,\quad\forall i\ge k. \tag{43}$$

Next we define the following four sets:

- $\mathcal{T}_n^+$: the set of $\mathbf{L}$ such that $\mathbf{L}\ge 0$, which is the set of $\mathbf{L}$ related to arrivals,
- $\mathcal{L}_n^+$: the set of $\mathbf{L}$ such that $\mathbf{L}\ge 0$ and $\mathbf{L}_i=0$ for $i\ge k+1$.
- $\mathcal{T}_n^-$: the set of $\mathbf{L}$ such that $\mathbf{L}\le 0$, which is the set of $\mathbf{L}$ related to departures.
- $\mathcal{L}_n^-$: the set of $\mathbf{L}\le 0$ and $\mathbf{L}_i=0$ for $i\ge m$.

We further define $\bar{N}_{\mathbf{L}}(a)=N_{\mathbf{L}}(a)-a$, which is a *centered* Poisson process. Then we have

$$\begin{aligned}&\boldsymbol{\gamma}^{(n)}(t)\\ =\;&\boldsymbol{\gamma}^{(n)}(0)+\\ &\sum_{\mathbf{L}\in(\mathcal{T}_n^+\cup\mathcal{T}_n^-)\setminus(\mathcal{L}_n^+\cup\mathcal{L}_n^-)}\frac{\mathbf{L}}{n}N_{\mathbf{L}}\left(\int_0^t R_{\mathbf{L}}^{(n)}(n\boldsymbol{\gamma}^{(n)}(u))\,\mathrm{d}u\right)+\\ &\sum_{\mathbf{L}\in\mathcal{L}_n^+\cup\mathcal{L}_n^-}\frac{\mathbf{L}}{n}\bar{N}_{\mathbf{L}}\left(\int_0^t R_{\mathbf{L}}^{(n)}(n\boldsymbol{\gamma}^{(n)}(u))\,\mathrm{d}u\right)+\\ &\sum_{\mathbf{L}\in\mathcal{L}_n^+\cup\mathcal{L}_n^-}\frac{\mathbf{L}}{n}\int_0^t R_{\mathbf{L}}^{(n)}(n\boldsymbol{\gamma}^{(n)}(u))\,\mathrm{d}u.\end{aligned}$$

Define $\mathbf{s}(t)$ to be the solution of the differential equations (7) with initial condition $\mathbf{s}(0)$, and $\mathbf{F}(\mathbf{s})$ such that the nonlinear differential equations in (7) are given by

$$\frac{\mathrm{d}\mathbf{s}}{\mathrm{d}t} = \mathbf{F}(\mathbf{s}).$$

Following the idea behind the proof of Kurtz's theorem (see [5] for an easy exposition), we have

$$\begin{aligned}
&\sup_{0\le u\le t}\left|\boldsymbol{\gamma}^{(n)}(u)-\mathbf{s}(u)\right| && (44)\\
\le&\left|\boldsymbol{\gamma}^{(n)}(0)-\mathbf{s}(0)\right| && (45)\\
&+\sup_{0\le u\le t}\left|\sum_{\mathbf{L}\notin(\mathcal{L}_n^+\cup\mathcal{L}_n^-)}\frac{\mathbf{L}}{n}N_{\mathbf{L}}\left(\int_0^u R_{\mathbf{L}}^{(n)}(n\boldsymbol{\gamma}^{(n)}(\tau))\,\mathrm{d}\tau\right)\right| && (46)\\
&+\sup_{0\le u\le t}\left|\sum_{\mathbf{L}\in\mathcal{L}_n^+\cup\mathcal{L}_n^-}\frac{\mathbf{L}}{n}\bar{N}_{\mathbf{L}}\left(\int_0^u R_{\mathbf{L}}^{(n)}(n\boldsymbol{\gamma}^{(n)}(\tau))\,\mathrm{d}\tau\right)\right| && (47)\\
&+\sup_{0\le u\le t}\left|\sum_{\mathbf{L}\in\mathcal{L}_n^+\cup\mathcal{L}_n^-}\frac{\mathbf{L}}{n}\int_0^u R_{\mathbf{L}}^{(n)}(n\boldsymbol{\gamma}^{(n)}(\tau))\,\mathrm{d}\tau\right.\\
&\qquad\qquad\left.-\int_0^u\mathbf{F}(\boldsymbol{\gamma}^{(n)}(\tau))\,\mathrm{d}\tau\right| && (48)\\
&+\sup_{0\le u\le t}\left|\int_0^u\mathbf{F}(\boldsymbol{\gamma}^{(n)}(\tau))\,\mathrm{d}\tau-\int_0^u\mathbf{F}(\mathbf{s}(\tau))\,\mathrm{d}\tau\right|. && (49)
\end{aligned}$$

According to Lemmas 13-15, we obtain that there exists $\bar{n}$ such that for any $n\ge\bar{n}$,

$$\begin{aligned}
&\mathbb{P}\left(\sup_{0\le u\le t}\left(\left|\boldsymbol{\gamma}^{(n)}(u)-\mathbf{s}(u)\right|-\right.\right.\\
&\qquad\left.\left.\left|\int_0^u\mathbf{F}(\boldsymbol{\gamma}^{(n)}(\tau))-\mathbf{F}(\mathbf{s}(\tau))\,\mathrm{d}\tau\right|\right)\ge 4\delta\right)\\
\le&\,\mathbb{P}\left(\left|\boldsymbol{\gamma}^{(n)}(0)-\mathbf{s}(0)\right|>\delta\right)+3(1-\mathbb{P}\left(\mathcal{B}_{n,\alpha}\right))\\
&+4m^k\max\left\{e^{-\frac{n}{m}\lambda t h(\frac{\delta}{(m+1)^k\lambda t})},e^{-nth(\frac{\delta}{2m^k t})}\right\}\\
&+\frac{\lambda t}{\delta}e^{-\frac{(d-1)^2}{2(d+1)}m}+\frac{\lambda t C_\alpha}{\delta m},
\end{aligned}$$

which converges to zero as $n\to\infty$ since $m=\Theta(\log n)$.

Let

$$\begin{aligned}
B_n=&\left\{\sup_{0\le u\le t}\left(\left|\boldsymbol{\gamma}^{(n)}(u)-\mathbf{s}(u)\right|-\right.\right.\\
&\left.\left.\left|\int_0^u\mathbf{F}(\boldsymbol{\gamma}^{(n)}(\tau))-\mathbf{F}(\mathbf{s}(\tau))\mathrm{d}\tau\right|\right)\le 4\delta\right\}.
\end{aligned}$$

Then $\mathbb{P}(B_n)\to 1$ as $n\to\infty$. When B_n occurs, for any $u\in[0,t]$,

$$\begin{aligned}
\left|\boldsymbol{\gamma}^{(n)}(u)-\mathbf{s}(u)\right|&\le 4\delta+\left|\int_0^u\mathbf{F}(\boldsymbol{\gamma}^{(n)}(\tau))-\mathbf{F}(\mathbf{s}(\tau))\mathrm{d}\tau\right|\\
&\le 4\delta+M\int_0^u\left|\boldsymbol{\gamma}^{(n)}(\tau)-\mathbf{s}(\tau)\right|\mathrm{d}\tau,
\end{aligned}$$

where the last inequality holds because $\mathbf{F}(\mathbf{s})$ is Lipschitz as shown in Lemma 16. By Gronwall's inequality we have $\left|\boldsymbol{\gamma}^{(n)}(u)-\mathbf{s}(u)\right|\le 4\delta e^{Mu}$ for any $u\in[0,t]$. Thus

$$\mathbb{P}\left(\sup_{0\le u\le t}\left|\boldsymbol{\gamma}^{(n)}(u)-\mathbf{s}(u)\right|\le 4\delta e^{Mt}\right)\ge\mathbb{P}(B_n)\to 1$$

as $n\to\infty$.

