Performance Analysis of Low-Complexity Resource-Allocation Algorithms in Stochastic Networks Using Fluid Models

by

Xiaohan Kang

A Dissertation Presented in Partial Fulfillment
of the Requirements for the Degree
Doctor of Philosophy

Approved October 2015 by the
Graduate Supervisory Committee:

Lei Ying, Chair
Douglas Cochran
Jim Dai
Junshan Zhang

ARIZONA STATE UNIVERSITY

December 2015

ABSTRACT

Resource allocation in communication networks aims to assign various resources such as power, bandwidth and load in a fair and economic fashion so that the networks can be better utilized and shared by the communicating entities. The design of efficient resource-allocation algorithms is, however, becoming more and more challenging due to the precipitously increasing scale of the networks. This thesis strives to understand how to design such low-complexity algorithms with performance guarantees.

In the first part, the link scheduling problem in wireless ad hoc networks is considered. The scheduler is charge of finding a set of wireless data links to activate at each time slot with the considerations of wireless interference, traffic dynamics, network topology and quality-of-service (QoS) requirements. Two different yet essential scenarios are investigated: the first one is when each packet has a specific deadline after which it will be discarded; the second is when each packet traverses the network in multiple hops instead of leaving the network after a one-hop transmission. In both scenarios the links need to be carefully scheduled to avoid starvation of users and congestion on links. One greedy algorithm is analyzed in each of the two scenarios and performance guarantees in terms of throughput of the networks are derived.

In the second part, the load-balancing problem in parallel computing is studied. Tasks arrive in batches and the duty of the load balancer is to place the tasks on the machines such that minimum queueing delay is incurred. Due to the huge size of modern data centers, sampling the status of all machines may result in significant overhead. Consequently, an algorithm based on limited queue information at the machines is examined and its asymptotic delay performance is characterized and it is shown that the proposed algorithm achieves the same delay with remarkably less sampling overhead compared to the well-known power-of-two-choices algorithm.

Two messages of the thesis are the following: greedy algorithms can work well in a stochastic setting; the fluid model can be useful in "derandomizing" the system and reveal the nature of the algorithm.

TABLE OF CONTENTS

CHAPTER Page

LIST OF TABLES

LIST OF FIGURES

Figure Page

LIST OF NOTATION

$\mathbb{Z}$	The set of integers
$\mathbb{N}$	The set of nonnegative integers
$\mathbb{R}$	The set of real numbers
$\mathbb{R}_+$	The set of nonnegative real numbers
$(x_n \colon n \in \mathbb{N})$	A sequence
(x_n)	Shorthand of a sequence
$\mathcal{L}(X)$	The probability distribution/law of the random element X
$\{x \in X \colon P(x)\}$	Set builder: the subset of X that satisfies the property P
A^{T}	The transpose of matrix A
e	The all-one column vector with appropriate dimension

Chapter 1

INTRODUCTION

Over the past decades, stochastic modeling has been successfully adopted to analyze various computing and communication networks for performance evaluation. Researchers developed optimal or near-optimal scheduling or routing algorithms in terms of throughput, delay and other metrics in various scenarios based on the stochastic models. However, in practice, many of the algorithms are seldom used due to their computational complexity or communication overhead. Instead, many heuristic algorithms are preferred in reality based on their empirical good performance and low complexity, and are becoming even more attractive for large scale computing and communication systems. These heuristic algorithms are often difficult to analyze in the stochastic settings, so the gap between theory and application for the low-complexity greedy algorithms needs to be filled in order to understand the systems and to provide performance guarantee.

In Chapter 2 we consider the stability of the longest-queue-first (LQF) scheduling policy in wireless networks with multihop traffic under the one-hop interference model. Although it is well known that the back-pressure algorithm achieves the maximal stability, its computational complexity is prohibitively high. So instead we consider LQF, a low-complexity scheduling algorithm which has been shown to have near-optimal throughput performance in many networks with single-hop traffic flows. We are interested in the performance of LQF for multihop traffic flows, in which scenario the coupling between queues due to multihop traffic flows makes the local-pooling-factor analysis difficult to perform. Using the fluid-limit techniques, we show that

LQF achieves the maximal stability for linear networks with multihop traffic and a single destination on the boundary of the network under the one-hop interference model.

In Chapter 3 we consider the problem of scheduling real-time traffic in wireless networks. We consider ad hoc wireless networks with general conflict graph–based interference model and single-hop traffic. Each packet is associated with a deadline and will be dropped if it is not transmitted before the deadline. The number of packet arrivals in each time slot and the maximum delay before the deadline are independent and identically distributed across time. We require a minimum fraction of packets to be delivered. At each link, we assume the link keeps track of the difference between the minimum number of packets that need to be delivered so far and the number of packets that are actually delivered, which we call the deficit. The largest-deficit-first (LDF) policy schedules links in descending order according to their deficit values, which is a variation of the longest-queue-first (LQF) policy for non-real-time traffic. We prove that the efficiency ratio of LDF, which is the fraction of the throughput region that LDF can achieve for given traffic distributions, can be lower bounded by a quantity that we call the real-time local-pooling factor (R-LPF). We further prove that a lower bound on the R-LPF can be related to the weighted sum of the service rates, with a special case of $1/(\beta+1)$ by considering the uniform weight, where β is the interference degree of the conflict graph. We also propose a heuristic consensus algorithm that can be used to obtain a good weight vector for such lower bounds for given network topology.

In Chapter 4 we consider the randomized load-balancing problem with large number of servers. 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. 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 dramatically reduces the sample complexity compared to previously proposed algorithms.

Chapter 2

LINK SCHEDULING IN WIRELESS NETWORKS WITH MULTIHOP TRAFFIC

2.1 Background

The scheduling problem in wireless networks with multihop traffic has gained significant attention over the last few decades. One fundamental goal of the design of scheduling policies, among many others, is to decide which set of links to schedule at each time slot in accordance with the underlying interference model so that the system is stable. The back-pressure algorithm has been proved to be throughput-optimal for general multihop-traffic settings (Tassiulas and Ephremides, 1992); i.e., it stabilizes the network as long as the arrival rates are within the network throughput region. The algorithm, however, requires the network to solve a maximum-weight independent set problem at each time instance and requires the nodes to exchange queue lengths with their neighbors constantly.

In this chapter, we study the stability of the longest-queue-first (LQF) scheduling policy, which selects links according to the queue lengths in a greedy fashion. LQF has been extensively studied as a low-complexity approximation of the MaxWeight scheduling, and has great throughput and delay performance in many networks. The conditions under which LQF is throughput-optimal have been established by Dimakis and Walrand (2006) and the performance guarantee of LQF in general networks has been characterized by Joo *et al.* (2009b) and estimated under different scenarios (Joo *et al.*, 2009b,a; Birand *et al.*, 2012; Leconte *et al.*, 2011). An asynchronous version of LQF has also been proved to be throughput-optimal under the local-pooling condition

by Maguluri *et al.* (2014) However, these results all assume single-hop traffic flows in the network. For networks with multihop traffic, transmitted packets at one link may become the *internal* arrivals to another link. Hence links with small queues may affect the ones with large queues by providing internal arrivals, which makes it difficult to analyze the system using local-pooling-factor analysis since the maximum fluid among the large queues does not always decrease, as we will show in the example in Section 2.3.2. Although Brzezinski *et al.* (2008) developed conditions for networks with multihop traffic under which a back-pressure-based greedy algorithm achieves the maximal throughput, the performance of LQF for networks with multihop traffic flows is still unknown. We are interested in tackling the problem of throughput performance of LQF, since it can shed light on the implementation of low-complexity scheduling algorithms in wireless multihop networks. While the original LQF is a centralized scheduling policy, a queue-length-based CSMA-type algorithm, called D-GMS, has been proposed by Ni *et al.* (2012) to approximate LQF in a distributed fashion, which does not require constant exchange of queue lengths. Thus, LQF can be used as the foundation to a low-complexity, distributed scheduling algorithm.

We focus on the scheduling problem under multihop traffic in a simple network, i.e., a linear network with single destination on the boundary of the network (also known as a tandem network) and one-hop interference model (also known as primary or node-exclusive interference model)[1]. Such networks have been well studied in the literature to provide insights in understanding the fundamental scaling properties of multihop traffic (Tassiulas and Ephremides, 1994; Stolyar, 2011; Bui *et al.*, 2011; Hellings *et al.*, 2011). In particular, Stolyar (2011) and Bui *et al.* (2011) analyzed the asymptotic

[1]While the one-hop interference model is indeed a mathematical simplification of wireless interference in reality, it has been used as a reasonable approximation to Bluetooth or FH-CDMA networks (Joo *et al.*, 2009b). It may also be used to model half-duplex communication.

delay performance of the back-pressure algorithm in large linear networks when no interference is present. To the best of our knowledge, however, neither throughput nor delay performance guarantee of LQF has been obtained under multihop traffic scenario for linear networks.

This chapter proves the throughput optimality of LQF in a special case of linear networks. While the result is only for linear networks, it is the first step to understand the following question: to achieve throughput optimality in a wireless network with multihop traffic flows that have fixed routes, is it sufficient to use queue lengths as weights instead of using differential queues? If the answer is positive, then nodes do not need to constantly exchange queue lengths, which eliminates a significant amount of communication overhead.

The novelty in this chapter lies in the techniques we adopt to show the stability of the fluid model after the standard construction of fluid limits. Instead of using an explicit Lyapunov function, we follow the observations from the simulation trajectories of an example network and examine the evolution of the states of the deterministic fluid limits. We first show that the system will eventually stay in the state where the fluid at the first queue is zero. Then by combining the first two queues into one using a coupled-network argument, we reduce the size of the network by one and conclude that fluids at all queues eventually become zero by induction.[2]

The chapter is organized as follows. We introduce the basic model in Section 2.2. In Section 2.3 we present our result of throughput optimality of LQF, as well as an intuitive example, formal notation and network equations, construction of fluid limits, and the proof.

[2]We remark that while a properly chosen Lyapunov function would suffice to show stability, finding such a function may be difficult.

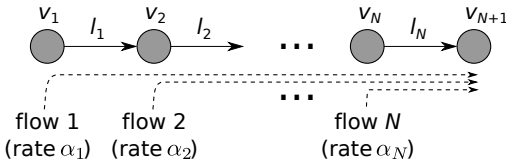


Figure 1: A linear network with N links

Note: The ith dashed line indicates the flow with source node v_i and destination node v_{N+1} and exogenous packet arrival rate α_i.

Figure 2: The one-hop interference graph of the linear network

2.2 Model

Consider a linear network represented by a directed graph $G = (V, L)$ with $|V| = N+1$ nodes and $|L| = N$ links as shown in Figure 1. Let $V = \{v_1, v_2, \ldots, v_{N+1}\}$ and $L = \{l_1, l_2, \ldots, l_N\}$, where l_i is the link from node v_i to node v_{i+1}. We assume v_i is the origin node of flow f_i with exogenous (or external) packet arrival rate α_i for $1 \leq i \leq N$, and all flows have the same destination v_{N+1}. In the chapter we focus on the one-hop interference model, so the interference graph is as shown in Figure 2.

We assume time is slotted, and in each time slot a subset of the links can be scheduled. Once scheduled, a packet at link l_i is transmitted from node v_i to node v_{i+1} and join the queue at node v_{i+1} if it has not reached the destination v_{N+1}, or leave the network otherwise. As a result, besides exogenous packet arrivals, there can also be *internal* packet arrivals to a node according to the schedule of other links.

The scheduler decides a subset of the links $s \subseteq L$ to be activated in every time slot, called a schedule, such that the schedule is feasible (scheduled links do not interfere with each other) and maximal (no other link can be added to the schedule without

violating the feasibility constraint), and then the queue length at each transmitter in the activated subset reduces by 1 if there are any packets to schedule, or remains 0 otherwise. A schedule (also known as an activation set) s is represented by an *activation vector* m, which is a binary column vector with N elements. According to the interference model shown in Figure 2, a schedule s is feasible if no two adjacent links are activated at the same time; i.e., the activation vector m does not contain two consecutive 1's. For example, the activation vectors for the four maximal schedules when $N = 5$ are 10101, 10010, 01010 and 01001. The number of maximal schedules grows exponentially with N.

In the chapter we are interested in LQF with arbitrary tie-breaking rules, and we define it as follows. At each time slot, let Z_i be the queue length at link l_i for $1 \leq i \leq N$. The set of links are sorted with arbitrary tie-breaks such that $Z_{\sigma_1} \geq Z_{\sigma_2} \geq \cdots \geq Z_{\sigma_N}$, where $(\sigma_1, \sigma_2, \ldots, \sigma_N)$ is the sorted index vector. LQF starts with the schedule $\mathcal{E} = \{\sigma_1\}$, and proceeds to consider $i = 2, 3, \ldots, N$ inductively and appends σ_i to $\mathcal{E}$ if σ_i does not interfere with any link that is already in $\mathcal{E}$. This procedure ends after the link l_{σ_N} is considered and the resulting schedule $\mathcal{E}$ is the schedule chosen by LQF.

2.3 Stability

In this section we analyze the stability property of LQF in the linear network under the one-hop interference model. We say the system is *stable* if the queue length process, as a Markov process, is positive recurrent. We first state the main theorem with the proof outline and an illustrative example, and then proceed with the formal proof.

2.3.1 Main Result

Theorem 1. *LQF is throughput-optimal on linear networks with multihop traffic and a single destination on the boundary of the network under the one-hop interference model.* ◇

Theorem 1 states that LQF can stabilize a linear multihop traffic network with a single destination on the boundary and the one-hop interference model. Thus using queue lengths instead of queue differences is sufficient. This result may also shed light on the throughput performance of LQF in other networks with multihop traffic, in which the routes are fixed.

The proof consists of the following steps. We first follow the standard construction of the fluid limits. Then we show that eventually the fluids should be such that each fluid is less than or equal to at least one neighbor fluid; i.e., no fluid dominates all its neighbors. After that we prove that the first fluid must decrease with rate at least $\epsilon > 0$. Finally we use a coupled-network argument to show that all fluids eventually go to zero under admissible arrival rates, which implies throughput optimality.

We next demonstrate the key ideas of the proof using an example.

2.3.2 Three-Link Linear Network

We consider the simple linear network example with four nodes $\{v_1, v_2, v_3, v_4\}$ and three links $\{l_1, l_2, l_3\}$. Suppose flow f_i has origin v_i and destination v_4 with Bernoulli arrival of rate α_i for $i = 1, 2, 3$. The interference is such that two adjacent links cannot be scheduled at the same time, so either $\{l_1, l_3\}$ or $\{l_2\}$ is scheduled in each time slot. Let $Z_i(n)$ be the queue length on link l_i at time slot n. Then at each time slot, the

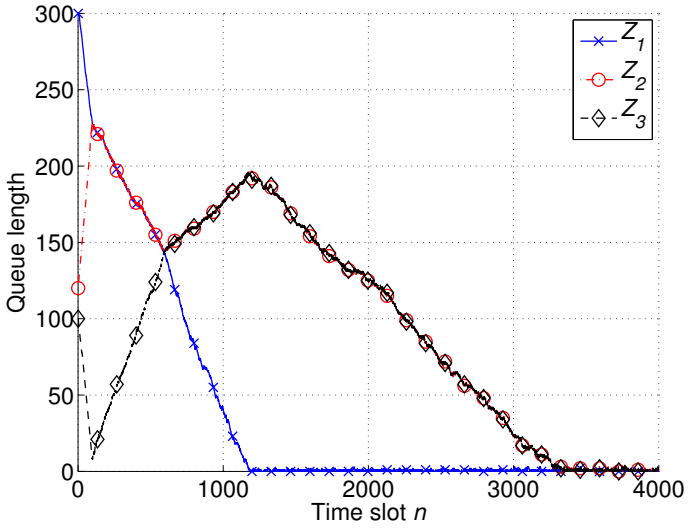


Figure 3: Queue evolution of the three-link linear network under LQF

LQF scheduler first selects the longest queue with arbitrary tie-breaking, and then chooses either $\{l_1, l_3\}$ or $\{l_2\}$ according to the first selection.

A typical queue evolution graph for the three-link linear network under LQF is shown in Figure 3. Here the initial queue lengths are $Z_1(0) = 300$, $Z_2(0) = 120$ and $Z_3(0) = 100$, with arrival rates $\alpha_1 = 0.25$, $\alpha_2 = 0.1$ and $\alpha_3 = 0.05$. We make several interesting observations from the figure:

1. The queue lengths look like piecewise-linear functions (this is partially due to the law of large numbers over the arrival process).
2. The queue dynamics are somewhat complex for the first time slots (largely due to the internal arrival from other links).
3. The first queue eventually drops to close to zero, and the behavior of the other queues become more predictable.
4. Finally all queues seem to be close to zero, so the system is expected to be stable.

In light of the above findings, we first conjecture that after some time we have either $Z_1(t) \approx Z_2(t) \geq Z_3(t)$ or $Z_1(t) \leq Z_2(t) \approx Z_3(t)$, since otherwise one queue will be larger than all its neighbors, resulting in a higher scheduling priority under LQF that will force the queue to start decreasing. We can then see that if $Z_1(t)$ and $Z_2(t)$ stick together then they must both decrease since the sum of the nominal total arrival rates to links l_1 and l_2 due to both the exogenous and internal arrivals is $2\alpha_1 + \alpha_2 = 0.6 < 1$. If $Z_2(t)$ and $Z_3(t)$ stick together and $Z_1(t)$ is positive, then the service rate of link l_i for all i, denoted by μ_i, should satisfy

$$\mu_1 = \mu_3,$$

$$\mu_1 + \mu_2 = 1,$$

and

$$\mu_1 + \alpha_2 - \mu_2 = \mu_2 + \alpha_3 - \mu_3.$$

We can then get $\mu_1 - \alpha_1 = \frac{1}{2} - \alpha_1 - \frac{1}{4}\alpha_2 + \frac{1}{4}\alpha_3 = 0.2375 > 0$. So in either case the first queue will decrease. We also argue that when the first queue drops to close to zero, it cannot rise again since if it did it would be "forced back" to zero immediately since the potential service rate of link l_1 is larger than its nominal (exogenous only) arrival rate. So at last the three-link linear network is reduced to a two-link linear network and the remaining two queues go to close to zero as well. The above intuition will lead our way to the rigorous proof for the general linear network case in the rest of this section.

2.3.3 Notation and Network Equations

The following notation is used throughout the chapter:

- R: the N-by-N routing matrix that is similar to the one defined by Tassiulas and Ephremides (1992), where $R_{ik} = -1$ if link l_k goes from node v_i, $R_{ik} = 1$ if link l_k goes to node v_i, or $R_{ik} = 0$ otherwise, for $1 \leq i \leq N$ and $1 \leq k \leq N$. Then in the linear network case the routing matrix is given by

$$R = \begin{pmatrix} -1 & 0 & 0 & \cdots & 0 \\ 1 & -1 & \ddots & \ddots & \vdots \\ 0 & 1 & \ddots & \ddots & \vdots \\ \vdots & \ddots & \ddots & -1 & 0 \\ 0 & \cdots & 0 & 1 & -1 \end{pmatrix}. \tag{2.1}$$

- M: the N-by-r binary-entry matrix whose columns are the activation vectors of the possible maximal schedules, where r is the total number of possible maximal schedules. We then have $M = (m_1, m_2, \ldots, m_r)$ where m_j is the activation vector of a maximal schedule for $1 \leq j \leq r$.
- $Z_i(n)$ for $1 \leq i \leq N$: the queue length at link l_i at time slot $n \in \mathbb{N}$ (before arrivals and departures happen in time slot n).
- $E_i(n)$ for $1 \leq i \leq N$: the cumulative exogenous arrival to link l_i up to time slot $n \in \mathbb{N}$. We assume the increment process of $\big(E_i(n)\big)$, i.e., the process $\big(E_i(n+1) - E_i(n) \colon n \in \mathbb{N}\big)$, is i.i.d. across n. We also assume the processes $\big(E_1(n)\big)$, $\big(E_2(n)\big)$, $\ldots$, $\big(E_N(n)\big)$ are independent. The exogenous arrival rate is $\mathbb{E}[E_i(n+1) - E_i(n)] = \alpha_i$ for all n.
- $A_i(n)$ for $1 \leq i \leq N$: the cumulative arrival to link l_i up to time slot $n \in \mathbb{N}$. This includes both exogenous and internal arrivals.
- $D_i(n)$ for $1 \leq i \leq N$: the actual cumulative departure from link l_i up to time slot $n \in \mathbb{N}$.

- $T_j(n)$ for $1 \leq j \leq r$: the cumulative service time (in number of time slots) of schedule m_j up to time slot $n \in \mathbb{N}$.
- $Y_i(n)$ for $1 \leq i \leq N$: the cumulative idle time (in number of time slots) of link l_i up to time slot $n \in \mathbb{N}$ (when link l_i is chosen by the scheduler but does not actually send packets). Note that even if the queue at link l_i is empty at the time of scheduling, the scheduler can still choose the schedule m_k such that $l_i \in m_k$, in which case $Y_i(n)$ will increase instead of $D_i(n)$. For non-idling (or work-conserving) scheduling policies $Y_i(n)$ can only increase when the queue at link l_i is empty.

Let $Z(n), E(n), A(n), D(n), T(n), Y(n)$ be the corresponding column vectors and let $\mathbb{X}(n) = (Z(n), E(n), A(n), D(n), T(n), Y(n))$ for any $n \in \mathbb{N}$. Then we refer to $\big(\mathbb{X}(n)\big)$ as the *queueing network process*. Let $\mathcal{X} = \mathbb{N}^{5N+r}$ be the state space of $\big(\mathbb{X}(n)\big)$. Then $\big(\mathbb{X}(n)\big)$ is an $\mathcal{X}$-valued stochastic process defined on $\mathbb{N}$. Let Ω be the set of sample paths specifying the exogenous arrival processes $\big(E(n)\big)$ and the possible tie-breaks of the scheduler. Note that under the LQF policy $\big(\mathbb{X}(n)\big)$ is a discrete Markov chain. The dynamics of the network are described by the following *queueing network equations*:

$$A(n) = E(n) + (R + I)D(n-1) \tag{2.2}$$

$$Z(n) = Z(0) + A(n) - D(n) \tag{2.3}$$

$$\sum_{j=1}^{r} T_j(n) = n \quad \Big(\text{or } e^{\mathrm{T}} T(n) = n\Big) \tag{2.4}$$

$$D(n) = MT(n) - Y(n) \tag{2.5}$$

for any nonnegative integer n, where I is the identity matrix. Moreover, if the

scheduling is non-idling, then we also have

$$Y_i(n) - Y_i(n-1) = \begin{cases} 1 & \text{if } Z_i(n-1) = 0 \text{ and} \\ & \sum_{j:\, i \in m_j} \big(T_j(n) - T_j(n-1)\big) = 1 \\ 0 & \text{otherwise} \end{cases} \tag{2.6}$$

for $1 \le i \le N$ and $n = 1, 2, 3, \ldots$. All of the variables take nonnegative integers in each component, and $\big(E_i(n)\big)$, $\big(A_i(n)\big)$, $\big(D_i(n)\big)$, $\big(T_j(n)\big)$, $\big(Y_i(n)\big)$ are all nondecreasing in n for any i and j. Also we assume the initial conditions are

$$E(0) = A(0) = D(0) = Y(0) = 0 \text{ and } T(0) = 0. \tag{2.7}$$

For the LQF policy, we have in addition to (2.2), (2.3), (2.4), (2.5), (2.6) and (2.7):

$$T_j(n) - T_j(n-1) = 1 \Rightarrow m_j \in \mathrm{LQF}(Z(n-1)), \tag{2.8}$$

where $\mathrm{LQF}(Z)$ is the set of possible LQF maximal schedules given queue length vector Z. We assume that the schedule is always maximal regardless of the queues being empty or not, so $\mathrm{LQF}(Z) \subseteq \{m_1, m_2, \ldots, m_r\}$.

2.3.4 Fluid Limits

We define the scaled systems based on the queueing network process for each sample path, and show that the scaled systems converge along some subsequence to deterministic systems called fluid limits.

We first extend the definition of $\mathbb{X}$ for arbitrary nonnegative time $t \in \mathbb{R}_+$ by piecewise linear interpolation

$$\mathbb{X}(t) = (1 + \lfloor t \rfloor - t)\mathbb{X}(\lfloor t \rfloor) + (t - \lfloor t \rfloor)\mathbb{X}(\lfloor t \rfloor + 1),$$

where $\lfloor t \rfloor$ is the largest integer less than or equal to t. Then $\mathbb{X}$ is an $\bar{\mathcal{X}}$-valued stochastic process with $\bar{\mathcal{X}} = \mathbb{R}_+^{5N+r}$, and is continuous for $t \in \mathbb{R}_+$ given any fixed sample path $\omega \in \Omega$.

Let $|\cdot|$ be the L^1-norm of $\mathbb{R}_+^{5N+r}$. Fix $\omega \in \Omega$, and let $\mathbb{X}^x(t)$ be the queueing network process with initial state $\mathbb{X}(0) = x$ for $x \in \mathcal{X}^0 = \left\{y \in \mathcal{X} \colon |y| > 0\right\}$ and define the *scaled system*

$$\bar{\mathbb{X}}^x(t) = \frac{1}{|x|}\mathbb{X}^x(|x|t).$$

We then have the following proposition giving the existence of the fluid limits, which is similar to Theorem 4.1 in Dai (1995).

Proposition 1. *For a work-conserving scheduling policy, for almost any sample path $\omega \in \Omega$ and any sequence of initial states $(x_k \colon k \in \mathbb{N})$ with $x_k \in \mathcal{X}^0$ for all k and $|x_k| \to \infty$ as $k \to \infty$, there exists a subsequence $\left(x_{k_p} \colon p \in \mathbb{N}\right)$ with $|x_{k_p}| \to \infty$ as $p \to \infty$ such that*

$$\bar{\mathbb{X}}^{x_{k_p}}(0) \to \bar{\mathbb{X}}(0) \quad \text{as } p \to \infty$$

and

$$\bar{\mathbb{X}}^{x_{k_p}} \to \bar{\mathbb{X}} \quad \text{u.o.c. as } p \to \infty$$

for some $\bar{\mathbb{X}} \colon \mathbb{R}_+ \to \bar{\mathcal{X}}$ with $\bar{\mathbb{X}}(t) = (\bar{Z}(t), \bar{E}(t), \bar{A}(t), \bar{D}(t), \bar{T}(t), \bar{Y}(t))$ for any $t \in \mathbb{R}_+$, where "u.o.c." stands for uniform convergence over compact sets (Royden and Fitzpatrick, 2010). Furthermore, for any $t \in \mathbb{R}_+$,

$$\bar{A}(0) = \bar{D}(0) = \bar{Y}(0) = 0 \text{ and } \bar{T}(0) = 0 \tag{2.9}$$

$$\bar{A}(t) = \alpha t + (R + I)\bar{D}(t) \tag{2.10}$$

$$\bar{Z}(t) = \bar{Z}(0) + \bar{A}(t) - \bar{D}(t) \tag{2.11}$$

$$e^{\mathrm{T}}\bar{T}(t) = t \tag{2.12}$$

$$\bar{D}(t) = M\bar{T}(t) - \bar{Y}(t) \tag{2.13}$$

and

$$\int_0^\infty \bar{Z}_i(t)\,\mathrm{d}\bar{Y}_i = 0 \quad i = 1, 2, \ldots, N. \tag{2.14}$$

Moreover, all components of $\bar{\mathbb{X}}$ *are absolutely continuous (Royden and Fitzpatrick, 2010) because they are Lipschitz continuous, and* $\big(\bar{A}_i(t)\big)$, $\big(\bar{D}_i(t)\big)$, $\big(\bar{T}_j(t)\big)$ *and* $\big(\bar{Y}_i(t)\big)$ *are nondecreasing in* t *for all* i *and* j.

Particularly, the fluid limits under LQF satisfy

$$\frac{\mathrm{d}\bar{T}_j}{\mathrm{d}t}(t) > 0 \Rightarrow m_j \in \mathrm{LQF}(\bar{Z}(t)) \quad j = 1, 2, \ldots, r, \tag{2.15}$$

where $\mathrm{LQF}(\bar{Z})$ *is the set of LQF schedules for the vector* $\bar{Z}$*, and* t *is assumed regular so the derivatives exist.* ◇

Remark 1. Basically, (2.9) is the initial condition assumption. (2.10) says the arrival rates consist of exogenous part and internal part. (2.11) is the queue evolution equation. (2.12) comes from the fact that in any period of time $(t, t+\delta)$ of the scaled systems the total increase of the cumulative service time is at most δ. (2.13) gives the relation among departures, serving time of schedules and idling time. (2.14) means a link can be idle (when it is chosen by the scheduler) only if the queue length at the link is 0. (2.15) states that only maximal schedules satisfying the LQF property given the queue fluids $\bar{Z}(t)$ can be chosen at time t, but it does not specify the fractions of the schedules that LQF could choose. The proof is similar to those in Dai and Prabhakar (2000), Shah and Wischik (2012) and Chen and Yao (2001), and can be found in Appendix A.1.

2.3.5 Transient States with Dominating Fluids

We first identify a set of transient states of the space of the queue fluid vectors. Let

$$B_1 = \left\{ \bar{Z} \in \mathbb{R}_+^N \colon \bar{Z}_1 > \bar{Z}_2 \right\}$$

$$B_2 = \left\{ \bar{Z} \in \mathbb{R}_+^N \colon \bar{Z}_2 > \bar{Z}_1, \bar{Z}_2 > \bar{Z}_3 \right\}$$

$$\vdots$$

$$B_{N-1} = \left\{ \bar{Z} \in \mathbb{R}_+^N \colon \bar{Z}_{N-1} > \bar{Z}_{N-2}, \bar{Z}_{N-1} > \bar{Z}_N \right\}$$

$$B_N = \left\{ \bar{Z} \in \mathbb{R}_+^N \colon \bar{Z}_N > \bar{Z}_{N-1} \right\}$$

and let

$$B = \bigcup_{i=1}^{N} B_i.$$

Then we have

$$\mathbb{R}_+^N \backslash B = \{ \bar{Z} \in \mathbb{R}_+^N : \bar{Z}_1 \leq \bar{Z}_2, \bar{Z}_{N-1} \geq \bar{Z}_N, \\ \bar{Z}_i \leq \max\{\bar{Z}_{i-1}, \bar{Z}_{i+1}\} \text{ for } 2 \leq i \leq N-1 \}.$$

So B is the set of queue fluid vectors such that some queue strictly dominates all of its neighbors (one link has at most two neighbors in linear networks), while $\mathbb{R}_+^N \backslash B$ is the set of queue length vectors without any queue strictly dominating all of its neighbors. We then have the following lemma.

Lemma 1. *B is transient. Formally, given $\alpha_i < 1$ for any $i \in \{1, 2, \ldots, N\}$, for any initial conditions $\bar{Z}(t_0) \in \mathbb{R}_+^N$ at time t_0, we have $\bar{Z}(t) \notin B$ for any $t \geq t_0 + \frac{\max_i \bar{Z}_i(t_0)}{(1-\max_i \alpha_i)}$ under LQF.* ◇

Remark 2. The outline of the proof is as follows, and the proof can be found in Appendix A.2.

1. If $\bar{Z} \in B$, then there are no adjacent dominating nodes.
2. Each dominating node loses its domination in time $\frac{\max_i \bar{Z}_i(t_0)}{(1-\max_i \alpha_i)}$.
3. Once a node loses domination, it cannot regain it.

2.3.6 Stability of the First Fluid $\bar{Z}_1$

We now further divide $\mathbb{R}_+^N \backslash B$ into several partitions:

$$C_0 = \left\{ \bar{Z} \in \mathbb{R}_+^N \backslash B \colon \bar{Z}_1 = 0 \right\}$$

$$C_1 = \left\{ \bar{Z} \in \mathbb{R}_+^N \backslash B \colon 0 < \bar{Z}_1 = \bar{Z}_2 \right\}$$

$$C_2 = \left\{ \bar{Z} \in \mathbb{R}_+^N \backslash B \colon 0 < \bar{Z}_1 < \bar{Z}_2 = \bar{Z}_3 \right\}$$

$$\vdots$$

$$C_{N-1} = \left\{ \bar{Z} \in \mathbb{R}_+^N \backslash B \colon 0 < \bar{Z}_1 < \cdots < \bar{Z}_{N-1} = \bar{Z}_N \right\}.$$

Then $\{C_0, C_1, \ldots, C_{N-1}, B\}$ forms a partition of $\mathbb{R}_+^N$. We then use the following two lemmas to show C_1, C_2, …, C_{N-1} are all transient under admissible arrival rates, so the system has to eventually go to state C_0 where $\bar{Z}_1$ stays at 0.

Lemma 2. *If the arrival rate vector α is admissible, then there exists $\epsilon > 0$ such that for any regular time $t_1 \geq \frac{\max_i \bar{Z}_i(0)}{1-\max_i \alpha_i}$ and $\bar{Z}(t_1) \notin C_0$ we have*

$$\frac{\mathrm{d}\bar{Z}_1}{\mathrm{d}t}(t_1) \leq -\epsilon.$$

◇

Remark 3. The idea of the proof is that for any sufficiently large regular time t_1 we show that if the fluid of the first queue is positive, then it must decrease with lower-bounded rate. Hence the first fluid reaches zero eventually.

Proof. Let $t' = \frac{\max_i \bar{Z}_i(0)}{1-\max_i \alpha_i}$. Then by Lemma 1, we have $\bar{Z}(t) \notin B$ for any $t \geq t'$. We let

$$\bar{W}_1(t) = \bar{Z}_1(t)$$

and

$$\bar{W}_i(t) = \bar{Z}_i(t) - \bar{Z}_{i-1}(t) \quad i = 2, 3, \ldots, N.$$

We further let

$$J_0(t) = \left\{ j \colon \bar{W}_j(t) = 0 \right\}$$

and for a regular time t,

$$J(t) = \left\{ j \in J_0(t) \colon \frac{\mathrm{d}\bar{W}_j}{\mathrm{d}t}(t) = 0 \right\}.$$

Note that $\bar{Z}(t) \in \mathbb{R}_+^N \backslash B$ implies $J_0(t) \neq \emptyset$. We claim that $J(t)$ is also nonempty in the following proposition, the proof of which can be found in Appendix A.3.