Lemma 13.

$$\mathbb{P}\left((46)>\delta\right)\le\frac{\lambda t}{\delta}e^{-\frac{(d-1)^2}{2(d+1)}m}+\frac{\lambda t C_\alpha}{\delta m}+2(1-\mathbb{P}\left(\mathcal{B}_{n,\alpha}\right)).$$

Proof. Note that $\mathbf{L}\in\mathcal{T}_n^+\setminus\mathcal{L}_n^+$ occurs when a task is dispatched to a queue with size at least k. Under condition (43), when a batch arrival occurs,

$$\begin{aligned}
&\mathbb{P}\left(\bigcup_{\mathbf{L}\in\mathcal{T}_n^+\setminus\mathcal{L}_n^+}\{n\boldsymbol{\gamma}\to n\boldsymbol{\gamma}+\mathbf{L}\}\right)\\
\le&\;\mathbb{P}\left(dm-Z_k<m\right)=\mathbb{P}\left(Z_k>(d-1)m\right)\\
\le&\;e^{-\frac{(d-1)^2}{2(d+1)}m},
\end{aligned}$$

where Z_k is the number of servers probed with queue size at least k and the last inequality is obtained from the Hoeffding's inequality for sampling without replacement. Therefore, we have

$$\begin{aligned}
&\mathbb{P}\left(\sup_{0\le u\le t}\left|\sum_{\mathbf{L}\in\mathcal{T}_n^+\setminus\mathcal{L}_n^+}\frac{\mathbf{L}}{n}N_{\mathbf{L}}\left(\int_0^u R_{\mathbf{L}}^{(n)}(n\boldsymbol{\gamma}^{(n)}(\tau))\,\mathrm{d}\tau\right)\right|\ge\delta\right)\\
\le&\,\mathbb{P}\left(\sup_{0\le u\le t}\frac{m}{n}N\left(\int_0^u\sum_{\mathbf{L}\in\mathcal{T}_n^+\setminus\mathcal{L}_n^+}R_{\mathbf{L}}^{(n)}(n\boldsymbol{\gamma}^{(n)}(\tau))\,\mathrm{d}\tau\right)\ge\delta\right)\\
\le_{(a)}&\,\mathbb{P}\left(\frac{m}{n}N\left(\int_0^t\sum_{\mathbf{L}\in\mathcal{T}_n^+\setminus\mathcal{L}_n^+}R_{\mathbf{L}}^{(n)}(n\boldsymbol{\gamma}^{(n)}(\tau))\,\mathrm{d}\tau\right)\ge\delta\right)\\
\le&\,\mathbb{P}\left(\frac{m}{n}N\left(\int_0^t\sum_{\mathbf{L}\in\mathcal{T}_n^+\setminus\mathcal{L}_n^+}R_{\mathbf{L}}^{(n)}(n\boldsymbol{\gamma}^{(n)}(\tau))\,\mathrm{d}\tau\right)\ge\delta\bigcap\mathcal{B}_{n,\alpha}\right)\\
&+1-\mathbb{P}(\mathcal{B}_{n,\alpha})\\
\le&\,\mathbb{P}\left(\frac{m}{n}N\left(\frac{n}{m}\lambda t e^{-\frac{(d-1)^2}{2(d+1)}m}\right)\ge\delta\right)+1-\mathbb{P}(\mathcal{B}_{n,\alpha})\\
\le&\,\frac{\lambda t}{\delta}e^{-\frac{(d-1)^2}{2(d+1)}m}+1-\mathbb{P}(\mathcal{B}_{n,\alpha}),
\end{aligned}$$

where inequality (a) holds because $N(t)$ is nondecreasing with t and the last inequality is obtained from the Markov inequality.

Similarly, we can also obtain

$$\begin{aligned}
&\mathbb{P}\left(\sup_{0\le u\le t}\left|\sum_{\mathbf{L}\in\mathcal{T}_n^-\setminus\mathcal{L}_n^-}\frac{\mathbf{L}}{n}N_{\mathbf{L}}\left(\int_0^u R_{\mathbf{L}}^{(n)}(n\boldsymbol{\gamma}^{(n)}(\tau))\,\mathrm{d}\tau\right)\right|\ge\delta\right)\\
\le&\mathbb{P}\left(\frac{1}{n}N\left(\int_0^t\sum_{\mathbf{L}\in\mathcal{T}_n^-\setminus\mathcal{L}_n^-}R_{\mathbf{L}}^{(n)}(n\boldsymbol{\gamma}^{(n)}(\tau))\,\mathrm{d}\tau\right)\ge\delta\right)\\
\le&\mathbb{P}\left(\mathcal{B}(n,\alpha)\bigcap\frac{1}{n}N\left(\int_0^t\sum_{\mathbf{L}\in\mathcal{T}_n^-\setminus\mathcal{L}_n^-}R_{\mathbf{L}}^{(n)}(n\boldsymbol{\gamma}^{(n)}(\tau))\,\mathrm{d}\tau\right)\ge\delta\right)\\
&+1-\mathbb{P}(\mathcal{B}_{n,\alpha})\\
\le&\mathbb{P}\left(\frac{1}{n}N\left(n\lambda t\frac{C_\alpha}{m}\right)\ge\delta\right)+1-\mathbb{P}(\mathcal{B}_{n,\alpha})\\
\le&\frac{\lambda tC_\alpha}{\delta m}+1-\mathbb{P}(\mathcal{B}_{n,\alpha}).
\end{aligned}$$

□

Lemma 14.

$$\mathbb{P}\left((47)>\delta\right)\le 4m^k\max\left\{e^{-\frac{n}{m}\lambda th(\frac{\delta}{(m+1)^k\lambda t})},e^{-nth(\frac{\delta}{2m^kt})}\right\}.$$

Proof. Note that $|\mathcal{L}_n^+\cup\mathcal{L}_n^-|\le m^k+m\le 2m^k$. For $\mathbf{L}\in\mathcal{L}_n^+$,

$$\begin{aligned}
&\mathbb{P}\left(\sup_{0\le u\le t}\left|\frac{\mathbf{L}}{n}\bar{N}_{\mathbf{L}}\left(\int_0^u R_{\mathbf{L}}^{(n)}(n\boldsymbol{\gamma}^{(n)}(\tau))\,\mathrm{d}\tau\right)\right|>\frac{\delta}{2m^k}\right)\\
\le&\mathbb{P}\left(\sup_{0\le u\le t}\frac{m}{n}\left|\bar{N}_{\mathbf{L}}\left(\frac{n}{m}\lambda u\right)\right|>\frac{\delta}{2m^k}\right)\\
\le&2e^{-\frac{n}{m}\lambda th(\frac{\delta}{2m^k\lambda t})},
\end{aligned}$$

where the last inequality follows from Proposition 5.2 in [5]. Similarly, for $\mathbf{L}\in\mathcal{L}_n^-$,

$$\begin{aligned}
&\mathbb{P}\left(\sup_{0\le u\le t}\left|\frac{\mathbf{L}}{n}\bar{N}_{\mathbf{L}}\left(\int_0^u R_{\mathbf{L}}^{(n)}(n\boldsymbol{\gamma}^{(n)}(\tau))\,\mathrm{d}\tau\right)\right|>\frac{\delta}{2m^k}\right)\\
\le&\mathbb{P}\left(\sup_{0\le u\le t}\frac{1}{n}\left|\bar{N}_{\mathbf{L}}\left(nu\right)\right|>\frac{\delta}{2m^k}\right)\\
\le&2e^{-nth(\frac{\delta}{2m^kt})}.
\end{aligned}$$