Proposition 2. *For any $t \geq t'$, we have $J(t) \neq \emptyset$.* ◇

Now we fix a regular time $t_1 \geq t'$ with $\bar{Z}_1(t_1) > 0$ and let

$$u = \min_{j \in J(t_1)} j.$$

Then $u \geq 2$. Let the service rate on link l_i at time t be $\mu_i(t) = \frac{\mathrm{d}}{\mathrm{d}t}\bar{D}_i(t)$ for regular time t and any $i \in \{1, 2, \ldots, N\}$. Then we claim that the service rates up to u at time t_1 satisfy the following proposition, the proof of which is presented in Appendix A.4.

Proposition 3. *For $i = 1, 2, \ldots, u-2$, $\mu_i(t_1) = \mu_{i+2}(t_1)$.* ◇

Due to the one-hop interference model, we have

$$\mu_1(t_1) + \mu_2(t_1) = 1$$

since at each time slot in the real system either link l_1 or link l_2 must be scheduled. Then by the definition of u, we have $\frac{\mathrm{d}\bar{Z}_{u-1}}{\mathrm{d}t}(t_1) = \frac{\mathrm{d}\bar{Z}_u}{\mathrm{d}t}(t_1)$, i.e.,

$$\mu_{u-2}(t_1) + \alpha_{u-1} - \mu_{u-1}(t_1) = \mu_{u-1}(t_1) + \alpha_u - \mu_u(t_1)$$

where $\mu_0(t_1) = 0$ by convention. Then if $u = 2$, we have

$$\begin{aligned}
&\begin{cases} \mu_1(t_1) + \mu_2(t_1) = 1 \\ \alpha_1 - \mu_1(t_1) = \mu_1(t_1) + \alpha_2 - \mu_2(t_1) \end{cases} \\
\Rightarrow &\begin{cases} \mu_1(t_1) = \frac{1}{3} + \frac{1}{3}\alpha_1 - \frac{1}{3}\alpha_2 \\ \mu_2(t_1) = \frac{1}{3} - \frac{1}{3}\alpha_1 + \frac{1}{3}\alpha_2 \end{cases} \\
\Rightarrow &\frac{\mathrm{d}\bar{Z}_1}{\mathrm{d}t}(t_1) = \alpha_1 - \mu_1(t_1) = -\frac{1}{3} + \frac{2}{3}\alpha_1 + \frac{1}{3}\alpha_2.
\end{aligned}$$

Similarly, if $u = 3$,

$$\begin{aligned}
&\begin{cases} \mu_1(t_1) + \mu_2(t_1) = 1 \\ \mu_1(t_1) = \mu_3(t_1) \\ \mu_1(t_1) + \alpha_2 - \mu_2(t_1) = \mu_2(t_1) + \alpha_3 - \mu_3(t_1) \end{cases} \\
\Rightarrow &\begin{cases} \mu_1(t_1) = \mu_3(t_1) = \frac{1}{2} - \frac{1}{4}\alpha_2 + \frac{1}{4}\alpha_3 \\ \mu_2(t_1) = \frac{1}{2} + \frac{1}{4}\alpha_2 - \frac{1}{4}\alpha_3 \end{cases} \\
\Rightarrow &\frac{\mathrm{d}\bar{Z}_1}{\mathrm{d}t}(t_1) = \alpha_1 - \mu_1(t_1) = -\frac{1}{2} + \alpha_1 + \frac{1}{4}\alpha_2 - \frac{1}{4}\alpha_3,
\end{aligned}$$

and if $u = 4$ we have,

$$\begin{cases} \mu_1(t_1) + \mu_2(t_1) = 1 \\ \mu_1(t_1) = \mu_3(t_1) \\ \mu_2(t_1) = \mu_4(t_1) \\ \mu_2(t_1) + \alpha_3 - \mu_3(t_1) = \mu_3(t_1) + \alpha_4 - \mu_4(t_1) \end{cases}$$

$$\Rightarrow \begin{cases} \mu_1(t_1) = \mu_3(t_1) = \frac{1}{2} + \frac{1}{4}\alpha_3 - \frac{1}{4}\alpha_4 \\ \mu_2(t_1) = \mu_4(t_1) = \frac{1}{2} - \frac{1}{4}\alpha_3 + \frac{1}{4}\alpha_4 \end{cases}$$

$$\Rightarrow \frac{\mathrm{d}\bar{Z}_1}{\mathrm{d}t}(t_1) = \alpha_1 - \mu_1(t_1) = -\frac{1}{2} + \alpha_1 - \frac{1}{4}\alpha_3 + \frac{1}{4}\alpha_4.$$

We can then get the derivative of $\bar{Z}_1(\cdot)$ at t_1 as

$$\frac{\mathrm{d}\bar{Z}_1}{\mathrm{d}t}(t_1) = \begin{cases} -\frac{1}{3} + \frac{2}{3}\alpha_1 + \frac{1}{3}\alpha_2 & \text{if } u = 2 \\ -\frac{1}{2} + \alpha_1 + \frac{1}{4}\alpha_{u-1} - \frac{1}{4}\alpha_u & \text{if } u = 3, 5, \ldots \cdot \\ -\frac{1}{2} + \alpha_1 - \frac{1}{4}\alpha_{u-1} + \frac{1}{4}\alpha_u & \text{if } u = 4, 6, \ldots \end{cases}$$

Since α is admissible, we have (Tassiulas and Ephremides, 1992)

$$-R^{-1}\alpha < M\gamma \tag{2.16}$$

for some convex combination coefficients γ. Then by (2.1), we have

$$-R^{-1} = \begin{pmatrix} 1 & 0 & 0 & \cdots & 0 \\ 1 & 1 & 0 & \ddots & 0 \\ \vdots & \ddots & \ddots & \ddots & \vdots \\ 1 & 1 & \ddots & \ddots & 0 \\ 1 & 1 & 1 & \cdots & 1 \end{pmatrix}.$$

Note that the ith row of $-R^{-1}\alpha$ is the *total workload* of link l_i, while the ith row of $M\gamma$ is the service rate on link l_i. Since the total workload of the last two links

are $\sum_{i=1}^{N-1}\alpha_i$ and $\sum_{i=1}^{N}\alpha_i$ respectively, and exactly one of the last two links must be chosen at each time slot, we have by combining the last two rows of (2.16)

$$2\alpha_1 + 2\alpha_2 + \cdots + 2\alpha_{N-1} + \alpha_N < 1.$$

So $\frac{\mathrm{d}\bar{Z}_1}{\mathrm{d}t}(t_1) \leq -\epsilon$, where

$$\begin{aligned}\epsilon = \min\Bigg\{ & \frac{1}{3} - \frac{2}{3}\alpha_1 - \frac{1}{3}\alpha_2, \\ & \frac{1}{2} - \alpha_1 - \frac{1}{4}\alpha_2 + \frac{1}{4}\alpha_3, \\ & \frac{1}{2} - \alpha_1 + \frac{1}{4}\alpha_3 - \frac{1}{4}\alpha_4, \\ & \ldots, \\ & \frac{1}{2} - \alpha_1 - \frac{1}{4}(-1)^{N-1}\alpha_{N-1} - \frac{1}{4}(-1)^{N}\alpha_N \Bigg\} \\ > 0.\end{aligned}$$

□

Corollary 1. *Given the initial conditions and arrival rates in Lemma 2, there exists* $\epsilon > 0$ *such that* $\bar{Z}_1(t) = 0$ *for any* $t \geq \frac{\bar{Z}_1(0)}{\epsilon} + \frac{\max_i \bar{Z}_i(0)}{1-\max_i \alpha_i}\left(\frac{1}{\epsilon} + 1\right)$. ◇

Remark 4. This comes directly from Lemma 2 and a similar proof of Proposition 5 in Appendix A.2. Basically, $\bar{Z}_1(t)$ has to drop to 0, after which it cannot rise since otherwise the negative derivative forces it to go back to 0.

2.3.7 Coupled-Network Argument

Based on Corollary 1, we use induction and a coupled-network argument to show the following lemma stating the stability of the fluid system, which leads to the stability of the original queueing system using a similar argument as presented by Dai (1995).

Lemma 3. *Given the initial conditions and arrival rates in Lemma 2, there exists* $c_3 > 0$ *such that* $\bar{Z}_i(t) = 0$ *for any* $t \geq \max_j \bar{Z}_j(0)c_3$ *and any* $i = 1, 2, \ldots, N$. $\diamond$

Proof. We use induction. First, by Corollary 1, there exists $\tilde{c} > 0$ such that $\bar{Z}_1(t) = 0$ for any $t \geq \max_j \bar{Z}_j(0)\tilde{c}$. Now suppose there exists c and k such that $\bar{Z}_i(t) = 0$ and $\bar{Z}_{k+1}(t) > 0$ for any $t \geq \max_j \bar{Z}_j(0)c$ and $i \leq k$. We consider a coupled linear network under the LQF scheduling with $N - k$ links, initial fluids $\bar{Z}'_i(\max_j \bar{Z}_j(0)c) = \bar{Z}_{i+k}(\max_j \bar{Z}_j(0)c)$ for $1 \leq i \leq N - k$, and arrival rates

$$\alpha'_1 = \alpha_1 + \alpha_2 + \cdots + \alpha_{k+1}$$

and

$$\alpha'_j = \alpha_{k+j} \qquad j = 2, 3, \ldots, N - k.$$

Thus $\{\bar{Z}'_i(t), 1 \leq i \leq N\}$ are the fluids of the original network with the first $k+1$ links combined into one link. Since the fluids satisfy $\bar{Z}_{k+1}(t) > \bar{Z}_k(t)$, we have that the queue length at link l_{k+1} is larger than that at link l_k in the actual system. Then by the LQF scheduling, the schedule of the first k links does not affect the schedule of the last $N-k$ links. Also notice that the fluid arrival to $\bar{Z}_{k+1}(t)$ is $\alpha_1+\alpha_2+\cdots+\alpha_{k+1} = \alpha'_1$ since all fluids $\bar{Z}_i(t)$'s prior to $\bar{Z}_{k+1}(t)$ remain zero, transferring their exogenous arrival to $\bar{Z}_{k+1}(t)$. Hence, $\bar{Z}_{i+k}(t) = \bar{Z}'_i(t)$ for all $t \geq \max_j \bar{Z}_j(0)c$.

Taking the last $N - k$ rows of (2.16), we can get

$$-R'^{-1}\alpha' < M'\gamma',$$

where R' is the routing matrix for the coupled network, M' is the maximal scheduling matrix of the coupled network, and γ' is a set of convex combination coefficients induced from γ. Note that M' consists of the maximal columns of the matrix formed by the last $N-k$ rows of M. Hence α' is also admissible. Let $\bar{Z}'_{\max} = \max_i \bar{Z}'_i(0)$. Then

by the Lipschitz continuity we have $\bar{Z}'_{\max} \leq \max_j \bar{Z}_j(0)c_4$ for some $c_4 > 0$. Again by Corollary 1, there exists $c_5 > 0$ such that $\bar{Z}'_1(t) = 0$ for any $t \geq \bar{Z}'_{\max}c_5$. Consequently, $\bar{Z}_{k+1}(t) = \bar{Z}'_1(t) = 0$ is also true for $t \geq \max_j \bar{Z}_j(0)c_4c_5$. By mathematical induction, we get that there exists $c_3 > 0$ such that $\bar{Z}_i(t) = 0$ for any $t \geq \max_j \bar{Z}_j(0)c_3$ and $1 \leq i \leq N$. □

2.3.8 Uniform Integrability

According to Lemma 4.5 in Dai (1995), we now have for sufficiently large t,

$$\frac{1}{n}\max_i Z_i(nt) \to 0$$

as $n \to \infty$. Following a similar argument given in Dimakis and Walrand (2006), to get the stability of the system, we only need to show

$$\mathbb{E}\left(\frac{1}{n}\max_i Z_i(nt)\right) \to 0$$

as $n \to \infty$. I.e., we only need to show $\left(\frac{1}{n}\max_i Z_i(nt) \colon n \in \mathbb{N}\right)$ are uniformly integrable (UI). Note

$$\begin{aligned}\frac{1}{n}\max_i Z_i(nt) &\leq \frac{1}{n}\sum_i Z_i(nt) \\ &\leq \frac{1}{n}\sum_i E_i(nt).\end{aligned}$$

Then by the law of large numbers, $\frac{1}{n}\sum_i E_i(nt)$ converges to $\sum_i \alpha_i t$ in probability as $n \to \infty$. Also note

$$\mathbb{E}\left(\frac{1}{n}\sum_i E_i(nt)\right) = \sum_i \alpha_i t.$$

Then by Theorem 4.5.4 in Chung (2001) we have UI of $\left(\frac{1}{n}\sum_i E_i(nt)\right)$ and thus that of $\left(\frac{1}{n}\max_i Z_i(nt)\right)$.

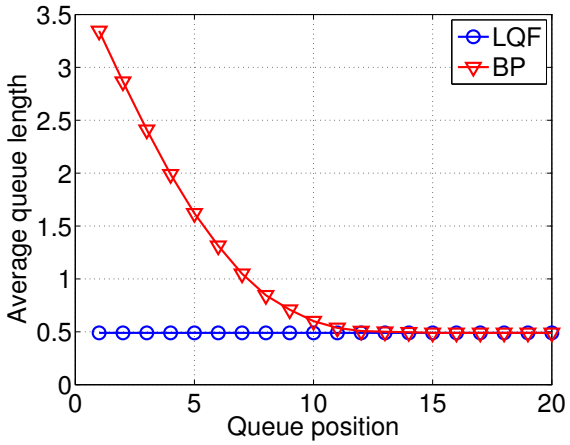
2.3.9 Simulations

Now we proved that LQF is throughput optimal in a linear network with one-hop interference model. In this subsection, we simulate the LQF policy and the back-pressure policy over a linear network with multihop traffic under different interference models. The goal of the simulations is twofold: 1) to examine the throughput performance of LQF on linear networks under other interference model; 2) to evaluate the delay performance of both LQF and BP.

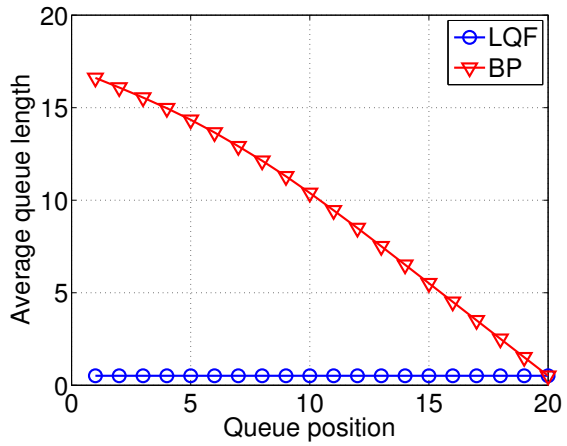
Throughout the simulations we fix the network size to be $N = 20$. We assume that there is a single flow with source node v_1, destination node v_{N+1}, and Bernoulli arrivals with mean α_1. We consider three interference models: no interference (all links can be transmitting simultaneously), one-hop interference (each link interferes with its direct neighbors) and two-hop interference (each link interferes with its two-hop neighbors). We simulate $1,000,000$ time slots for each setting.

2.3.9.1 No Interference

When no interference is present, we can easily see that the stability region of the network is $[0, 1)$. LQF is throughput optimal in this case since all links will always try to transmit and the queue length is at most 1 for any link due to the Bernoulli arrival and the zero initial state. It has however been noticed that there exists a critical point of the arrival rate for linear multihop networks, above which the average total queue lengths (and hence the average delay) will increase quadratically as the network size becomes larger (Stolyar, 2011; Bui *et al.*, 2011). In Figure 4 we demonstrate that the



(a) Undercritical scenario when $\alpha_1 = 0.49$



(b) Supercritical scenario when $\alpha_1 = 0.51$

Figure 4: Stationary queue lengths for different link positions under no interference

critical arrival rate is 1/2 in our discrete-time constant service setting, as opposed to 1/4 obtained in the continuous-time exponential-service setting by Stolyar (2011).

Figure 4a shows the stationary queue lengths of both policies in the undercritical scenario when $\alpha_1 = 0.49 < 1/2$. We note that the stationary queue length of BP decreases quickly as the position increases, and stays at α_1 for the tail positions. Hence the average total backlog only increases linearly with the network size. By Little's law the delay is also linearly dependent on the network size.

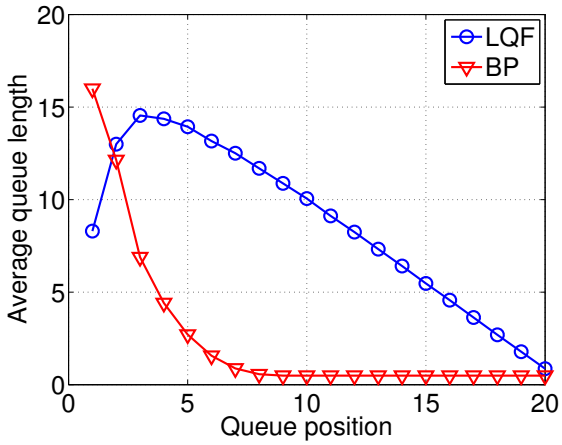


Figure 5: Stationary queue lengths for different link positions under one-hop interference when $\alpha_1 = 0.49$

On the other hand, Figure 4b shows the stationary queue lengths in the supercritical scenario when $\alpha_1 = 0.51 > 1/2$. We see that the stationary queue length of BP decreases linearly as the position increases. Then as the network size increases the average total backlog will increase quadratically, resulting in bad delay performance.

2.3.9.2 One-Hop Interference

Under one-hop interference the stability region becomes $[0, 1/2)$. The stationary queue lengths for both policies when $\alpha_1 = 0.49$ are shown in Figure 5. We notice that in this scenario the average total queue length of LQF is about three times that of BP, yielding comparable delay performances. We conjecture that the delay performance of BP under one-hop interference is good since the restriction of interference forces BP to choose good schedules.

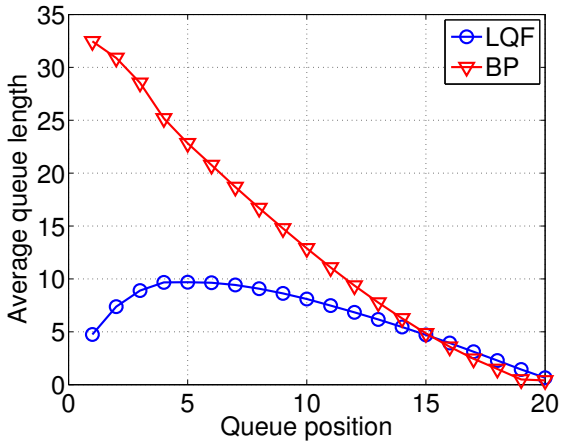


Figure 6: Stationary queue lengths for different link positions under two-hop interference when $\alpha_1 = 0.32$

2.3.9.3 Two-Hop Interference

Under two-hop interference the stability region further shrinks to $[0, 1/3)$. We show the stationary queue lengths for both policies when $\alpha_1 = 0.32$ in Figure 6. Note that in this particular case the average total queue length of LQF is less than half of that of BP, so LQF achieves better delay performance than BP. Whether LQF is throughput optimal for general linear networks under two-hop interference is unknown.

Chapter 3

LINK SCHEDULING IN WIRELESS NETWORKS WITH REAL-TIME TRAFFIC

3.1 Background

With the increasing number of real-time applications in wireless networks, scheduling traffic of packets with hard deadlines has become a very important problem. However, the problem is very challenging due to the stochastic nature of the traffic arrivals and deadlines. Hou *et al.* (2009) first proposed a frame-based analytical framework for studying scheduling real-time traffic in wireless networks. In the frame-based framework it is assumed that each frame is a number of consecutive time slots, and all packets arrive at the beginning of a frame and have to be scheduled before the end of the frame. They also characterized the real-time capacity region and developed the optimal scheduling algorithm for collocated networks. Later, the frame-based framework has been generalized to networks with heterogeneous delays, fading, congestion control, etc. (Hou and Kumar, 2009, 2010a,b; Jaramillo and Srikant, 2011; Jaramillo *et al.*, 2011) In particular, Jaramillo *et al.* (2011) extended the idea to general arrival/deadline patterns within a frame and general non-collocated network topology, and found the optimal scheduling policy, where they assumed that packets can arrive at any time slot during a frame, and the deadline of a packet can be any time after its arrival and before the end of the frame. Their paper assumes that the arrival and deadline information is available at the beginning of the frame, so future knowledge is assumed. Furthermore, the computational complexity of the

optimal algorithm is prohibitively high except for some special cases such as collocated networks.

In this chapter, we consider the case of general real-time traffic patterns without the assumption of frames and with a general conflict graph–based interference model. Under these settings, the stability region is difficult to characterize, and the optimal policy is unknown. In this chapter, we are interested in the performance of a low-complexity greedy policy called the largest-deficit-first (LDF) policy (Hou *et al.*, 2009), which is the real-time variation of the longest-queue-first (LQF) policy that iteratively selects the link with the largest deficit that does not interfere with those links that are already selected. It has been shown that the largest-deficit-first policy is optimal for scheduling real-time traffic in collocated networks (Hou *et al.*, 2009; Jaramillo *et al.*, 2011) under the frame-based model. The performance of the LDF in general non-collocated networks has not been studied.

Since LDF can be directly applied to networks with non-frame-based real-time traffic, we are interested in characterizing the performance of LDF. We investigate the efficiency ratio of LDF, which is the fraction of the throughput region guaranteed by LDF for given traffic distributions. Although the capacity region and optimal scheduling algorithm for networks with non-frame-based real-time traffic remain unknown, we are able to establish the efficiency ratio of LDF by connecting it to the frame-based optimal scheduling algorithm, and obtain a lower bound on the efficiency ratio in terms of a new quantity, called the real-time local-pooling factor (R-LPF). The R-LPF extends the idea of the local-pooling factor for non-real-time traffic (Joo *et al.*, 2007) and its extension for fading channels (Reddy *et al.*, 2012).

We show using the fluid limit technique (Dai, 1995) that this R-LPF can be successfully used to provide a minimum performance guarantee of LDF under real-

time traffic. While the R-LPF depends on the traffic pattern, we lower-bound the R-LPF by purely topological quantities based on the network, which in particular connects the R-LPF with the interference degree of the conflict graph[3] (Chaporkar *et al.*, 2005). Our contributions are therefore fourfold:

1. We formulate the construction of the R-LPF and prove that it is a lower bound on the efficiency ratio of LDF in the presence of non-frame-based real-time traffic.
2. We show that by assigning a nonnegative weight to each link we can get a lower bound on the R-LPF regardless of the traffic pattern. This translates to an R-LPF at least $1/(\beta+1)$ for a network with interference degree β, and in particular an R-LPF at least $1/3$ in a network with one-hop interference model.
3. We also propose a heuristic consensus algorithm that intelligently assigns the weights to compute a good lower bound based on the network topology.
4. We evaluate the performance of the LDF policy and the proposed consensus algorithm via simulations.

We would like to emphasize again that for general (non-frame-based) real-time traffic, to the best of our knowledge, there are no known theoretical results on any scheduling policy in ad hoc wireless networks. This makes the lower bounds obtained in this chapter a novel contribution.

[3]The interference degree of a network with a conflict graph is the maximum number of links that interfere with some single link and can be scheduled simultaneously.

3.2 Model

In this chapter, we consider a wireless network consisting of K links. The set of links is denoted by $\mathcal{K}$. Assume time is slotted, and at each time slot one packet can be successfully transmitted over a link if no interfering links are transmitting at the same time. We remark that the constant service rate assumption has been widely used in the literature, e.g., Dimakis and Walrand (2006); Joo *et al.* (2009b). We consider a general interference model. We call a set of links $\mathcal{Z} \subseteq \mathcal{K}$ a maximal link schedule if links in $\mathcal{Z}$ can be scheduled at the same time without interfering with each other, but no other link can be further scheduled without interfering with links in $\mathcal{Z}$. We assume that there are R possible maximal link schedules and the set of maximal link schedules is represented by a maximal link schedule matrix M, which is a K-by-R matrix with binary entries such that each column represents a distinct maximal link schedule and the set of links that are included in this schedule have value 1 in that column. For example, let M_r be the rth column of matrix M, then the set of links $\big\{l \in \mathcal{K} \colon M_{l,r} = 1\big\}$ is a maximal link schedule, where $M_{l,r}$ is the (l, r) entry of the matrix. By abuse of notation we also let $M = \{M_1, M_2, \ldots, M_R\}$. It is easy to see that any subset of a maximal link schedule is itself a feasible link schedule (i.e., all links in that set can be scheduled at the same time).

We consider single hop traffic with deadline constraints. Let $a_l(t)$ denote the number of packets that arrive at the beginning of time slot t at link l, where we assume that all packets have the same size and can be transmitted in a single time slot. We assume that $\big(a(t) \colon t \geq 1\big)$ is a stochastic process that is temporally independent and identically distributed (i.i.d.) and independent across links, with probability mass function (p.m.f.) $f_l \colon \mathbb{N} \to \mathbb{R}$, where $\mathbb{N}$ is the set of nonnegative integers and $\mathbb{R}$ is the

set of real numbers. We also assume that $f_l(i) = 0$ for $i > a_{\max}$; i.e., the number of packets arriving on a link at each time slot is at most $a_{\max}$. Denote by $\bar{a}_l$ the arrival rate on link l; i.e., $\bar{a}_l = \mathbb{E}[a_l(t)] = \sum_{i=1}^{N} i f_l(i)$ for any t.

Each packet is associated with a maximum delay τ, which is a random variable with integer value between $\tau_{\min}$ and $\tau_{\max}$ and follows a p.m.f. $g_l \colon \{\tau_{\min}, \tau_{\min}+1, \ldots, \tau_{\max}\} \to \mathbb{R}$. Furthermore, let $A_l(t)$ be the cumulative number of packet arrivals to link l up to time slot t for any $l \in \mathcal{K}$ and any nonnegative integer t; i.e., $A_l(t) = \sum_{t'=1}^{t} a_l(t')$, and by convention $A_l(0) = 0$. We order the packets arriving on link l according to the arriving time with arbitrary tie-breakings. Then we let $b_l(n)$ be the time slot during which the nth packet arrives on link l; i.e., $b_l(n) = \min\big\{t \colon A_l(t) \geq n\big\}$. We also let $e_l(n)$ be the deadline of the nth packet on link l. Note that $e_l(n) = b_l(n) + \tau_l(n) - 1$, where $\tau_l(n)$ is the maximum delay associated with the nth packet on link l. Then the nth arriving packet on link l will be immediately dropped if the deadline is missed. Note that $\big(A(t) \colon t \geq 0\big)$, $\big(\tau(n) \colon n \geq 1\big)$, $\big(b(n) \colon n \geq 1\big)$ and $\big(e(n) \colon n \geq 1\big)$ are all stochastic processes, and $\big(A(t) \colon t \geq 0\big)$ and $\big(\tau(n) \colon n \geq 1\big)$ determine $\big(b(n) \colon n \geq 1\big)$ and $\big(e(n) \colon n \geq 1\big)$. Denote the space of sample paths of the cumulative arrival process $\big(A(t) \colon t \geq 0\big)$ and the maximum delay process $\big(\tau(n) \colon n \geq 1\big)$ by $\mathcal{A}$. An example of a sample path of the arrival and maximum delay processes on a link during the first 10 time slots is shown in Figure 7.

We assume each link l in $\mathcal{K}$ is associated with a minimum delivery rate p_l (sometimes called the *quality of service* or *QoS*), which is the minimum fraction of packets that should be delivered on link l. The goal of a scheduling policy is to keep the long-term delivery rate on link l at least p_l.

Now consider a scheduling policy μ. Denote by $S^{\mu}(t)$ the cumulative service up

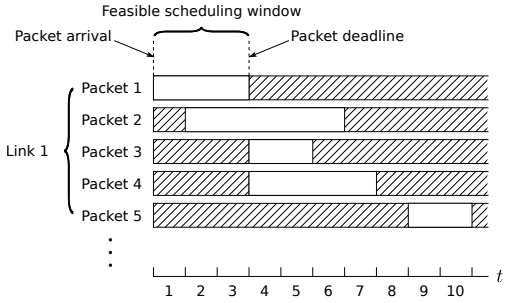


Figure 7: An example of the arrival and maximum delay pattern of packets on a link

Note: For each packet, the beginning of the blank bar is the time slot when that packet arrives, and the end of the blank bar is the deadline associated with that packet. So the feasible scheduling window denoted by the blank bar represents the time slots when the packet is available for transmission, while the shaded part indicates that the packet is not available, either because it has not arrived or because its deadline has passed. Note that here the cumulative numbers of packet arrivals to link 1 are $\big(A_1(t)\colon t \geq 0\big) = (0, 1, 2, 2, 4, 4, 4, 4, 4, 5, 5, \dots)$ and the maximum delays are $\big(\tau_1(n)\colon n \geq 1\big) = (3, 5, 2, 4, 2, \dots)$.

to time t, in which $S_l^\mu(t)$ is the service link l received up to time slot t. For any scheduling policy, it is easy to see that the following three conditions hold:

1. (Initialization) $S_l^\mu(0) = 0$ for all $l \in \mathcal{K}$.
2. (Feasibility) The incremental service vector is a feasible schedule; i.e., $0 \preceq S^\mu(t) - S^\mu(t-1) \preceq M_r$ for some $M_r \in M$, for any positive integer t, where $\preceq$ denotes entrywise less than or equal to.
3. (Deadline constraint) All served packets are served before their deadlines. Formally, let $\zeta_l^\mu(n)$ be the time slot in which the nth packet on link l is scheduled by μ if that packet is ever scheduled, and $\zeta_l^\mu(n) = 0$ if that packet is never scheduled by μ. Then the deadline constraint can be stated as follows: For any n and any l with $\zeta_l^\mu(n) > 0$,

$$b_l(n) \leq \zeta_l^\mu(n) < b_l(n) + \tau_l(n).$$

In this chapter, we will consider a greedy scheduling policy, called Largest-Deficit-

First (LDF) (Hou *et al.*, 2009) based on the following *deficit process* $D^{\mu}(t)$ (also known as debts or virtual queues)

1. (Initialization) $D_l^{\mu}(0) = 0$ for all $l \in \mathcal{K}$.
2. (Dynamics) The dynamics of the deficits include the arrival and departure of the deficits.

   The arrival of the deficits are based on the arrival of the real packets and a *coin tossing process* that determines whether the arriving packet is counted as a deficit arrival or not. Let the coin tossing process for link l, denoted by $\big(C_l(n)\colon n \geq 1\big)$, be an i.i.d. Bernoulli process with mean p_l. It is assumed that $\big(C_l(n)\colon n \geq 1\big)$ is independent across l. Let $B_l(t)$ be the cumulative deficit arrival on link l given by $B_l(0) = 0$ and

   $$B_l(t) - B_l(t-1) = \sum_{n=A_l(t-1)+1}^{A_l(t)} C_l(n),$$

   where by definition $B_l(t) - B_l(t-1) = 0$ if $A_l(t-1) = A_l(t)$. Then each packet arrival is counted in the cumulative deficit arrival with probability p_l.

   The deficit decreases by one each time a packet is scheduled until it reaches zero. Hence, the evolution of the deficit process for link l is then given by

   $$\begin{aligned} D_l^{\mu}(t) = [D_l^{\mu}(t-1) + (B_l(t) - B_l(t-1)) \\ - (S_l^{\mu}(t) - S_l^{\mu}(t-1))]^+, \end{aligned}$$

   where $(\cdot)^+ = \max\{0, \cdot\}$.

Observe from the definition that the deficit process keeps track on the amount of service we owe to a link in order to fulfill the minimum delivery rate. To see that, note that the arrival rate of deficit on link l is $\bar{a}_l p_l$. The deficit of link l reduces by one when a packet is successfully transmitted over link l before its deadline. So if all

deficits are bounded, then the requirements on packet minimum delivery rates are fulfilled.

The LDF scheduling policy is defined as follows. At each time slot, LDF first sorts the links $\mathcal{K}$ according to the current deficits D with arbitrary tie-breaks, and gets the index vector I such that $D_{I_1} \geq D_{I_2} \geq \cdots \geq D_{I_K}$. LDF starts with the *selection* $\mathcal{E} = \{I_1\}$, which only consists of the link with the largest deficit. Then LDF repeatedly considers the link with the next largest deficit I_i for i from 2 to K and adds it into the selection $\mathcal{E}$ if the following two conditions are satisfied:

1. Link I_i does not interfere with any link in $\mathcal{E}$; i.e., there exists some $M_r \in M$ such that M_r schedules both $\mathcal{E}$ and I_i.
2. There is at least one packet available for transmission on link I_i; i.e., $Q_{I_i} > 0$, where Q_l is the number of available packets on link l.

The procedure ends when all links have been considered, and the final selection of links is the desired LDF schedule.

3.3 Preliminaries

In this section, we introduce basic definitions on stability and efficiency ratio that will be used in the following sections. First note that a scheduling algorithm may depend on both the actual queue information and the deficit, so the deficit process alone is not a Markov chain. We thus introduce a weaker notion of stability than positive recurrence over the entire Markov chain of actual queue and deficit. This notion of stability was first proposed by Loynes (1962) as substability, and was used by, e.g., Maguluri *et al.* (2011) and Srikant and Ying (2014) (Chapter 4.2).

Definition 1. The system is *stable* under a scheduling policy μ if the corresponding deficit process $\left(D^{\mu}(t)\colon t \geq 0\right)$ satisfies

$$\lim_{C\to\infty} \limsup_{t\to\infty} \mathbb{P}\left(\sum_{l\in\mathcal{K}} D_l^{\mu}(t) \geq C\right) = 0.$$

Note that if the deficit $\left(D(t)\colon t \geq 0\right)$ is a part of an aperiodic, irreducible and positive recurrent Markov chain, then the system is stable as defined in Definition 1, since the sum of the deficits converges in distribution as time goes to infinity.