Combining the results above and using the union bound, we obtain

$$\mathbb{P}\left((47)>\delta\right)\le 4m^k\max\left\{e^{-\frac{n}{m}\lambda th(\frac{\delta}{(m+1)^k\lambda t})},e^{-nth(\frac{\delta}{2m^kt})}\right\}.$$

□

Lemma 15. *There exists $\bar{n}$ such that for any $n\ge\bar{n}$,*

$$\mathbb{P}\left((48)>\delta\right)\le 1-\mathbb{P}\left(\mathcal{B}_{n,\alpha}\right).$$

Proof. To study (48) under condition (42), we define

$$\mathbf{F}^{(n)}(\boldsymbol{\gamma})=\frac{1}{n}\sum_{\mathbf{L}\in\mathcal{L}_n^+\cup\mathcal{L}_n^-}\mathbf{L}R_{\mathbf{L}}^{(n)}(n\boldsymbol{\gamma}),$$

and consider

$$\left|\mathbf{F}^{(n)}(\boldsymbol{\gamma})-\mathbf{F}(\boldsymbol{\gamma})\right|=\sum_i\left|F_i^{(n)}(\boldsymbol{\gamma})-F_i(\boldsymbol{\gamma})\right|. \tag{50}$$

We divide the analysis into the following cases:

- For $i>m$, $L_i=0$ for any $\mathbf{L}\in\mathcal{L}_n^+\cup\mathcal{L}_n^-$, which implies $F_i^{(n)}(\boldsymbol{\gamma})=0$ and
$$\sum_{i>m}\left|F_i^{(n)}(\boldsymbol{\gamma})-F_i(\boldsymbol{\gamma})\right|=\sum_{i>m}F_i(\boldsymbol{\gamma})=\gamma_{m+1}\le\frac{C_\alpha}{m}.$$
- For $m\ge i>k$,
$$F_i^{(n)}(\boldsymbol{\gamma}) \;=\; -\gamma_i+\gamma_{i+1},$$
which implies that
$$\left|F_i^{(n)}(\boldsymbol{\gamma})-F_i(\boldsymbol{\gamma})\right|=0.$$
- For $k\ge i$,
$$\begin{aligned}
F_i^{(n)}(\boldsymbol{\gamma}) &= \frac{1}{n}\left(\frac{\lambda n}{m}\,\mathbb{E}[D_i|\boldsymbol{\gamma}]-n\gamma_i+n\gamma_{i+1}\right)\\
&= \lambda d\,\mathbb{E}\left[\frac{D_i}{dm}\middle|\boldsymbol{\gamma}\right]-\gamma_i+\gamma_{i+1},
\end{aligned}$$
where D_i is a random variable denoting the change in the number of servers with queue size at least i after water filling. Therefore,
$$\left|F_i^{(n)}(\boldsymbol{\gamma})-F_i(\boldsymbol{\gamma})\right|=\lambda d\,\mathbb{E}\left[\frac{D_i}{dm}\middle|\boldsymbol{\gamma}\right].$$
Recall Z_i to be the number of probed servers with queue size at least i, so D_i is a function of Z_j ($j\le i$). Specifically,
$$D_i=\min\left\{dm-Z_i,\left(m-\sum_{j=0}^{i-1}(dm-Z_j)\right)^+\right\}. \tag{51}$$
Therefore,
$$\frac{D_i}{dm}=\min\left\{\frac{dm-Z_i}{dm},\left(\frac{1}{d}-\sum_{j=0}^{i-1}\left(1-\frac{Z_j}{dm}\right)\right)^+\right\}.$$
Applying the Hoeffding's inequality for sampling without replacement, we have that
$$\mathbb{P}\left(|Z_i-\gamma_i dm|\ge\sqrt{m\log m}\right)\le 2e^{-2\frac{\log m}{d}}=\frac{2}{m^{2/d}},$$
which implies that
$$\mathbb{P}\left(|Z_i-\gamma_i dm|\le\sqrt{m\log m}\quad\forall i\le k\right)\ge 1-\frac{2k}{m^{2/d}}.$$
Given $|Z_i-\gamma_i dm|\le\sqrt{m\log m}$ for all $i\le k$, we can obtain
$$\begin{aligned}
&\left|\mathbb{E}\left[\frac{D_i}{dm}\middle|\boldsymbol{\gamma}\right]-\min\left\{1-\gamma_i,\left(\frac{1}{d}-\sum_{j=0}^{i-1}(1-\gamma_j)\right)^+\right\}\right|\\
\le&\frac{k\sqrt{\log m}}{d\sqrt{m}}.
\end{aligned}$$

By summarizing the cases above, we obtain that under condition (42)

$$\left|\mathbf{F}^{(n)}(\boldsymbol{\gamma})-\mathbf{F}(\boldsymbol{\gamma})\right|\le\frac{C_\alpha}{m}+\frac{k\sqrt{\log m}}{d\sqrt{m}}.$$

Therefore, given δ, there exists m_δ such that for any $m \geq m_\delta$,

$$\begin{aligned}&\sup_{0\leq u\leq t}\left|\int_0^u \mathbf{F}^{(n)}(\boldsymbol{\gamma}^{(n)}(\tau))\,\mathrm{d}\tau - \int_0^u \mathbf{F}(\boldsymbol{\gamma}^{(n)}(\tau))\,\mathrm{d}\tau\right|\\ \leq& t\left(\frac{C_\alpha}{m} + \frac{k\sqrt{\log m}}{d\sqrt{m}}\right) \leq \delta.\end{aligned}$$

So for sufficient large n,

$$\begin{aligned}&\mathbb{P}\left(\sup_{0\leq u\leq t}\left|\int_0^u \mathbf{F}^{(n)}(\boldsymbol{\gamma}^{(n)}(\tau))\,\mathrm{d}\tau - \int_0^u \mathbf{F}(\boldsymbol{\gamma}^{(n)}(\tau))\,\mathrm{d}\tau\right| > \delta\right)\\ \leq& 1 - \mathbb{P}\left(\mathcal{B}_{n,\alpha}\right).\end{aligned}$$