Obviously, the stability of the system depends on the arrival distributions given by $f(\cdot)$, the maximum delay distribution given by $g(\cdot)$, and the required minimum delivery rate $p = (p_l\colon l \in \mathcal{K})$. Without loss of generality, we fix f and g and consider the stability of the system in terms of the deficit arrival rate $\lambda = (\lambda_l\colon l \in \mathcal{K})$ with $\lambda_l = \bar{a}_l p_l$. We then have the following definition for characterizing such a relation.

Definition 2. The deficit arrival rate vector λ is *supportable* by a scheduling policy if the system is stable under that policy with deficit rate λ_l for each link l.

Definition 3. The *stability region* of a scheduling policy μ is

$$\Lambda_{\mu} = \{\lambda \succeq 0\colon \lambda \text{ is supportable by } \mu\},$$

where $\succeq$ denotes pairwise greater than or equal to.

Note that the stability region defined here is different from the conventional stability region for non-real-time traffic as it is in terms of deficit arrival rates rather than packet arrival rates. This is due to the constraint that packets cannot be scheduled after their deadlines, which makes the stability of the system depend on the specific distributions of packet arrivals and deadlines. As a result, we investigate the stability by fixing f and g while varying the QoS p.

Let the set of all causal scheduling policies be $\mathcal{M}$, where a causal scheduling policy, also known as an online policy, is one that makes decision based on current information but not future information. We then have the following characterization.

Definition 4. The *maximum stability region* of the system is

$$\Lambda = \bigcup_{\mu \in \mathcal{M}} \Lambda_\mu.$$

That is, the maximum stability region of the system is the set of deficit arrival vectors that can be supported by some causal scheduling policy.

For a given scheduling policy μ, the efficiency ratio of the scheduling policy is defined as follows.

Definition 5. The *efficiency ratio* of a scheduling policy μ is

$$\gamma_\mu^* = \sup \left\{ \gamma \colon \gamma \Lambda \subseteq \Lambda_\mu \right\}.$$

While refined characterizations of the stability region are possible (Li *et al.*, 2011, 2012), the efficiency ratio is still a critical metric to evaluate the throughput performance of a scheduling policy.

3.4 Main Results

In this section we present the main results of the LDF policy for scheduling real-time traffic in wireless networks. The first result is Theorem 2, which provides a lower bound on the efficiency ratio of the LDF policy, called the real-time local-pooling factor (R-LPF), in the case when the traffic distributions are known. The second result is Theorem 3, which gives lower bounds on the R-LPF regardless of the traffic distributions by assigning weights to the links and calculating the ratio of the weighted sum of the LDF schedule to that of the optimal schedule.

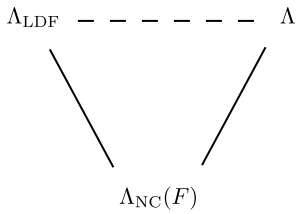


Figure 8: A roadmap of the proof of the lower bound in Theorem 2

We provide a roadmap of the proof of Theorem 2 in Figure 8. The goal of Theorem 2 is to establish the connection between Λ, the maximum stability region of the system, and Λ_{LDF}, the stability region of the LDF policy. However, characterizing Λ turns out to be extremely difficult due to the general arrival and maximum delay distributions. We therefore have to introduce a region called $\Lambda_{\text{NC}}(F)$, which is the maximum stability region by dividing the time into frames with length F and assume (i) all information within a frame (arrivals and maximum delays) are known at the beginning of a frame and (ii) at the end of a frame, all packets that have not been transmitted are dropped. The region is denoted by $\Lambda_{\text{NC}}(F)$ since the frame length is F and the system is non-causal because of condition (i).

This novel frame concept was first introduced by Hou *et al.* (2009) for real-time scheduling in wireless networks and provides an analytical framework for understanding real-time communication in wireless networks. The framework has then been extended to general networks and traffic patterns. In particular, the capacity of the non-causal system and heterogeneous deadlines has been characterized by Jaramillo *et al.* (2011); i.e., $\Lambda_{\text{NC}}(F)$ is known.

We will use $\Lambda_{\text{NC}}(F)$ to bridge Λ and Λ_{LDF}. In the theorem, we will first show that

$$\text{int}(\Lambda) \subseteq \liminf_{F\to\infty} \Lambda_{\text{NC}}(F),$$

where $\text{int}(\Lambda)$ is the interior of the set Λ and $\liminf_{F\to\infty} \Lambda_{\text{NC}}(F)$ is the limit set of

$\Lambda_{\mathrm{NC}}(F)$ as F goes to infinity. After that, we will prove that

$$\sigma^* \operatorname{int}\left(\Lambda_{\mathrm{NC}}(F)\right) \subseteq \Lambda_{\mathrm{LDF}},$$

where σ^* is a constant, called the real-time local-pooling factor whose definition is presented in Section 3.4.1. Combining the two results together, we will be able to prove that σ^* is a lower bound on the efficiency ratio. *We remark that the second step is non-trivial since we will compare the time-slot-based, causal LDF (not frame-based LDF) with the frame-based, non-causal system.*

As for the second result regarding lower bounds on the R-LPF, we first reformulate the dual problem of solving the R-LPF as a weight assignment problem following Li *et al.* (2011) In the weight assignment problem we try to maximize the ratio of the smallest weighted sum of a schedule to the largest one over different weights within a frame. A similar result for the special case of all-one weight assignment has been observed by Reddy *et al.* (2012) for characterizing the local-pooling factor for fading channels. We then look at the multigraph of the network for any given traffic pattern in a frame, and obtain further lower bounds on the R-LPF using the local weight ratios which do not require the traffic information. Our result immediately implies that the R-LPF is at least $1/(\beta+1)$ for networks with interference degree β, regardless of the distributions of the packet arrivals and the maximum delays.

3.4.1 Real-Time Local-Pooling Factor

We will define a quantity analogous to the local-pooling factor proposed by Joo *et al.* (2009b) and the fading local-pooling factor studied by Reddy *et al.* (2012) Before we do that, we need the following two definitions.

Definition 6. A non-causal frame-based scheduling policy μ with frame size F (called

an *F-framed* policy for abbreviation) is defined as follows. The packet arrivals and deadlines in the kth frame are known to the policy μ at the beginning of the frame and all packets that arrive during the kth frame are dropped at the end of the frame if not transmitted. Formally, for any $l \in \mathcal{K}$ and positive integer n with $\zeta_l^\mu(n) > 0$, there exists a positive integer k such that

$$kF + 1 \leq b_l(n) \leq \zeta_l^\mu(n) \leq (k+1)F,$$

where $\zeta_l^\mu(n)$ was defined in Section 3.2 in the deadline constraint condition.

Let the set of all F-framed policies be $\mathcal{M}_{\text{NC}}(F)$. Note that $\mathcal{M}_{\text{NC}}(F)$ is not a subset of $\mathcal{M}$ since policies in $\mathcal{M}_{\text{NC}}(F)$ can be non-causal. The frame concept (alternatively called intervals or periods) has been used in the literature for tractable analytical analysis of delay constrained traffic (Hou and Kumar, 2009, 2010a,b; Jaramillo and Srikant, 2011; Jaramillo *et al.*, 2011), where packets that arrive in a frame have deadlines in the same frame. In this section, we adopt this concept to derive the real-time local-pooling factor for the general traffic model.

Definition 7. The *maximum stability region of F-framed policies* for a positive integer F is

$$\Lambda_{\text{NC}}(F) = \bigcup_{\mu \in \mathcal{M}_{\text{NC}}(F)} \Lambda_\mu.$$

We now introduce some notations needed for the main results. Let $\mathcal{J}(F)$ be the set of arrival and maximum delay patterns in a frame of F time slots. We will call an element of $\mathcal{J}(F)$ an F-pattern. An F-pattern is represented by $J = (A^{(F)}, \tau^{(F)})$ with $A^{(F)} = \left(A_l^{(F)}(t) \colon l \in \mathcal{K}, 1 \leq t \leq F\right)$ and $\tau^{(F)} = \left(\tau_l^{(F)}(n) \colon l \in \mathcal{K}, 1 \leq n \leq A_l^{(F)}(F)\right)$, where $A_l^{(F)}(t)$ is the cumulative packet arrival to link l by time slot t in the frame, and $\tau_l^{(F)}(n)$ is the maximum delay associated with the nth packet on link l. Thus, $A_l^{(F)}(F)$

is the total number of arrivals in the frame on link l. Due to the i.i.d. distributions of the arrival and maximum delay given by f and g, there is a stationary distribution of the set of F-patterns, denoted by $\pi\colon \mathcal{J}(F) \to \mathbb{R}$.

For a given F-pattern $J = (A^{(F)}, \tau^{(F)})$, a *schedule* $s = \left(s_l(n)\colon l \in \mathcal{K}, 1 \le n \le A_l^{(F)}(F)\right)$ specifies the time slot at which each packet is scheduled to be transmitted (if it ever gets scheduled), where $s_l(n)$ is a nonnegative integer that indicates the nth packet on link l is scheduled at time slot $s_l(n)$ if $s_l(n) \in \{1, 2, \ldots, F\}$, and is never scheduled if $s_l(n) = 0$. A schedule s is *feasible* for the F-pattern J if 1) each scheduled packet is scheduled within its feasible scheduling window, 2) at most one packet is scheduled on each link in one time slot, and 3) the set of links with packets scheduled in each time slot forms a feasible link schedule; i.e., it is a subset of $\left\{l \in \mathcal{K}\colon M_{l,r} = 1\right\}$ for some r. Note that the schedule s here is different from the link schedules defined in Section 3.2 in that s specifies the scheduling of each packet in the whole frame, and that s needs to take into account the traffic so that no scheduling is allowed before the arrival or after the deadline of a packet. We also say that a schedule s is *maximal* for J if no more packets can be further scheduled (i.e., no $s_l(t)$ can be changed from 0 to a positive integer) without breaking feasibility. We denote the maximal feasible schedules for J by $S^*(J)$.

We define the *total service vector of schedule* s to be the column vector $W(s) = \left(W_i(s)\colon i \in \mathcal{K}\right)$ with $W_i(s) = \sum_{n=1}^{A_i(F)} \mathbb{1}_{\{s_i(n)\neq 0\}}$, where $\mathbb{1}_{\{s_i(n)\neq 0\}}$ is the indicator function. Then $W(s)$ is the vector of total number of scheduled packets on each link for the schedule s. Let the *maximal service matrix for* J be

$$M_J = \left\{W(s)\colon s \in S^*(J)\right\},$$

where again M_J represents both the set and the matrix consisting of the vectors as its columns, by abuse of notation. Then the columns of M_J are the total service vectors

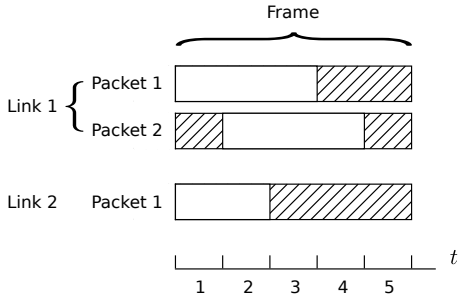


Figure 9: An example of a 5-pattern for two links

of the maximal schedules. We note that M_J does not contain all-zero columns if and only if J includes at least one packet arrival on some link, since schedules in $S^*(J)$ are maximal. Similarly, define $M_{J,L}$ to be the maximal service matrix restricted to the set of links L for given pattern J. Then $M_{J,L}$ has no all-zero columns if and only if the pattern J includes at least one packet on some link in L. Also note that $M_{J,L}$ has $|L|$ rows while M_J has K rows.

We use the example in Figure 9 to illustrate the above notations and the concept of the maximal service matrix. As shown in the figure, we consider a frame with size 5 and a 5-pattern J with two packets arriving to link 1 and one packet arriving to link 2, whose arriving times and deadlines are indicated by the blank bars in the figure. The corresponding pattern can be represented by $J = (A^{(5)}, \tau^{(5)})$, where $\left(A_1^{(5)}(t)\colon t \geq 0\right) = (0,1,2,2,2,2)$, $\left(A_2^{(5)}(t)\colon t \geq 0\right) = (0,1,1,1,1,1)$, $\left(\tau_1^{(5)}(n)\colon n \geq 1\right) = (3,3)$, and $\left(\tau_2^{(5)}(n)\colon n \geq 1\right) = (2)$. Assume the two links interfere with each other, so at each time slot only one of them can be scheduled. We can check that there are eight maximal feasible schedules in $S^*(J)$ as follows:

$$s^1 = \begin{pmatrix} (1,2) \\ (0) \end{pmatrix}, s^2 = \begin{pmatrix} (1,3) \\ (2) \end{pmatrix}, s^3 = \begin{pmatrix} (1,4) \\ (2) \end{pmatrix}, s^4 = \begin{pmatrix} (2,3) \\ (1) \end{pmatrix},$$

$$s^5 = \begin{pmatrix} (2,4) \\ (1) \end{pmatrix}, s^6 = \begin{pmatrix} (3,2) \\ (1) \end{pmatrix}, s^7 = \begin{pmatrix} (3,4) \\ (1) \end{pmatrix}, s^8 = \begin{pmatrix} (3,4) \\ (2) \end{pmatrix},$$

where the first row of s^i is the schedule for the two packets on link 1, and the second row is the schedule for the packet on link 2. Then the total service vectors are

$$W(s^1) = \begin{pmatrix} 2 \\ 0 \end{pmatrix} \text{ and } W(s^i) = \begin{pmatrix} 2 \\ 1 \end{pmatrix} \text{ for } 2 \leq i \leq 8.$$

Hence the maximal service matrix is

$$M_J = \begin{pmatrix} 2 & 2 \\ 0 & 1 \end{pmatrix}.$$

We remark from the above example that unlike in the scenario of non-real-time traffic (Dimakis and Walrand, 2006), the total service vector of one maximal schedule could be dominated by that of another in the real-time setting. Thus the maximal service matrix M_J can be huge and hard to compute, especially for large frame size F and complex traffic pattern J.

Definition 8. The *real-time local-pooling factor* (R-LPF) for the F-framed scheduling policies for the set of links $L \subseteq \mathcal{K}$ is

$$\sigma_L^*(F) = \inf\left\{\sigma \colon \exists \phi_1, \phi_2 \in \Phi_L(F) \text{ such that } \sigma\phi_1 \succeq \phi_2\right\},$$

where $\Phi_L(F)$ is the *service region restricted to the set of links* $L \subseteq \mathcal{K}$ defined by

$$\Phi_L(F) = \left\{\phi \colon \phi = \sum_{J \in \mathcal{J}(F)} \pi(J)\eta_J, \eta_J \in \mathcal{CH}(M_{J,L})\right\},$$

and $\mathcal{CH}(M_{J,L})$ defines the convex hull over the columns of the matrix $M_{J,L}$.

Definition 9. The R-LPF for the F-framed scheduling policies is

$$\sigma^*(F) = \min_{L \subseteq \mathcal{K}} \sigma_L^*(F).$$

Definition 10. The R-LPF for the system is

$$\sigma^* = \liminf_{F\to\infty} \sigma^*(F).$$

We then have the following theorem stating that the R-LPF is a lower bound on the efficiency ratio of LDF.

Theorem 2. $\gamma^*_{\mathrm{LDF}} \geq \sigma^*$.

Intuitively when the frame length goes to infinity the loss at the edge of frames becomes negligible. The proof of Theorem 2 uses the strictly separating hyperplane theorem (Boyd and Vandenberghe, 2004) and follows the fluid limit technique that was first proposed by Dai (1995) for multiclass queueing systems and later developed for discrete-time generalized switches by Andrews *et al.* (2004) and further used in wireless networks by Reddy *et al.* (2012) and Ji *et al.* (2013) The complete proof is presented in Appendix B.1.

By the definition of the R-LPF, we can get the R-LPF by solving the following linear program for each $L \subseteq \mathcal{K}$, as suggested by Li *et al.* (2011) and Reddy *et al.* (2012):

$$\begin{aligned} \sigma^*_L(F) = \min_{\sigma,\rho,\theta} \quad & \sigma \\ \text{s.t.} \quad & M_L(F)\theta - M_L(F)\rho \succeq 0 \\ & \mathbf{1}^{\mathrm{T}}\theta - \sigma = 0 \\ & \mathbf{1}^{\mathrm{T}}\rho - 1 = 0 \\ & \rho, \theta \in \mathbb{R}^r_+, \end{aligned} \tag{3.1}$$

where $M_L(F)$ is the vertices of the polygon $\sum_{J\in\mathcal{J}(F)} \pi(J)\mathcal{CH}(M_{J,L})$ (the summation is in the sense of the Minkowski sum and $\mathcal{CH}(M_{J,L})$ denotes the convex hull of the

column vectors in $M_{J,L}$, where $r_{J,L}$ denotes the number of columns), $r = \sum_{J\in\mathcal{J}(F)} r_{J,L}$ is the cardinality of $M_L(F)$, ρ and θ are nonnegative column vectors of length r. That said, computing the exact R-LPF is usually complex, as it involves roughly

$$\sum_{L\subseteq\mathcal{K}} \left(2 \sum_{J\in\mathcal{J}(F)} r_{J,L} + |L| \right)$$

constraints for each F, which increases exponentially with both the size of the network K and the frame size F. Thus, we seek lower bounds on the R-LPF in the next subsection.

3.4.2 Characterizing the R-LPF

3.4.2.1 Dual of the R-LPF

The dual problem of (3.1) is given by the following (Li *et al.*, 2011)

$$\begin{aligned} \omega_L^*(F) = \max_{\alpha_L,\omega} \quad & \omega \\ \text{s.t.} \quad & \mathbf{1}^{\mathrm{T}} \succeq \alpha_L^{\mathrm{T}} M_L(F) \succeq \omega \mathbf{1}^{\mathrm{T}} \\ & \alpha_L \in \mathbb{R}_+^{|L|} \\ = \max_{\alpha_L\in\mathbb{R}_+^{|L|}} & \frac{\min_{\phi\in M_L(F)} \alpha_L^{\mathrm{T}}\phi}{\max_{\phi\in M_L(F)} \alpha_L^{\mathrm{T}}\phi}, \end{aligned} \tag{3.2}$$

where we adopt the useful convention $\frac{0}{0} = 0$. By the strong duality of the linear program (3.1), finding the R-LPF for subset L and frame size F is equivalent to the weight assignment problem (3.2); i.e., $\sigma_L^*(F) = \omega_L^*(F)$.

Another interpretation of the dual problem is that since $\alpha_L^{\mathrm{T}}\phi$ is the length of the projection of ϕ along the vector α_L (module the length of α_L), the optimization problem (3.2) is to find the best projection direction α_L for each subset L such that

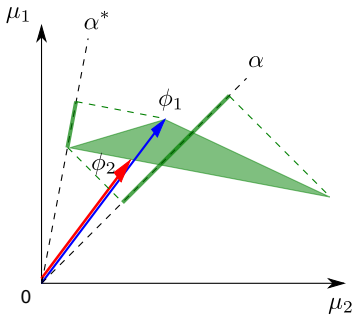


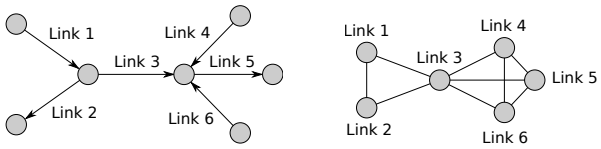
Figure 10: Illustration of the projection interpretation of the dual formulation of $\omega_L^*(F)$ with $L = \mathcal{K} = \{1, 2\}$

Note: The shaded triangular area is the convex hull of $M_L(F)$.

the ratio of the smallest projection to the largest one from $M_L(F)$ is maximized. Note that for any α_L we have a corresponding ratio of the smallest to the largest projections, which is a valid lower bound on $\sigma_L^*(F)$. We illustrate this interpretation in Figure 10 with $L = \mathcal{K} = \{1, 2\}$. In the figure, α^* is the optimal projection direction since the ratio of the smallest vector (ϕ_2 projected to the direction of α^*) to the largest vector (ϕ_1 projected to the direction of α^*) in the projection along the direction of α^* is the maximum (it equals $\omega_L^*(F)$) among all the possible projection directions. The vector α in the figure is an arbitrary direction. The ratio of the smallest vector projection to the largest vector projection is smaller than that along the direction α^*, as shown in the shaded segment along α, and thus provides a lower bound on $\omega_L^*(F)$.

3.4.2.2 Lower Bounds for Conflict Graph Interference Model

While the dual problem (3.2) gives a different view of the original problem for solving the R-LPF, the problem is not simplified since the size of $M_L(F)$, where $M_L(F)$ is taken as the set of columns, grows exponentially with F. As a result, we



(a) Original graph (b) Conflict graph

Figure 11: An original graph and its conflict graph

are interested in finding a lower bound on the R-LPF that can be computed efficiently (without calculating the Minkowski sum of the maximal service matrices for all traffic pattern and all frame size). In this subsection we introduce the lower bounds on R-LPF for networks with interference models that are represented by conflict graphs (Jain *et al.*, 2003) (also known as interference graph), where two links either interfere with each other exclusively, or do not interfere at all. An example of the original graph and its conflict graph is given in Figure 11. We introduce the ideas of pressure and minimum pressure in the following, which use the local information to estimate the global optimal value in (3.2).

For any $\alpha \in \mathbb{R}_+^K$, we define the *pressure* of link i to be

$$\kappa_i(\alpha) = \frac{\alpha_i}{\alpha_i + \max_{I \in \mathcal{I}(i)} \sum_{j \in I} \alpha_j}, \tag{3.3}$$

where $\mathcal{I}(i) \subseteq \mathcal{P}(N(i))$ is the collection of subsets of the neighbors of i that can be scheduled simultaneously ($N(i)$ is the set of links that interfere with link i, and $\mathcal{P}(\cdot)$ denotes the power set). We also define the *minimum pressure* for given vector $\alpha \in \mathbb{R}_+^K$ by

$$\psi(\alpha) = \min_{i \in \mathcal{K}} \kappa_i(\alpha). \tag{3.4}$$

So $\psi(\alpha)$ is just the lowest pressure for α over all links. Then we have the following lower bound on the R-LPF.

Theorem 3.

$$\sigma^* \geq \sup_{\alpha \in \mathbb{R}_+^K} \psi(\alpha).$$

Intuition and proof outline. Given an arbitrary vector of nonnegative weights α on the links, we define

$$G(\alpha, L, F) = \frac{\min_{\phi \in M_L(F)} \alpha_L^{\mathrm{T}} \phi}{\max_{\phi \in M_L(F)} \alpha_L^{\mathrm{T}} \phi},$$

where α_L is the vector of α restricted to the subset L. Then $G(\alpha, L, F)$ is the minimum global weight ratio of schedules for subset L and frame size F. By the dual representation in (3.2), the R-LPF is lower-bounded by the smallest $G(\alpha, L, F)$ over all possible L and F. Note that while $G(\alpha, L, F)$ is defined over $M_L(F)$ which is averaged over all possible traffic patterns in $\mathcal{J}(F)$ according to the traffic distributions, we can work on an arbitrary traffic $J \in \mathcal{J}(F)$ to establish a universal lower bound that holds for any $J \in \mathcal{J}(F)$, which will also be a lower bound on $G(\alpha, L, F)$. Since

$$G(\alpha, L, F) \geq \frac{\min_{\phi \in M_{J,L}} \alpha_L^{\mathrm{T}} \phi}{\max_{\phi \in M_{J,L}} \alpha_L^{\mathrm{T}} \phi}$$

for any $J \in \mathcal{J}(F)$, we only need to lower-bound the ratio of the weights of two maximal schedules given the specific traffic. Using a *multigraph* representation of the network, for any $J \in \mathcal{J}(F)$ and any $\phi_1, \phi_2 \in M_{J,L}$, we can divide the scheduled packets of ϕ_1 and ϕ_2 into groups such that

$$\frac{\alpha_L^{\mathrm{T}} \phi_1}{\alpha_L^{\mathrm{T}} \phi_2} \geq \frac{\sum_{i=1}^N x_i}{\sum_{i=1}^N y_i},$$

where N is the total number of packets scheduled in ϕ_1, x_i is the total weight of the ith packet in ϕ_1, and y_i is the weight of neighboring packets of the ith packet of ϕ_1 scheduled by ϕ_2. By the group construction and the definition of minimum pressure $\psi(\alpha)$, we have $\frac{x_i}{y_i} \geq \psi(\alpha)$ for any i regardless of L and F. Then $G(\alpha, L, F)$ is lower-bounded by the minimum pressure $\psi(\alpha)$. Since the minimum pressure $\psi(\alpha)$ is

determined by the network topology and does not depend on L or F, we get a lower bound on the R-LPF, which can then be strengthened to the form of Theorem 3 by optimizing over α. The detailed proofs can be found in Appendix B.2.

Remark 5. We emphasize that while the R-LPF involves computing the maximal service matrices for all subsets of $\mathcal{K}$ and all traffic patterns under all frame sizes, Theorem 3 states that the maximum value of the minimum pressure, which is a purely topological quantity for the entire network, serves as a lower bound on the R-LPF.

Remark 6. Theorem 3 implies that any $\alpha \in \mathbb{R}_+^K$ will give a lower bound on the R-LPF by the corresponding minimum pressure $\psi(\alpha)$. In particular, we have the following corollary.

Corollary 2. *For a network with an interference degree of* β,

$$\sigma^* \geq \frac{1}{\beta + 1}.$$

Corollary 2 follows directly from Theorem 3 by setting $\alpha = \mathbf{1}$.

We note that this translates to $\sigma^* \geq 1/3$ in the one-hop interference model, where the interference degree is at most 2 (this can be easily proved by noticing that the neighboring links of link 3 in Fig 11 form two cliques in the conflict graph). This is related to the well-known result that LQF has efficiency ratio at least 1/2 in packet switches (Weller and Hajek, 1997; Dai and Prabhakar, 2000) and in wireless networks under the one-hop interference (Lin and Shroff, 2005; Wu and Srikant, 2005), where $\beta = 2$. The lower bound on the efficiency ratio of the greedy scheduling policy decreases from 1/2 in the non-real-time case to 1/3 in the real-time case due to the temporal correlation among packets brought by deadlines, which does not exist in non-real-time traffic. This can be illustrated by considering a star network with one center link and two leaf links, where the center link interferes with any leaf link but

the two leaf links do not interfere each other. Suppose one packet arrives at the center link at the beginning of time slot 1 with deadline at the end of time slot 2, and one packet arrives at each of the two leaf links at the beginning of time slot 1 with deadline at the end of time slot 1. Then the optimal scheduler will schedule the two leaf links at time slot 1 and the center link at time slot 2, while LDF may schedule the center link at time slot 1 and nothing at time slot 2 (since the packets on the leaf links have already expired), which results in an efficiency ratio of 1/3. Note that if the above arrival pattern is for non-real-time packets (i.e., there is no deadline of the packets), then the longest-queue-first (LQF) can schedule at least one packet at each time slot when at least one of the queues is not empty, while the optimal scheduler may schedule at most two packets at each time slot, so LQF guarantees at least half throughput. To sum up, LDF has a smaller efficiency ratio lower bound than its non-real-time counterpart LQF because it may inevitably schedule the "wrong" packets due to its inability to take into account the consequences in the future of its current decisions, and this cannot be compensated by future actions.

3.4.2.3 Lower Bounds on R-LPF for Special Networks

We now do a case study of lower bounds on R-LPF for some special networks via the minimum pressure technique stated in Theorem 3.

3.4.2.3.1 Collocated Network

We first consider the scenario of the collocated network, where at most one link can be scheduled at each time slot. Notice that the interference degree of the network

is $\beta = 1$ since in any subset of the links there is at most one link that can be scheduled. Then by Theorem 2 and Theorem 3, the efficiency ratio is at least 1/2.

3.4.2.3.2 Star Networks

Consider a star network with interference degree β. Then this network consists of one center link and β leaf links, where at each time slot either the center link or all the leaves can be scheduled. By setting the weight of the center link to be $\sqrt{\beta}$ and the weight of each leaf to be 1, we get the minimum pressure $\psi = 1/(\sqrt{\beta}+1)$. Hence the efficiency ratio is at least $1/(\sqrt{\beta}+1)$ for the star network with interference degree β.

3.4.2.3.3 Tree Networks

For tree networks with interference degree β, a lower bound on the R-LPF is given in the following corollary.

Corollary 3. $\sigma^* \geq \frac{1}{2\sqrt{\beta-1}+1}$.

The proof can be found in Appendix B.3.

Remark. Note that the interference degree is equal to the largest degree of a link in the trees. Also note that this lower bound is better than the $1/(\beta+1)$ minimum pressure bound given by the all-one vector for $\beta \geq 3$.

3.5 The Consensus Algorithm

We design an algorithm that can be used to compute a lower bound on the R-LPF. Let each link i maintain a weight α_i and a *pressure* κ_i. In each time slot all links

broadcast (α_i, κ_i) to its neighbors and update its weight and pressure by

$$\Delta\alpha_i = z \sum_{j \in N(i)} (\kappa_j - \kappa_i)$$

and

$$\kappa_i = \frac{\alpha_i}{\alpha_i + \max_{I \in \mathcal{I}(i)} \sum_{j \in I} \alpha_j}.$$

The constant z can be interpreted as the step size. The intuition behind the simple heuristic algorithm is that if under current weight assignment one link has pressure greater than those of its neighbors, then the weight of that link should be transferred to its neighbors so that the minimum of their pressures can increase. Likewise, when one link has pressure less than those of its neighbors, then the weights on its neighbors should be transferred to that link to make the minimum pressure higher. We would expect the algorithm to converge as time goes by when the step size is sufficiently small; i.e., the weights and pressures for all links remain unchanged eventually. Then the weight vector that our algorithm converges to yields a minimum pressure that lower-bounds the R-LPF. Note that we call this algorithm "consensus" because the pressures for the links will usually converge to the same value and reach consensus. We evaluate the performance of the consensus algorithm in Section 3.7.

3.6 Discussions

3.6.1 Efficiency Ratios Under Adversarial Traffic

In this section we discuss the performance of LDF under adversarial traffic. We consider a more general type of traffic, where, instead of i.i.d., the packet arrival, maximum delay, deficit arrival and tie-breaking processes are relaxed to be irreducible

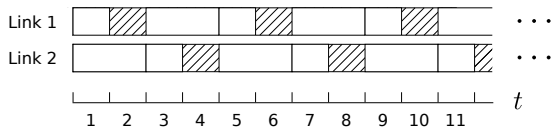


Figure 12: An adversarial traffic pattern for a collocated network with two links

Note: Each blank bar indicates the arriving time and deadline of a real-time packet. The packets arrive on link 1 at the beginning of time slots 1, 3, 5, 7, ..., and must be scheduled before the end of time slots 1, 4, 5, 8, ... The packets arrive on link 2 at the beginning of time slots 1, 3, 5, 7, ..., and must be scheduled before the end of time slots 2, 3, 6, 7, ...

positive recurrent Markov chains as by Andrews *et al.* (2004) We illustrate that when the traffic is adversarial in this general type, the efficiency ratio of LDF can be as low as $1/(\sqrt{\beta}+1)$, where β is the interference degree. In particular, for collocated networks the efficiency ratio of LDF under adversarial traffic is consistent with the lower bound given in Corollary 2.

We start with a collocated network with two links. Consider the adversarial traffic given in Figure 12. Assume that the deficits for both links are the same at the beginning of time slot 1. Also assume that when there is a packet arriving to each link (time slots 1, 3, 5, 7, ...), the deficits on both links increase by one with probability $1/2+\epsilon$ for some small positive ϵ, and remain unchanged with probability $1/2-\epsilon$. This results in minimum delivery rates $p_i = 1/2+\epsilon$ for $i = 1, 2$. We further assume that when the deficits on the two links are equal, the tie-breaking rule of LDF gives priority to link 2. Then one can easily see that LDF schedules link 2 at time slots 1, 5, 9, ..., and schedules link 1 at time slots 3, 7, 11, ..., while LDF idles at even time slots. Then the average deficit arrival to each link per time slot is $1/4+\epsilon/2$, and the average deficit departure from each link per time slot is $1/4$. Hence the deficits are not stable under LDF given this traffic pattern. However, one would notice that

the optimal scheduler could schedule link 1 in time slots $4k$ and $4k+1$ and schedule link 2 in time slots $4k+2$ and $4k+3$, for all positive integer k. Hence the optimal scheduler can stabilize the system when the minimum delivery rates are $p_i = 1$ for $i = 1, 2$. By making ϵ arbitrarily small we can see that the efficiency ratio of LDF is at most $1/2$ in this two-link collocated network, which meets with the lower bound given in Corollary 2.

We now consider a general network. In the following theorem we construct an adversarial traffic process.

Theorem 4. *There exists a traffic pattern distribution such that* $\gamma^*_{\mathrm{LDF}} \leq \frac{1}{\sqrt{\beta}+1}$, *where* β *is the interference degree.*

The theorem can be proved by finding the link with interference degree β and constructing a specific traffic pattern on that graph. The detailed proof can be found in Appendix B.4.

3.6.2 Extension to Heterogeneous Link Rates

The LDF policy can be generalized to heterogeneous integer-valued link rate scenario following Dimakis and Walrand (2006). Assume the link rate for link i is $c_i \in \mathbb{N}$ for $i \in \mathcal{K}$. LDF now schedules c_i packets instead of 1 packet on each selected link i. Then Theorem 2 still holds by replacing the 1's on the ith row of M with c_i, and Theorem 3 still holds after redefining the pressure by

$$\kappa_i(\alpha) = \frac{\alpha_i}{\alpha_i + \max_{I \in \mathcal{I}(i)} \sum_{j \in I} \alpha_j c_j}.$$

Note that in the summation of the denominator all the weights α_j's are multiplied by the corresponding link rate c_j, while α_i in both the denominator and the numerator

are not multiplied by c_i. Intuitively this is due to the fact that in the worst case LDF can schedule only one packet on link i as opposed to $\max_{I \in \mathcal{I}(i)} \sum_{j \in I} c_j$ packets scheduled by the optimal policy on the neighboring links of i. The consensus algorithm can also be modified according to the new definition of pressure.

3.7 Simulations

In this section we use simulations to evaluate the stability performance of LDF, as well as the consensus algorithm we proposed.