□

Lemma 16. $\mathbf{F}(\mathbf{s})$ *is Lipschitz.*

Proof. Consider $\mathbf{s}, \mathbf{s}' \in \mathbb{N}^\infty$. Without loss of generality $\bar{X}_{\mathbf{s}} \leq \bar{X}_{\mathbf{s}'}$. Define

$$h_i(\mathbf{s}) = F_i(\mathbf{s}) - s_i + s_{i+1}.$$

Then

$$\begin{aligned}&|\mathbf{F}(\mathbf{s}) - \mathbf{F}(\mathbf{s}')|\\ =&\sum_{i=1}^\infty |F_i(\mathbf{s}) - F_i(\mathbf{s}')|\\ \leq&\sum_{i=1}^\infty \left(|s_i - s_i'| + |s_{i+1} - s_{i+1}'| + |h_i(\mathbf{s}) - h_i(\mathbf{s}')|\right)\\ \leq&2|\mathbf{s} - \mathbf{s}'| + \sum_{i=1}^\infty |h_i(\mathbf{s}) - h_i(\mathbf{s}')|.\end{aligned}$$

Recall that $F_i(\mathbf{s}) = -s_i + s_{i+1}$ for $i > \bar{X}_{\mathbf{s}}$ and $F_i(\mathbf{s}) = \lambda d - (1+\lambda d)s_i + s_{i+1}$ for $i < \bar{X}_{\mathbf{s}}$, so

$$\begin{aligned}&|\mathbf{F}(\mathbf{s}) - \mathbf{F}(\mathbf{s}')|\\ \leq&2|\mathbf{s} - \mathbf{s}'| + \lambda d \sum_{i=1}^{\bar{X}_{\mathbf{s}}-1} |s_i - s_i'| + \sum_{i=\bar{X}_{\mathbf{s}}}^{\bar{X}_{\mathbf{s}'}} |h_i(\mathbf{s}) - h_i(\mathbf{s}')|.\end{aligned}$$

We next consider two cases. If $h_{\bar{X}_{\mathbf{s}}}(\mathbf{s}) \leq h_{\bar{X}_{\mathbf{s}}}(\mathbf{s}')$, then

$$\begin{aligned}&\sum_{i=\bar{X}_{\mathbf{s}}}^{\bar{X}_{\mathbf{s}'}} |h_i(\mathbf{s}) - h_i(\mathbf{s}')|\\ =&\lambda d - \lambda d s_{\bar{X}_{\mathbf{s}}}' - \lambda + \lambda d \sum_{j=1}^{\bar{X}_{\mathbf{s}}-1}(1 - s_j)\\ &+ \lambda d \sum_{i=\bar{X}_{\mathbf{s}}+1}^{\bar{X}_{\mathbf{s}'}-1}(1 - s_i')\\ &+ \lambda - \lambda d \sum_{j=1}^{\bar{X}_{\mathbf{s}'}-1}(1 - s_j')\\ =&\lambda d \sum_{j=1}^{\bar{X}_{\mathbf{s}}-1}(s_j' - s_j)\\ \leq&\lambda d \sum_{j=1}^{\bar{X}_{\mathbf{s}}-1}|s_j' - s_j|\\ \leq&\lambda d|\mathbf{s} - \mathbf{s}'|.\end{aligned}$$

If $h_{\bar{X}_{\mathbf{s}}}(\mathbf{s}) > h_{\bar{X}_{\mathbf{s}}}(\mathbf{s}')$, then

$$\begin{aligned}&\sum_{i=\bar{X}_{\mathbf{s}}}^{\bar{X}_{\mathbf{s}'}} |h_i(\mathbf{s}) - h_i(\mathbf{s}')|\\ =&-\lambda d + \lambda d s_{\bar{X}_{\mathbf{s}}}' + \lambda d - \lambda d s_{\bar{X}_{\mathbf{s}}} + \lambda - \lambda d \sum_{j=1}^{\bar{X}_{\mathbf{s}}}(1 - s_j)\\ &+ \lambda - \lambda d \sum_{j=1}^{\bar{X}_{\mathbf{s}}}(1 - s_j') - \lambda + \lambda d \sum_{j=1}^{\bar{X}_{\mathbf{s}}}(1 - s_j)\\ &+ \lambda - \lambda d \sum_{j=1}^{\bar{X}_{\mathbf{s}}}(1 - s_j)\\ \leq&\lambda d|s_{\bar{X}_{\mathbf{s}}}' - s_{\bar{X}_{\mathbf{s}}}| + \lambda d \sum_{j=1}^{\bar{X}_{\mathbf{s}}}|s_j' - s_j|\\ \leq&2\lambda d|\mathbf{s} - \mathbf{s}'|,\end{aligned}$$

where the first inequality holds because

$$\lambda - \lambda d \sum_{j=1}^{\bar{X}_{\mathbf{s}}}(1 - s_j) \leq 0$$

according to the definition of $\bar{X}_{\mathbf{s}}$.

Combining the results above, we obtain that

$$|\mathbf{F}(\mathbf{s}) - \mathbf{F}(\mathbf{s}')| \leq (2 + 3\lambda d)|\mathbf{s} - \mathbf{s}'|.$$

Therefore, the lemma holds. □

Appendix E
Proof of Theorem 9

Let $\hat{\mathbf{X}}^{(n_k)}$ denote the weak convergence subsequence in assumption (A4). By (A1) and the Skorohod representation theorem, there exists $\{\tilde{\mathbf{X}}^{(n_k)}\}$ and $\tilde{\mathbf{X}}$ such that

- $\tilde{\mathbf{X}}^{(n_k)} =_d \hat{\mathbf{X}}^{(n_k)}$,
- $\tilde{\mathbf{X}} =_d \bar{\mathbf{X}}$, and
- $\tilde{\mathbf{X}}^{(n_k)}$ converges to $\tilde{\mathbf{X}}$ almost surely.

Now let $\mathbf{X}^{(n_k)}(0) = \tilde{\mathbf{X}}^{(n_k)}$, i.e., the n_kth system starts at a random initial condition specified by its stationary distribution, which implies that

$$\mathbf{X}^{(n_k)}(t) =_d \tilde{\mathbf{X}}^{(n_k)} \quad \forall t.$$

Denote by $\mathbf{X}(t)$ the random state of the dynamical system starting from the random initial condition $\tilde{\mathbf{X}}$. According to (A2), for any deterministic initial condition in $\mathcal{X}$,

$$\mathbf{X}^{(n_k)}(t) \xrightarrow{w} \mathbf{X}(t).$$

By the definition of weak convergence, for a bounded continuous function f,

$$\begin{aligned}&\lim_{n\to\infty} \mathbb{E}\left[f(\mathbf{X}^{(n_k)}(t))|\mathbf{X}^{(n_k)}(0) = \tilde{\mathbf{X}}^{(n_k)}\right]\\ =&\mathbb{E}\left[f(\mathbf{X}(t))|\mathbf{X}^{(n_k)}(0) = \tilde{\mathbf{X}}\right].\end{aligned}$$

Since f is bounded, further by the bounded convergence theorem and the fact that $\mathbb{P}\left(\tilde{\mathbf{X}} \in \mathcal{X}\right) = \mathbb{P}\left(\bar{\mathbf{X}} \in \mathcal{X}\right) = 1$, we have

$$\lim_{n\to\infty} \mathbb{E}\left[f(\mathbf{X}^{(n_k)}(t))\right] = \mathbb{E}\left[f(\mathbf{X}(t))\right],$$

which implies that $\mathbf{X}^{(n_k)}(t)$ converges weakly to $\mathbf{X}(t)$ for any t.