3.7.1 Stability Performance

Since to the best of our knowledge, neither the maximum stability region nor an optimal scheduling policy has been obtained in the literature, we do not have a benchmark for the stability performance of LDF. As a result, we compare LDF to two other scheduling policies that do not depend on frames and evaluate the performance using simulations. The first simple scheduling policy we consider is RandMax, which randomly chooses a maximal schedule over the links with packets in each time slot. The other one is MaxWeight, which chooses a maximal schedule with the maximum deficit sum over the links with packets in each time slot.

We first considered a 4-link linear network with one-hop interference. We assumed the packet arrival distribution is binomial with number of trials 2 and success probability 0.5, and the maximum delay distribution is uniform over $\{2, 3, 4\}$. This gives us packet arrival rate $\bar{a} = 1$ and mean maximum delay $\bar{\tau} = 3$. We varied the minimum delivery rate to vary the deficit arrival rate. We compare the average deficit sums

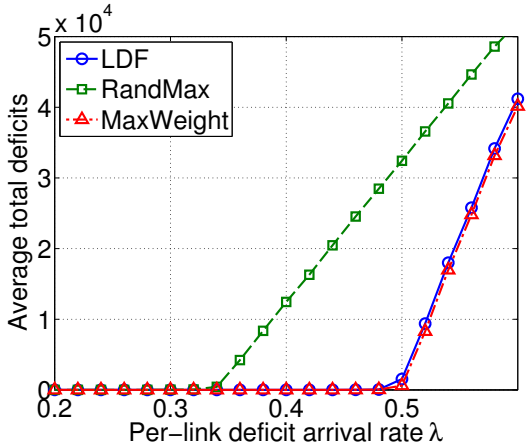


Figure 13: Comparison of the three scheduling policies on a four-linear network with one-hop interference

of the last 1,000 iterations under the three policies, where each simulation is run for 100,000 iterations. The results are shown in Figure 13.

As can be observed from the figure, LDF and MaxWeight have similar stability performance, achieving a maximum deficit arrival rate of roughly 0.5 and significantly outperform the simple RandMax policy, which achieves a maximum deficit arrival rate of roughly 0.33. We further remark that for non-real-time traffic, the maximum deficit arrival rate is 0.5. Thus both LDF and MaxWeight have a near-optimal performance in this case.

We also consider a nine-cycle network with two-hop interference, whose non-real-time local-pooling factor is 2/3. The arrival and deadline distributions are the same as the previous case, and the number of iterations is 100,000. The results are shown in Figure 14.

Note that in this example, RandMax is still the worst of the three, achieving a maximum deficit arrival rate roughly 0.12, while MaxWeight is slightly better than LDF, both of which achieve a maximum deficit arrival rate roughly 0.16. We note

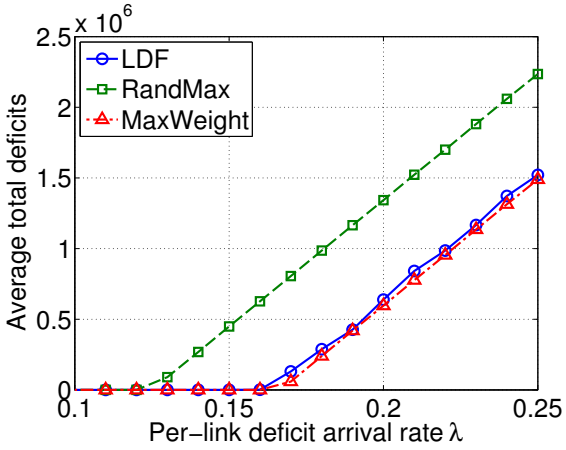


Figure 14: Comparison of the three scheduling policies on a nine-cycle network with two-hop interference

that for non-real-time traffic the maximum deficit arrival rate is 1/3. As we have been trying to convey in this chapter, the maximum stability region for the specific packet arrival and deadline distribution is unknown. We only know that the maximum rate for the real-time traffic is $\bar{\lambda} \leq 1/3$. Note that the nine-cycle has an interference degree of 2, so by Theorem 3, LDF has an efficiency ratio of at least 1/3, which agrees with the simulation result since $0.16 > \frac{1}{3} \times \frac{1}{3} \geq \frac{1}{3}\bar{\lambda}$.

Therefore, both simulations imply good throughput performance of LDF and validate our lower bound on the efficiency ratio.

3.7.2 Performance of the Consensus Algorithm

We now study the performance of the consensus algorithm we proposed in Section 3.5. We run the consensus algorithm on random networks with the unit-disk interference model used by Joo *et al.* (2009b)

We place 32 nodes randomly in a unit square area. Any two nodes with distance less than the communication range $r_c = 0.25$ may form a link. The default maximal

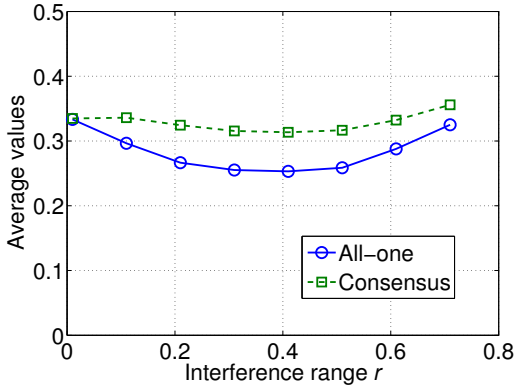
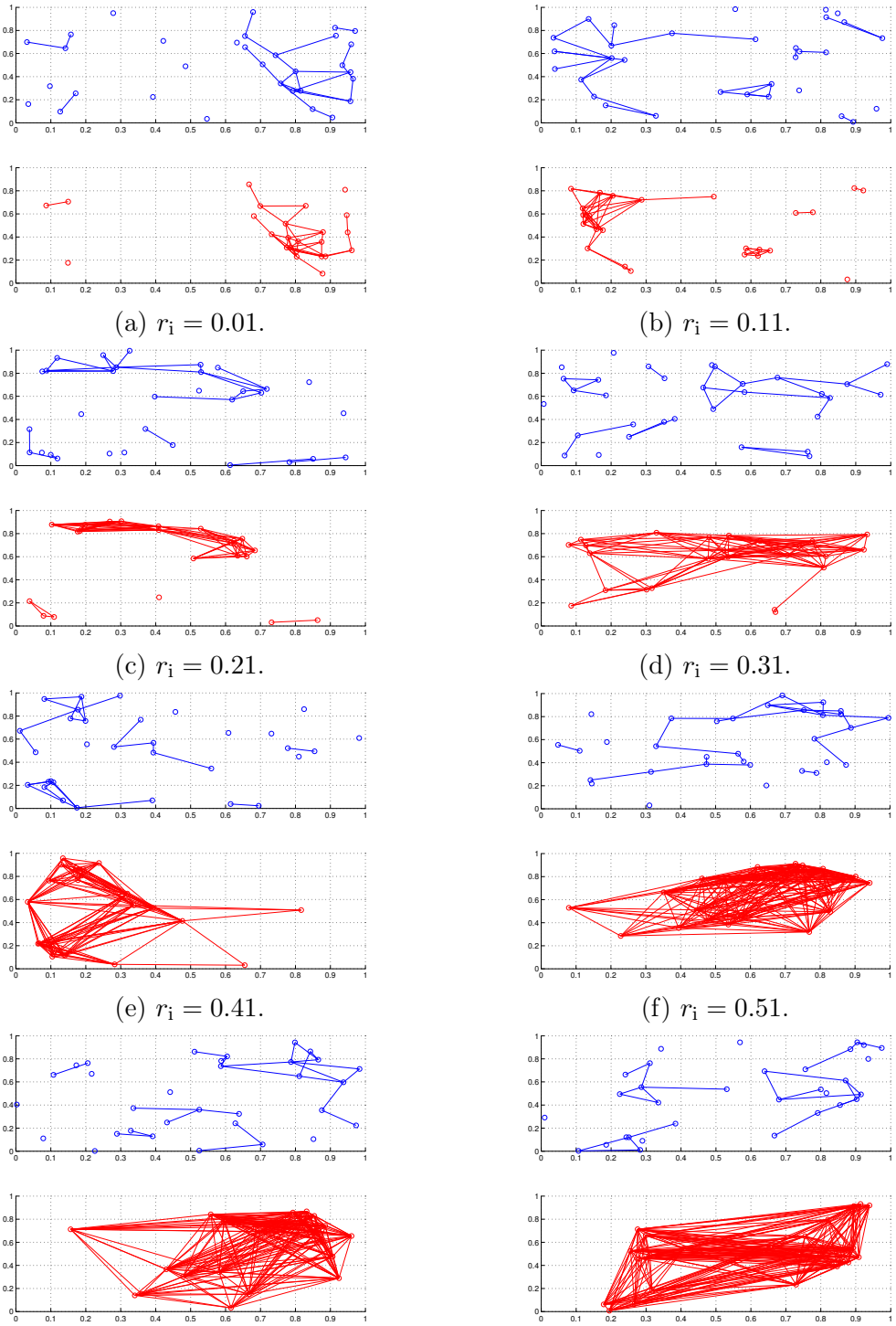


Figure 15: Lower bounds on the R-LPF given by the consensus algorithm and the all-one algorithm for different interference range

number of links is 24 (uniformly chosen from possible links). Any two links with minimum node distance less than the interference range r_{i} interfere each other, and the default interference range is $r_{\mathrm{i}} = 0.4$.

We run the consensus algorithm for 3000 iterations with step size $z = 0.1$ for each network. We compare the average lower bound obtained via the consensus algorithm to the all-one algorithm in Figure 15. Each point is the average of 100 random networks. We also attach one example network for each interference range r_{i} in Figure 16. We see from Figure 15 that our consensus algorithm achieves a much better lower bound than the one given by the interference degree alone.

We show an example of traces of the consensus algorithm for a random unit-disk network with $r_{\mathrm{c}} = 0.25$ and $r_{\mathrm{i}} = 0.11$ in Figure 17. The lines in the upper figure of Figure 17 are the weights α_i's for the links $i \in \mathcal{K}$ with respect to the iterations, and the lower figure of Figure 17 shows the minimum pressure given by the weights with respect to the iterations. We see that the weights become unchanged after about 300 iterations, which indicates that the consensus algorithm converges reasonably fast.



(a) $r_{\mathrm{i}} = 0.01$. (b) $r_{\mathrm{i}} = 0.11$.

(c) $r_{\mathrm{i}} = 0.21$. (d) $r_{\mathrm{i}} = 0.31$.

(e) $r_{\mathrm{i}} = 0.41$. (f) $r_{\mathrm{i}} = 0.51$.

(g) $r_{\mathrm{i}} = 0.61$. (h) $r_{\mathrm{i}} = 0.71$.

Figure 16: Example networks for different interference range r_{i}

Note: The top figures are the original node-link graphs, while the bottom ones are the corresponding conflict graphs.

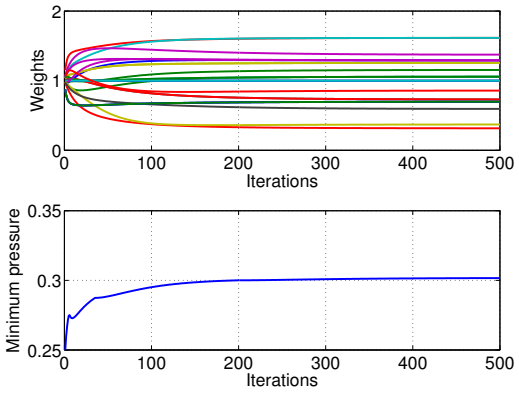


Figure 17: Traces of weights and minimum pressure under the consensus algorithm

We also note that the minimum pressure given by the consensus algorithm after 500 is about 0.3, while by Corollary 2 the all-one vector gives a minimum pressure of 0.25 since the interference degree of the tested network is 3.

Chapter 4

RANDOMIZED LOAD BALANCING

4.1 Background

In many computing and networking applications, including routing, hashing, and load balancing (see Mitzenmacher *et al.* (2001)), 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 Eryilmaz and Srikant (2012) 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 Mitzenmacher (1996, 2001); Vvedenskaya *et al.* (1996). 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 Bramson *et al.* (2012, 2013), the results have been extended to the case of heavy-tailed distributions, in Tsitsiklis and Xu (2012, 2013), the effect of resource pooling has been considered, and the case of heterogeneous servers operating under

the processor-sharing discipline has been treated in Mukhopadhyay and Mazumdar (2013).

In this chapter, 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 that 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 Ousterhout *et al.* (2013), 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, Ousterhout *et al.* (2013) 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 chapter, we observe that this basic algorithm for batch arrivals suggested in Ousterhout *et al.* (2013) 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-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* (see Ethier and Kurtz (2005)), 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 Ousterhout *et al.* (2013).

4.2 Problem Statement and Main Results

We consider a computing cluster with n identical servers and a central scheduler as shown in Figure 18. 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 $m = \Theta(\log n)$. For simplicity, we consider a deterministic batch size here, but the results in the chapter can be extended to random batch sizes as well in a straightforward manner. Furthermore, the results of this chapter 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

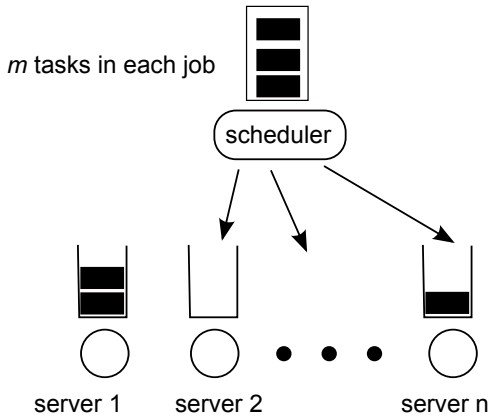


Figure 18: A computing cluster with n servers and a central scheduler

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** (see Mitzenmacher (1996); Vvedenskaya *et al.* (1996)): 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.
- **Batch-Sampling** (see Ousterhout *et al.* (2013)): 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 chapter, 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

	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 1: Summary of the expected per-task delays and the maximum queue sizes of the three scheduling algorithms

Note: 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 $d \geq 2$.

one to the least loaded server, where the queue length of a server is updated after it receives a task.

Remark 7. 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 chapter, d is called the *probe ratio*, which is assumed to be a constant independent of n. As in Mitzenmacher (1996); Vvedenskaya *et al.* (1996), 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 chapter are summarized in Table 1, 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 chapter.
- The size of the longest-queue in the system under the-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.

4.3 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 4.4. 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 $Q^{(n)}(t)$ is an irreducible and nonexplosive Markov chain, and using the standard Foster–Lyapunov theorem (see, for example, Srikant and Ying (2014)) it can be verified that the Markov chain is positive recurrent and hence, has a unique stationary distribution.

Theorem 5. *The Markov chain $Q^{(n)}(t)$ is positive recurrent under batch-filling. Furthermore, there exists a constant $c > 0$, independent of n, such that*

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

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

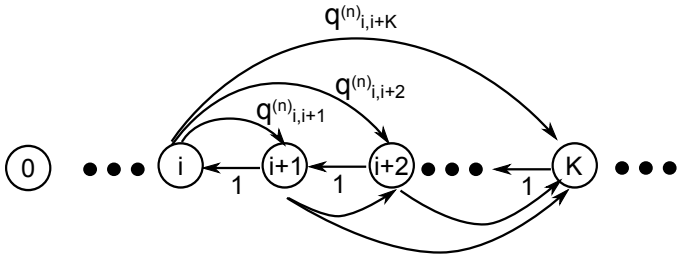


Figure 19: The Markov chain representing the nth system in the mean field analysis

The proof of this theorem is presented in Appendix C.1. 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 π. This i.i.d. assumption in the mean-field analysis will be validated later in Section 4.4 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 19, 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.

4.3.1 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

$$\begin{aligned} q_{i,j}^{(n)} &= \frac{n}{m}\lambda \times \frac{dm}{n} \times \sum_{\phi} \mathbb{P}(\phi) \times \mathbb{P}(j|\phi, i) \\ &= d\lambda \sum_{\phi} \mathbb{P}(\phi)\, \mathbb{P}(j|\phi, i). \end{aligned} \tag{4.1}$$

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;
- ϕ is a $(dm-1)$-vector that denotes the queue lengths of the other $dm-1$ sampled servers, so
$$\mathbb{P}(\phi) = \prod_{k=1}^{dm-1} \pi_{\phi_k};$$
and
- $\mathbb{P}(j|\phi, i)$ 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 ϕ.

Without loss of generality, assume $\phi_k \leq \phi_l$ if $k \leq l$, i.e., ϕ is ordered. Recall that batch-filling dispatches tasks using water filling among the sampled dm queues. Therefore, given i and ϕ, either $j = i$ if no task is assigned to the server, or j takes

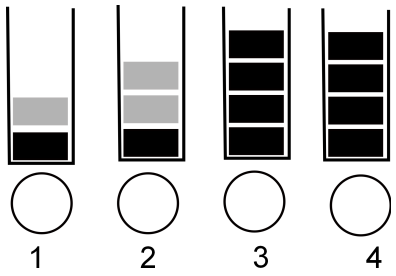


Figure 20: An example of water filling

two possible values. Consider a simple example in Figure 20 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.