Since $\mathbf{X}^{(n_k)}(t) =_d \hat{\mathbf{X}}^{(n_k)}\ \forall t$, we further have $\mathbf{X}(t) =_d \bar{\mathbf{X}}\ \forall t$. Now according to (A3), the dynamical system converges to $\hat{\mathbf{X}}$ starting from any initial condition in $\mathcal{X}$, which implies $\mathbf{X}(t)$ converges to $\hat{\mathbf{X}}$ almost surely and also implies that $\mathbf{X}(t)$ converges weakly to $\hat{\mathbf{X}}$. Therefore, $\bar{\mathbf{X}}$ is a point mass at $\hat{\mathbf{X}}$, which implies that $\hat{\mathbf{X}}^{(n_k)}$ converges weakly to $\hat{\mathbf{X}}$. Since this holds for any convergent subsequence, the theorem holds.

Appendix F
Uniform Convergence of the Series (16)

First since $\sum_{i=1}^{\infty}\mathbb{E}\left[\tilde{\gamma}_i\right] \le c$ according to (15) and $\mathbb{E}\left[\tilde{\gamma}_i\right] \ge 0$ for all i, the sequence $s_b = \sum_{i=1}^{b}\mathbb{E}\left[\tilde{\gamma}_i\right]$ is bounded above and increasing, so $\bar{s} = \lim_{b\to\infty} s_b$ exits. Therefore, given any ϵ, there exists $\tilde{b}_\epsilon$ such that for any $b \ge \tilde{b}_\epsilon$,

$$\sum_{i=b+1}^{\infty}\mathbb{E}\left[\tilde{\gamma}_i\right] = \bar{s} - s_b \le \epsilon. \tag{52}$$

Next we establish an upper bound on

$$\sum_{i=b+1}^{\infty}\mathbb{E}\left[\hat{\gamma}_i^{(n_k)}\right]$$

using the Lyapunov-drift analysis at the steady state. Since

$$\sum_{i=1}^{\infty}\mathbb{E}\left[\hat{\gamma}_i^{(n_k)}\right] \le c$$

for any n_k and $\mathbb{E}\left[\hat{\gamma}_i^{(n_k)}\right]$ is decreasing,

$$\mathbb{E}\left[\hat{\gamma}_i^{(n_k)}\right] \le \frac{c}{i}$$

for any i and any n_k. According to Markov's inequality, we have

$$\mathbb{P}\left(\hat{\gamma}_i^{(n_k)} \ge \frac{d-1}{2d}\right) \le \frac{c}{i}\frac{2d}{d-1}. \tag{53}$$

Now consider the n_kth system and define a Lyapunov function to be

$$V(\mathbf{Q}(t)) = \sum_{j=1}^{n_k}((Q_j(t)-b+1)^+)^2,$$

where $b > 0$ and the superscript (n_k) of Q is ignored to simplify the notation. Let $\mathbf{x}, \mathbf{y} \in \mathbb{N}^{n_k}$ denote the state of the Markov chains, and $q_{\mathbf{x},\mathbf{y}}$ denote the transition rate from state $\mathbf{x}$ to state $\mathbf{y}$. According to the Foster-Lyapunov theorem for continuous-time Markov chain (see, for example, Theorem 9.1.8 in [15]), we consider

$$\sum_{\mathbf{y}\neq\mathbf{x}} q_{\mathbf{x},\mathbf{y}}\left(V(\mathbf{y}) - V(\mathbf{x})\right). \tag{54}$$

Recall that $\mathbf{e}_j$ is a $1 \times n_k$ vector such that $\mathbf{e}_j[j] = 1$ and $\mathbf{e}_j[l] = 0$ for any $l \neq j$. Then

$$\begin{aligned}
&q_{\mathbf{x},\mathbf{x}-\mathbf{e}_j}\left(V((\mathbf{x}-\mathbf{e}_j)^+) - V(\mathbf{x})\right)\\
=&\begin{cases} 0, & \text{if } x_j \le b-1\\ -1, & \text{if } x_j = b\\ -2(x_j-b)-1, & \text{if } x_j > b\end{cases}\\
\le& -2(x_j-b)^+,
\end{aligned}$$

which corresponds to a departure at server j.

Next define $\Psi_{\mathbf{x}}$ to be the set of possible states of the Markov chain when a batch arrival occurs and the system is in state $\mathbf{x}$. Consider

$$\sum_{\mathbf{y}\in\Psi_{\mathbf{x}}} q_{\mathbf{x},\mathbf{y}}\left(V(\mathbf{y}) - V(\mathbf{x})\right).$$

We note that batch-filling is one of the optimal solutions to the following problem:

$$\begin{aligned}
&\min_{\mathbf{a}} \textstyle\sum_{k=1}^{dm}\left((a_k + x_k - b + 1)^+\right)^2\\
\text{subject to:}\quad &\textstyle\sum_{k=1}^{n} a_k = m\\
&a_k \in \mathbb{N} \quad \forall\ k,
\end{aligned}$$

where $\{x_k\}_{k=1,\cdots,dm}$ are the sizes of the probed dm queues. In other words, given $\mathbf{x}$ *and the set of* dm *probed servers*, the batch-filling minimize $V(\mathbf{y})$. This can be proved by showing that any task assignment can be modified to the batch-filling solution, by iteratively moving new tasks from large queues to small queues, without increasing the value of the objective function.

Given any $b > 2$, we consider the following two cases.

- First consider $\mathbf{x}$ such that

$$\mathbf{x} \in \Omega_b := \left\{\mathbf{x} : \sum_j \mathbb{I}_{x_j \le b-2} \ge n_k\frac{d+1}{2d}\right\}.$$

In other words, at least $(d+1)/2d$ fraction of servers with queue size at most $b-2$.
Define $\tilde{\Psi}_{\mathbf{x}}$ to be the set of possible states of the Markov chain under *batch-sampling* when a batch arrival occurs and the system is in state $\mathbf{x}$, and $\tilde{q}_{\mathbf{x},\mathbf{y}}$ to be the corresponding transition rate. Since batch-filling minimizes $V(\mathbf{y})$ for any given set of probed server, we have

$$\sum_{\mathbf{y}\in\Psi_{\mathbf{x}}} q_{\mathbf{x},\mathbf{y}}\left(V(\mathbf{y}) - V(\mathbf{x})\right) \le \sum_{\mathbf{y}\in\tilde{\Psi}_{\mathbf{x}}} \tilde{q}_{\mathbf{x},\mathbf{y}}\left(V(\mathbf{y}) - V(\mathbf{x})\right).$$

Now under batch-sampling, a server may receive one (and at most one) task if it is probed. Consider server j such that $x_j \ge b-1$. Server j is probed with probability dm/n_k, and will receive one task if it is among the m least loaded queues in the md probed queues. Conditioned on server j is probed, define G_{b-2} to be the number of probed servers with queue size at most $b-2$ among the other $dm-1$ servers. According to Hoeffding's inequality for sampling without replacement, we get

$$\mathbb{P}(G_{b-2} < m) \le e^{-\frac{(d-1)^2}{2d}m}.$$

Therefore, we conclude that

$$\begin{aligned}
&\sum_{\mathbf{y}\in\Psi_{\mathbf{x}}} q_{\mathbf{x},\mathbf{y}}\left(V(\mathbf{y}) - V(\mathbf{x})\right)\\
\le&\sum_j \frac{\lambda n_k}{m}\times\frac{dm}{n_k}\times e^{-\frac{(d-1)^2}{2d}m}\left(2(x_j-b+1)+1\right)^+\\
\le&\sum_j \lambda d e^{-\frac{(d-1)^2}{2d}m}\left(2(x_j-b)^+ + 3\right).
\end{aligned}$$

Note that $(y_j - b + 1)^+ = 0$ for any queue j such that $x_j \leq b-2$ since each server is given at most one task under batch-sampling.

- Consider $\mathbf{x}$ such that $\mathbf{x} \notin \Omega_b$, i.e.,

$$\sum_j \mathbb{I}_{x_j \leq b-2} < n_k \frac{d+1}{2d}.$$

In this case, we compare batch-filling with the randomized load-balancing algorithm that places m tasks in a set of randomly selected m servers, one for each server. According to the analysis in the proof of Theorem 1, we have

$$\begin{aligned}
&\sum_{\mathbf{y}\in\Psi_{\mathbf{x}}} q_{\mathbf{x},\mathbf{y}} \left(V(\mathbf{y}) - V(\mathbf{x})\right) \\
&\leq \frac{\lambda n_k}{m}\left(m + 2\frac{m}{n_k}\sum_j (x_j - b + 1)^+\right) \\
&= \lambda n_k + \sum_j 2\lambda (x_j - b + 1)^+ \\
&\leq 3\lambda n_k + \sum_j 2\lambda (x_j - b)^+
\end{aligned}$$

Combining the results above, we have that

$$\begin{aligned}
&\sum_{\mathbf{y}\neq\mathbf{x}} q_{\mathbf{x},\mathbf{y}} \left(V(\mathbf{y}) - V(\mathbf{x})\right) \\
\leq &\sum_j -2\left(1 - \lambda \max\{1, de^{-\frac{(d-1)^2}{2d}m}\}\right)(x_j - b)^+ + \\
&3n_k\lambda de^{-\frac{(d-1)^2}{2d}m}\,\mathbb{P}(x \in \Omega_b) + 2\lambda n_k\,\mathbb{P}(x \notin \Omega_b).
\end{aligned}$$

Recall that the Markov chain is positive recurrent according to Theorem 1. Assume the system is in the steady state, then we have

$$0 = \mathbb{E}\left[\sum_{\mathbf{y}\neq\mathbf{x}} q_{\mathbf{x},\mathbf{y}} \left(V(\mathbf{y}) - V(\mathbf{x})\right)\right],$$

which implies that

$$\begin{aligned}
&\mathbb{E}\left[\frac{1}{n_k}\sum_j (x_j - b)^+\right] \\
\leq &\frac{3\lambda de^{-\frac{(d-1)^2}{2d}m}\,\mathbb{P}(\mathbf{x} \in \Omega_b) + 2\lambda\,\mathbb{P}(\mathbf{x} \notin \Omega_b)}{2\left(1 - \lambda\max\{1, de^{-\frac{(d-1)^2}{2d}m}\}\right)}.
\end{aligned}$$

Now given any $0 < \epsilon < 1$, define n_ϵ such that for any $n \geq n_\epsilon$

$$\frac{3\lambda de^{-\frac{(d-1)^2}{2d}m}}{2(1-\lambda)} \leq \frac{\epsilon}{2} \text{ and } de^{-\frac{(d-1)^2}{2d}m} \leq 1.$$

Such n_ϵ exists because $m = \Theta(\log n)$ and m is increasing function of n. Furthermore, define b_ϵ such that for any $b \geq b_\epsilon$ and in the steady state of any n_kth system,

$$\frac{2\lambda\,\mathbb{P}(\mathbf{x} \notin \Omega_b)}{2(1-\lambda)} \leq \frac{\epsilon}{2}.$$

Such b_ϵ exists, independent of n_k, due to inequality (53).

Furthermore, we note that

$$\mathbb{E}\left[\frac{1}{n_k}\sum_j (x_j - b)^+\right] = \sum_{i=b+1}^{\infty} \mathbb{E}\left[\hat{\gamma}_i^{(n_k)}\right].$$

Therefore, given any $0 < \epsilon < 1$, there exist b_ϵ and n_ϵ such that for any $b \geq b_\epsilon$ and $n_k \geq n_\epsilon$, the following inequality holds:

$$\sum_{i=b+1}^{\infty} \mathbb{E}\left[\hat{\gamma}_i^{(n_k)}\right] \leq \epsilon.$$

Combining the result above and result (52), we conclude that given any $0 < \epsilon < 1$, for any $n_k \geq n_{\epsilon/2}$ and $b \geq \max\{b_{\epsilon/2}, \tilde{b}_{\epsilon/2}\}$,

$$\begin{aligned}
&\textstyle\sum_{i=b+1}^{\infty} \mathbb{E}\left[\left|\hat{\gamma}_i^{(n_k)} - \tilde{\gamma}_i\right|\right] \\
\leq\ &\textstyle\sum_{i=b+1}^{\infty} \mathbb{E}\left[\hat{\gamma}_i^{(n_k)}\right] + \sum_{i=b+1}^{\infty} \mathbb{E}\left[\tilde{\gamma}_i\right] \leq \epsilon,
\end{aligned}$$

which concludes the proof.

Appendix G
Proof of Corollary 12

To simplify the notation, we assume $k = 2$, the analysis for $k > 2$ is almost identical and hence omitted here. Now for the nth system, we define $\mathcal{S}^{(n)} = \{i : i > 2\}$, i.e., the set of all servers except servers 1 and 2. We consider the following Markov chain $(Q_1^{(n)}(t), Q_2^{(n)}(t), \boldsymbol{\eta}^{(n)}(t))$, where

$$\begin{aligned}
\eta_i^{(n)}(t) &= \frac{\sum_{i\in\mathcal{S}^{(n)}} \mathbf{I}_{Q_i^{(n)}(t)=i}}{n-2} \\
&= \frac{\Pi_i^{(n)}(t) - \mathbf{I}_{Q_1^{(n)}(t)=i} - \mathbf{I}_{Q_2^{(n)}(t)=i}}{n-2},
\end{aligned}$$

i.e., the fraction of servers with queue size i in $\mathcal{S}^{(n)}$. Recall that $Q_1^{(n)}(t)$ is the queue length of the first server in the nth system, $Q_2^{(n)}(t)$ is the queue length of the second server in the nth system, and $\hat{Q}_1^{(n)}$ and $\hat{Q}_2^{(n)}$ are the queue lengths in the steady state. Denote by

$$\pi^{(n)}(x, y, \boldsymbol{\eta}) = \mathbb{P}\left((\hat{Q}_1^{(n)}, \hat{Q}_2^{(n)}, \hat{\boldsymbol{\eta}}^{(n)}) = (x, y, \boldsymbol{\eta})\right),$$

i.e., the stationary distribution of the Markov chain. For the nth system, the global balance equation for a given state $(x, y, \boldsymbol{\eta})$ is

$$\begin{aligned}
&\pi^{(n)}(x, y, \boldsymbol{\eta}) \sum_{(\tilde{x},\tilde{y},\tilde{\boldsymbol{\eta}})\neq(x,y,\boldsymbol{\eta})} r^{(n)}_{(x,y,\boldsymbol{\eta})}(\tilde{x}, \tilde{y}, \tilde{\boldsymbol{\eta}}) \\
= &\sum_{(\tilde{x},\tilde{y},\tilde{\boldsymbol{\eta}})\neq(x,y,\boldsymbol{\eta})} \pi^{(n)}(\tilde{x}, \tilde{y}, \tilde{\boldsymbol{\eta}})\, r^{(n)}_{(\tilde{x},\tilde{y},\tilde{\boldsymbol{\eta}})}(x, y, \boldsymbol{\eta}),
\end{aligned}$$

where $r^{(n)}_{(x,y,\boldsymbol{\eta})}(\tilde{x}, \tilde{y}, \tilde{\boldsymbol{\eta}})$ is the transition rate from state $(x, y, \boldsymbol{\eta})$ to $(\tilde{x}, \tilde{y}, \tilde{\boldsymbol{\eta}})$ in the nth system, which further implies that

$$\begin{aligned}
&\sum_{\boldsymbol{\eta}} \sum_{(\tilde{x},\tilde{y},\tilde{\boldsymbol{\eta}})\neq(x,y,\boldsymbol{\eta})} \pi^{(n)}(x, y, \boldsymbol{\eta})\, r^{(n)}_{(x,y,\boldsymbol{\eta})}(\tilde{x}, \tilde{y}, \tilde{\boldsymbol{\eta}}) \\
= &\sum_{\boldsymbol{\eta}} \sum_{(\tilde{x},\tilde{y},\tilde{\boldsymbol{\eta}})\neq(x,y,\boldsymbol{\eta})} \pi^{(n)}(\tilde{x}, \tilde{y}, \tilde{\boldsymbol{\eta}})\, r^{(n)}_{(\tilde{x},\tilde{y},\tilde{\boldsymbol{\eta}})}(x, y, \boldsymbol{\eta}). \qquad (55)
\end{aligned}$$

Note that for $(\tilde{x},\tilde{y},\tilde{\boldsymbol{\eta}})$ such that $\tilde{x}=x$ and $\tilde{y}=y$,

$$\begin{aligned}&\sum_{\boldsymbol{\eta}}\sum_{\tilde{\boldsymbol{\eta}}\neq\boldsymbol{\eta}}\pi^{(n)}\left(x,y,\boldsymbol{\eta}\right)r^{(n)}_{(x,y,\boldsymbol{\eta})}(x,y,\tilde{\boldsymbol{\eta}})\\=&\sum_{\boldsymbol{\eta}}\sum_{\tilde{\boldsymbol{\eta}}\neq\boldsymbol{\eta}}\pi^{(n)}\left(x,y,\tilde{\boldsymbol{\eta}}\right)r^{(n)}_{(x,y,\tilde{\boldsymbol{\eta}})}(x,y,\boldsymbol{\eta})\end{aligned}\tag{56}$$

by exchanging the notations $\boldsymbol{\eta}$ and $\tilde{\boldsymbol{\eta}}$. Furthermore, to transit to a state with $\tilde{x}>x$ and $\tilde{y}>y$, server 1 and server 2 need to be both probed, so

$$\sum_{\tilde{x}>x,\tilde{y}>y}r^{(n)}_{(x,y,\boldsymbol{\eta})}(\tilde{x},\tilde{y},\tilde{\boldsymbol{\eta}})\leq\lambda\frac{n}{m}\frac{dm(dm-1)}{n(n-1)}=\Theta\left(\frac{m}{n}\right),$$

which implies that

$$\sum_{\boldsymbol{\eta}}\pi^{(n)}\left(x,y,\boldsymbol{\eta}\right)\sum_{\tilde{x}>x,\tilde{y}>y}r^{(n)}_{(x,y,\boldsymbol{\eta})}(\tilde{x},\tilde{y},\tilde{\boldsymbol{\eta}})=O\left(\frac{m}{n}\right)\tag{57}$$

since $\sum_{\boldsymbol{\eta}}\pi^{(n)}\left(x,y,\boldsymbol{\eta}\right)\leq 1$. Similarly, we have

$$\sum_{\boldsymbol{\eta}}\sum_{\tilde{x}<x,\tilde{y}<y}\pi^{(n)}\left(\tilde{x},\tilde{y},\tilde{\boldsymbol{\eta}}\right)r^{(n)}_{(\tilde{x},\tilde{y},\tilde{\boldsymbol{\eta}})}(x,y,\boldsymbol{\eta})=O\left(\frac{m}{n}\right).\tag{58}$$

Note that

$$r^{(n)}_{(x,y,\boldsymbol{\eta})}(x-1,y,\boldsymbol{\eta})=r^{(n)}_{(x,y,\boldsymbol{\eta})}(x,y-1,\boldsymbol{\eta})=1,$$

so

$$\begin{aligned}&\sum_{\boldsymbol{\eta}}\pi^{(n)}\left(x,y,\boldsymbol{\eta}\right)r^{(n)}_{(x,y,\boldsymbol{\eta})}(x-1,y,\boldsymbol{\eta})\\=&\sum_{\boldsymbol{\eta}}\pi^{(n)}\left(x,y,\boldsymbol{\eta}\right)r^{(n)}_{(x,y,\boldsymbol{\eta})}(x,y-1,\boldsymbol{\eta})\\=&\pi^{(n)}(x,y),\end{aligned}\tag{59}$$

$$\sum_{\boldsymbol{\eta}}\pi^{(n)}\left(x+1,y,\boldsymbol{\eta}\right)r^{(n)}_{(x+1,y,\boldsymbol{\eta})}(x,y,\boldsymbol{\eta})=\pi^{(n)}\left(x+1,y\right),\tag{60}$$

and

$$\sum_{\boldsymbol{\eta}}\pi^{(n)}\left(x,y+1,\boldsymbol{\eta}\right)r^{(n)}_{(x,y+1,\boldsymbol{\eta})}(x,y,\boldsymbol{\eta})=\pi^{(n)}\left(x,y+1\right).\tag{61}$$

Now we consider

$$\begin{aligned}&\sum_{\boldsymbol{\eta}}\sum_{\tilde{x}>x}\sum_{\tilde{\boldsymbol{\eta}}}\pi^{(n)}\left(x,y,\boldsymbol{\eta}\right)r^{(n)}_{(x,y,\boldsymbol{\eta})}(\tilde{x},y,\tilde{\boldsymbol{\eta}})\\=&\sum_{\tilde{x}>x}\sum_{\boldsymbol{\eta}}\pi^{(n)}\left(x,y,\boldsymbol{\eta}\right)\sum_{\tilde{\boldsymbol{\eta}}}r^{(n)}_{(x,y,\boldsymbol{\eta})}(\tilde{x},y,\tilde{\boldsymbol{\eta}})\\=&\pi^{(n)}(x,y)\sum_{\tilde{x}>x}\sum_{\boldsymbol{\eta}}\pi^{(n)}\left(\boldsymbol{\eta}|x,y\right)\sum_{\tilde{\boldsymbol{\eta}}}r^{(n)}_{(x,y,\boldsymbol{\eta})}(\tilde{x},y,\tilde{\boldsymbol{\eta}})\\=&\pi^{(n)}(x,y)\sum_{\tilde{x}>x}\mathbb{E}_{\boldsymbol{\eta}}\left[r^{(n)}_{(x,y,\boldsymbol{\eta})}(\tilde{x},y)\middle|x,y\right],\end{aligned}$$

where $r^{(n)}_{(x,y,\boldsymbol{\eta})}(\tilde{x},y)=\sum_{\tilde{\boldsymbol{\eta}}}r^{(n)}_{(x,y,\boldsymbol{\eta})}(\tilde{x},y,\tilde{\boldsymbol{\eta}})$.

Note that

$$\begin{aligned}&r^{(n)}_{(x,y,\boldsymbol{\eta})}(\tilde{x},y)=\\&\mathbb{E}\left[\min\left\{1,\left(\frac{\frac{1}{d}-(\tilde{x}-1-x)-\sum_{j=0}^{\tilde{x}-2}(\tilde{x}-1-j)\frac{X_j}{dm}}{1+\sum_{j=0}^{\tilde{x}-1}\frac{X_j}{dm}}\right)^+\right\}\middle|\boldsymbol{\eta}\right]\\&-\mathbb{E}\left[\min\left\{1,\left(\frac{\frac{1}{d}-(\tilde{x}-x)-\sum_{j=0}^{\tilde{x}-1}(\tilde{x}-j)\frac{X_j}{dm}}{1+\sum_{j=0}^{\tilde{x}}\frac{X_j}{dm}}\right)^+\right\}\middle|\boldsymbol{\eta}\right],\end{aligned}$$

which implies that

$$\begin{aligned}&\mathbb{E}_{\boldsymbol{\eta}}\left[r^{(n)}_{(x,y,\boldsymbol{\eta})}(\tilde{x},y)\middle|x,y\right]=\\&\mathbb{E}\left[\min\left\{1,\left(\frac{\frac{1}{d}-(\tilde{x}-1-x)-\sum_{j=0}^{\tilde{x}-2}(\tilde{x}-1-j)\frac{X_j}{dm}}{1+\sum_{j=0}^{\tilde{x}-1}\frac{X_j}{dm}}\right)^+\right\}\right]\\&-\mathbb{E}\left[\min\left\{1,\left(\frac{\frac{1}{d}-(\tilde{x}-x)-\sum_{j=0}^{\tilde{x}-1}(\tilde{x}-j)\frac{X_j}{dm}}{1+\sum_{j=0}^{\tilde{x}}\frac{X_j}{dm}}\right)^+\right\}\right].\end{aligned}$$

It is easy to show that X_j/dm converges weakly to $\hat{\gamma}_i$ because $\boldsymbol{\eta}$ converges weakly to $\hat{\boldsymbol{\gamma}}$. Hence, we have

$$\lim_{n\to\infty}\mathbb{E}_{\boldsymbol{\eta}}\left[r^{(n)}_{(x,y,\boldsymbol{\eta})}(\tilde{x},y)\middle|x,y\right]=q_{x,\tilde{x}}(\hat{\gamma}),\quad x<\tilde{x}\leq\bar{Q}_{BF},$$

and

$$\lim_{n\to\infty}\sum_{\tilde{x}>\bar{Q}_{BF}}\mathbb{E}_{\boldsymbol{\eta}}\left[r^{(n)}_{(x,y,\boldsymbol{\eta})}(\tilde{x},y)\middle|x,y\right]=0,\quad x<\bar{Q}_{BF},$$

where $q_{x,\tilde{x}}(\hat{\gamma})$ and $\bar{Q}_{BF}$ are defined in Lemma 2. Since $0\leq\pi^{(n)}(x,y)\leq 1$ and $0\leq\mathbb{E}_{\boldsymbol{\eta}}\left[r^{(n)}_{(x,y,\boldsymbol{\eta})}(\tilde{x},y)\middle|x,y\right]\leq d\lambda$, we can conclude that

$$\begin{aligned}&\lim_{n\to\infty}\sum_{\boldsymbol{\eta}}\sum_{\tilde{x}>x}\sum_{\tilde{\boldsymbol{\eta}}}\pi^{(n)}\left(x,y,\boldsymbol{\eta}\right)r^{(n)}_{(x,y,\boldsymbol{\eta})}(\tilde{x},y,\tilde{\boldsymbol{\eta}})\\=&\lim_{n\to\infty}\pi^{(n)}(x,y)\sum_{\tilde{x}>x}\mathbb{E}_{\boldsymbol{\eta}}\left[r^{(n)}_{(x,y,\boldsymbol{\eta})}(\tilde{x},y)\middle|x,y\right]\\=&\lim_{n\to\infty}\pi^{(n)}(x,y)\sum_{\bar{Q}_{BF}\geq\tilde{x}>x}\lim_{n\to\infty}\mathbb{E}_{\boldsymbol{\eta}}\left[r^{(n)}_{(x,y,\boldsymbol{\eta})}(\tilde{x},y)\middle|x,y\right]\\&+\lim_{n\to\infty}\pi^{(n)}(x,y)\lim_{n\to\infty}\sum_{\tilde{x}>x\geq\bar{Q}_{BF}}\mathbb{E}_{\boldsymbol{\eta}}\left[r^{(n)}_{(x,y,\boldsymbol{\eta})}(\tilde{x},y)\middle|x,y\right]\\=&\pi(x,y)\sum_{\bar{Q}_{BF}\geq\tilde{x}>x}q_{x,\tilde{x}}(\hat{\gamma}).\end{aligned}$$

Similarly, we have

$$\begin{aligned}&\lim_{n\to\infty}\sum_{\boldsymbol{\eta}}\sum_{\tilde{x}<x}\sum_{\tilde{\boldsymbol{\eta}}}\pi^{(n)}\left(\tilde{x},y,\tilde{\boldsymbol{\eta}}\right)r^{(n)}_{(\tilde{x},y,\tilde{\boldsymbol{\eta}})}(x,y,\boldsymbol{\eta})\\=&\sum_{\tilde{x}<x\leq\bar{Q}_{BF}}\pi(\tilde{x},y)q_{\tilde{x},x}(\hat{\boldsymbol{\gamma}}).\end{aligned}$$

Summarizing the results above, (55) implies that

$$\begin{aligned}&\pi(x,y)\left(\sum_{\bar{Q}_{BF}\geq\tilde{x}>x}q_{x,\tilde{x}}(\hat{\boldsymbol{\gamma}})+\sum_{\bar{Q}_{BF}\geq\tilde{y}>y}q_{y,\tilde{y}}(\hat{\boldsymbol{\gamma}})\right)\\=&\sum_{\tilde{x}<x\leq\bar{Q}_{BF}}\pi(\tilde{x},y)q_{\tilde{x},x}(\hat{\boldsymbol{\gamma}})+\sum_{\tilde{y}<y\leq\bar{Q}_{BF}}\pi(x,\tilde{y})q_{\tilde{y},y}(\hat{\boldsymbol{\gamma}}).\end{aligned}$$

It is easy to verify the equation above is the detailed balance equation for two independent and identical Markov chains with

transition rates given in Lemma 2, and the unique solution therefore is $\pi(x, y) = \hat{\pi}_x \hat{\pi}_y$ for $\hat{\boldsymbol{\pi}}$ defined in (4). This means that queue 1 and queue 2 are independent in the large-system limit.