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

- If
$$\sum_{k=1}^{dm-1} (i - \phi_k)\mathbb{I}_{\phi_k \leq i-1} \geq m, \tag{4.2}$$
which means that the tasks will be assigned to servers whose original queue sizes are smaller than i, then
$$\mathbb{P}(j|\phi, i) = \begin{cases} 1 & \text{if } j = i, \\ 0 & \text{if } j \neq i. \end{cases}$$

- If condition (4.2) does not hold, then the server with queue size i will receive some tasks, and
$$\mathbb{P}(j|\phi, i) = \begin{cases} 1 - \alpha_{\phi,i} & \text{if } j = \bar{Q}_{\phi,i} - 1, \\ \alpha_{\phi,i} & \text{if } j = \bar{Q}_{\phi,i}, \end{cases}$$

where

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

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

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

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

While the transition rate $q_{i,j}^{(n)}$ in (4.1) 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 Appendix C.2.

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

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

where for $j \neq i$,

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

$$\bar{Q}_\pi = \min\left\{ j \colon \sum_{l=0}^{j-1} (j-l)\pi_l \ge \frac{1}{d} \right\} \tag{4.3}$$

and

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

$\diamond$

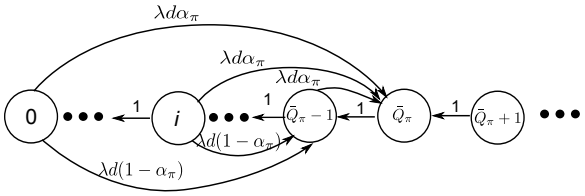


Figure 21: The queue-length Markov chain of a single-server, in the large-system limit, under batch-filing

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 21, where the up-crossing transitions are *into* only two states $\bar{Q}_\pi - 1$ and $\bar{Q}_\pi$ due to water filling. Based on Lemma 4, we can calculate the stationary distribution of the queue length of a single server in the large-system limit by finding $\hat{\pi}$ that satisfies the global balance equation (see, e.g., Srikant and Ying (2014)).

Theorem 6. *The stationary distribution of the queue length of a single server in the large-system limit under batch-filling is*

$$\hat{\pi}_i = \begin{cases} 1-\lambda & i = 0, \\ (1-\lambda)\lambda d(1+\lambda d)^{i-1} & 1 \le i \le \bar{Q}_{BF} - 1, \\ 1-(1-\lambda)(1+\lambda d)^{\bar{Q}_{BF}-1} & i = \bar{Q}_{BF}, \\ 0 & \textit{otherwise.} \end{cases} \tag{4.4}$$

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 $\bar{Q}_{BF} = \bar{Q}_{\hat{\pi}}$, where $\bar{Q}_\pi$ is defined in (4.3). Note that $\bar{Q}_{BF} \ge 1$. If

λ and d are such that $\bar{Q}_{BF} = 1$, then equivalently

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

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

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

If λ and d are such that $\bar{Q}_{BF} > 1$, according to (4.3), to show $\bar{Q}_{\hat{\pi}} = \bar{Q}_{BF}$ we only need to show

$$\sum_{l=0}^{\bar{Q}_{BF}-2} (\bar{Q}_{BF} - 1 - l)\hat{\pi}_l < \frac{1}{d} \leq \sum_{l=0}^{\bar{Q}_{BF}-1} (\bar{Q}_{BF} - l)\hat{\pi}_l. \tag{4.5}$$

Let LHS and RHS denote the left-hand-side and the right-hand-side of (4.5). 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 (4.5) is equivalent to

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

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

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

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

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

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

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

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

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

So the global balance equations hold.

Finally the expected queue length in stationary distribution is

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

□

4.3.2 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

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

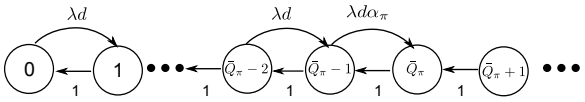


Figure 22: The Markov chain in the large-system limit under batch-sampling

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

Lemma 5. *Under batch-sampling, the transition rates given distribution π, denoted by $q_{i,j}^{(n)}(\pi)$ converge; and specifically,*

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

where

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

and

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

◇

The Markov chain in the large-system limit is shown in Figure 22. Given π, the Markov chain is a birth–death process up to state $\bar{Q}_\pi$. The stationary distribution can again be calculated using the global balance equations. The results are presented in Theorem 7, and the details are omitted.

Theorem 7. *The stationary distribution of the queue length of a single server in the large-system limit under batch-sampling is*

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

where

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

The expected queue length is

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

◇

4.3.3 The Stationary Distribution under the-Power-of-d-Choices

The stationary queue-length distribution of a single server in the large-system limit under the-power-of-d-choices has been established in Mitzenmacher (1996). We present the result below for comparison purposes.

Theorem 8. *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

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

◇

4.4 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 by following similar steps. We first consider the following non-linear system described by differential equations:

$$\frac{\mathrm{d}x_i}{\mathrm{d}t} = \begin{cases} -(1+\lambda d)x_i + x_{i+1} & i \le \bar{X}_x - 2, \\ \lambda d(1-\alpha_x)\sum_{j=0}^{i-1} x_j - (1+\lambda d\alpha_x)x_i + x_{i+1}, & i = \bar{X}_x - 1 \\ \lambda d\alpha_x \sum_{j=0}^{i} x_j - x_i + x_{i+1}, & i = \bar{X}_x \\ -x_i + x_{i+1} & \text{otherwise,} \end{cases} \tag{4.6}$$

where

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

and

$$\alpha_x = \frac{\frac{1}{d} - \sum_{j=0}^{\bar{X}_x - 2} (\bar{X}_x - 1 - j)x_j}{\sum_{j=0}^{\bar{X}_x - 1} x_j}.$$

These differential equations are derived from the Markov chain in Figure 21. View x_i as the fraction of queues with length i. Consider x_i for $i \le \bar{X}_x - 2$. According to Figure 21, 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 $\bar{X}_x - 1$ or $\bar{X}_x$ 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 α_x and $\bar{X}_x$ depend on the state x.

We further define

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

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

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

for $\hat{\pi}$ defined in (4.4). Note that $s_0(t) = 1$ for any t. The differential equations of the non-linear system can be written in terms of $s(t)$ as follows:

$$\frac{\mathrm{d}s_i}{\mathrm{d}t} = \begin{cases} \lambda d - (1+\lambda d)s_i + s_{i+1} & i \leq \bar{X}_s - 1, \\ \lambda - \lambda d \sum_{j=0}^{i-1}(1-s_j) - s_i + s_{i+1}, & i = \bar{X}_s \\ -s_i + s_{i+1} & \text{otherwise,} \end{cases} \tag{4.7}$$

where

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

The following theorem establishes the equilibrium point and the stability of this non-linear system. The proof is presented in Appendix C.3.

Theorem 9. *Assume the initial condition $s(0)$ satisfies $1 = s_1(0) \geq s_2(0) \geq \cdots \geq 0$ and (ii) $|s(0)| < \infty$. Starting from $s(0)$, the system converges to the equilibrium point $\hat{s}$ 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 $\pi^{(n)}(t)$ is a random vector instead of a distribution. Define

the vector $\Gamma^{(n)}(t) \in \mathbb{N}^\infty$ such that its ith 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, $\gamma^{(n)}(t) = \frac{\Gamma^{(n)}(t)}{n}$, and $\hat{\gamma}$ such that $\hat{\gamma}_i = \sum_{j=i}^\infty \hat{\pi}_j$ for $\hat{\pi}$ defined in (4.4).

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

$$1 = \gamma_0 \geq \gamma_1 \geq \cdots \geq 0 \tag{4.8}$$

with the 1-norm. The proof is presented in Appendix C.4.

Theorem 10. *Suppose that $\gamma^{(n)}(0) \to s(0)$ in probability, where $s(0)$ is a deterministic initial condition such that $s(0) \geq 0$ and $|s(0)| < \infty$. Then the following holds*

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

◇

This result is motivated by Kurtz's theorem (see Ethier and Kurtz (2005)). 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 $\gamma^{(n)}$ 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 $|s(0)| = \sum_i i x_i(0) < \infty$ 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 9 and Theorem 10 establish the following result:

$$\gamma^{(n)}(t) \xrightarrow{n\to\infty} s(t) \xrightarrow{t\to\infty} \hat{\gamma}, \tag{4.9}$$

which further implies that

$$\pi^{(n)}(t) \xrightarrow{n\to\infty} x(t) \xrightarrow{t\to\infty} \hat{\pi}. \tag{4.10}$$

A direct consequence of (4.10) is that if $\hat{\pi}^{(n)}$ converges to some $\tilde{\pi}$ or a subsequence of $\hat{\pi}^{(n)}$ converges to some $\tilde{\pi}$, then $\tilde{\pi} = \hat{\pi}$. The convergence of stationary distributions will be discussed in the next section.

4.5 Convergence of the Stationary Distributions

We first present a theorem on the interchange of limits. The theorem is similar to Theorem 5.1 in Anantharam and Benchekroun (1993). However, Anantharam and Benchekroun (1993) 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 chapter.

Theorem 11. *Consider a sequence of random processes $X^{(n)}$ indexed by a scaling parameter n, where $X^{(n)}$ is a vector that denotes value of the process at time t, and a dynamic system $\dot{X}(t) = F(X)$. Assume $X^{(n)}$ and $\hat{X}$ satisfy the following assumptions:*

- *(A1) Suppose that for any n,*

$$X^{(n)}(t) \xrightarrow{w} \hat{X}^{(n)}, \tag{4.11}$$

  where $\hat{X}^{(n)}$ is the stationary distribution of the random process and $\xrightarrow{w}$ denotes the weak convergence.

- *(A2) Suppose for each finite t,*

$$X^{(n)}(t) \xrightarrow{w} X(t), \tag{4.12}$$

  when

$$\lim_{n\to\infty} X^{(n)}(0) = X(0)$$

  where both $X^{(n)}(0)$ and $X(0)$ are deterministic initial conditions, and $X(0) \in \mathcal{X}$, where $\mathcal{X}$ is a set of initial conditions.

- *(A3) Starting from each initial condition $X(0) \in \mathcal{X}$, assume that*

$$\lim_{t\to\infty} X(t) = \hat{X}. \tag{4.13}$$

- *(A4) Any subsequence of $\hat{X}^{(n)}$ has a subsubsequence that weakly converges. The limit of any convergent subsequence, denoted by $\bar{X}$, satisfies $\mathbb{P}\left(\bar{X} \in \mathcal{X}\right) = 1$ and its support is separable.*

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

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

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

The theorem says that with additional assumptions, we further have

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

The proof is presented in Appendix C.5.

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

Theorem 12.

$$\hat{\gamma}^{(n)} \xrightarrow{w} \gamma.$$ ◇

Proof. Define

$$\mathcal{X} = \{\gamma : 1 = \gamma_0 \geq \gamma_1 \geq \gamma_2 \geq \cdots \geq 0, \sum_i \gamma_i < \infty\},$$

with metric

$$\rho(x, y) = \sum_{i=1}^{\infty} |x_i - y_i|.$$

Then $\mathcal{X}$ is separable because it is a subspace of the ℓ^1 space.

- (A1) holds due to Theorem 5.
- Note $\lim_{n\to\infty}\gamma^{(n)}(0) = s(0)$ for deterministic initial conditions $\gamma^{(n)}(0)$ and $s(0)$ implies that $\gamma^{(n)}(0) \to s(0)$ in probability. Therefore, according to Theorem 10, given deterministic initial conditions $\gamma^{(n)}(0)$ and $s(0)$ such that $\lim_{n\to\infty}\gamma^{(n)}(0) = s(0)$, we have

$$\lim_{n\to\infty}\sup_{0\le u\le t}|\gamma^{(n)}(u) - s(u)| = 0 \quad \text{in probability,}$$

  which implies weak convergence.
- (A3) is established in Theorem 9.
- To validate (A4), we consider the space

$$\tilde{\mathcal{U}} \triangleq \left\{\gamma = (\gamma_0, \gamma_1, \dots) \in [0,1]^\infty \colon 1 = \gamma_0 \ge \gamma_1 \ge \cdots \ge 0\right\}$$

  with the metric used in Vvedenskaya *et al.* (1996)

$$\rho'(\gamma, \gamma') = \sup_{i>0}\frac{|\gamma_i - \gamma_i'|}{i}.$$

  Then $(\tilde{\mathcal{U}}, \rho')$ is a compact metric space. Let $\mu^n \in \mathcal{P}(\tilde{\mathcal{U}})$ be the stationary distribution of the nth system, where $\mathcal{P}(\tilde{\mathcal{U}})$ is the set of probability measures on $\tilde{\mathcal{U}}$ with its Borel sets as the σ-algebra. Then the sequence (μ^n) is tight on $\tilde{\mathcal{U}}$. By Prokhorov's theorem (see Billingsley (1999)), any subsequence of (μ^n) has a subsubsequence that weakly converges in $\tilde{\mathcal{U}}$. Suppose the convergent subsubsequence is $(\mu^{n_k} \colon k)$ and its limit is $\mu^\infty \in \mathcal{P}(\tilde{\mathcal{U}})$. By Skorokhod's representation theorem, there exists a sequence of random elements with the same distributions that converge almost surely. By slight abuse of the notation, we assume $\hat{\gamma}^{(n_k)}$ converges to $\tilde{\gamma}$ almost surely in $(\tilde{\mathcal{U}}, \rho')$, where $\mathcal{L}(\hat{\gamma}^{(n_k)}) = \mu^{(n_k)}$ and $\mathcal{L}(\tilde{\gamma}) = \mu^\infty$. Since $0 \le \gamma_i \le 1$, we have by the dominated convergence theorem

$$\lim_{k\to\infty}\mathbb{E}[|\hat{\gamma}_i^{(n_k)} - \tilde{\gamma}_i|] = 0 \quad \forall i. \tag{4.14}$$

Consider

$$f_l(\gamma) = \sum_{i=1}^{l} \gamma_i.$$

Then f_l is continuous and bounded on $(\tilde{\mathcal{U}}, \rho')$. By the definition of weak convergence, we have

$$\lim_{k\to\infty} \mathbb{E}\, f_l(\hat{\gamma}^{(n_k)}) = \mathbb{E}\, f_l(\tilde{\gamma}).$$

Recall by Theorem 5, for any k and l,

$$\mathbb{E}\, f_l(\hat{\gamma}^{(n_k)}) \leq c.$$

We have by Tonelli's theorem,

$$\mathbb{E}\left(\sum_{i=1}^{\infty} \tilde{\gamma}_i\right) = \sum_{i=1}^{\infty} \mathbb{E}\, \tilde{\gamma}_i \leq c < \infty.$$

Consequently $\mathbb{P}(\tilde{\gamma} \in \mathcal{X}) = 1$, or equivalently $\text{supp}(\mu^\infty) \subseteq \mathcal{X}$. The following uniform convergence is established in Appendix C.6.

Lemma 6. *The series* $\sum_{i=1}^{\infty} \mathbb{E}\left|\hat{\gamma}_i^{(n_k)} - \tilde{\gamma}_i\right|$ *are uniformly convergent for all* k. $\diamond$

Then we get

$$\lim_{k\to\infty} \sum_{i=1}^{\infty} \mathbb{E}\left|\hat{\gamma}_i^{(n_k)} - \tilde{\gamma}_i\right| = \sum_{i=1}^{\infty} \lim_{k\to\infty} \mathbb{E}\left|\hat{\gamma}_i^{(n_k)} - \tilde{\gamma}_i\right| = 0. \tag{4.15}$$

So (μ^n) is tight in $(\mathcal{X}, \rho)$ and (A4) is verified.

□

Based on the theorem above, we further have the following results according to using the same analysis for getting (4.14) and (4.15).

Corollary 4.

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

$$\lim_{n\to\infty} \mathbb{E}\left[\sum_i \hat{\gamma}_i^{(n)}\right] = \sum_i \hat{\gamma}_i, \tag{4.16}$$

and

$$\lim_{n\to\infty} \mathbb{E}\left[|\hat{\gamma}^{(n)} - \hat{\gamma}|\right] = 0. \tag{4.17}$$

In the next corollary, we show that any k queues are independently and identically distributed with distribution $\hat{\pi}$ in the large-system limit, where k is a constant independent of n. Then the system is said to be $\hat{\pi}$-chaotic (see Sznitman (1991)). 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. The proof is in Appendix C.7.

Corollary 5. *Consider a set of k servers, and without loss of generality, assume the servers are $1, 2, \ldots, k$. Let $\pi^{(n)}(Q_1, Q_2, \ldots, Q_k)$ denote the stationary distribution of the queue lengths of these k servers. In the large-system limit, we have*

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

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

4.6 Simulations

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

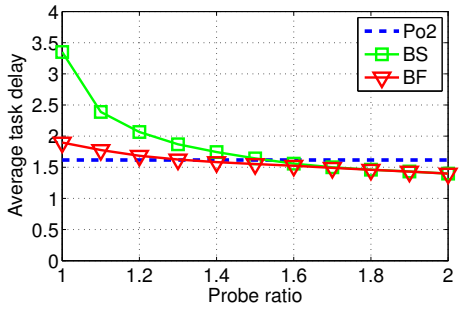


Figure 23: The average task delays for power-of-two-choices (Po2), batch-sampling (BS) and batch-filling (BF) with $\lambda = 0.7$ and deterministic batch sizes

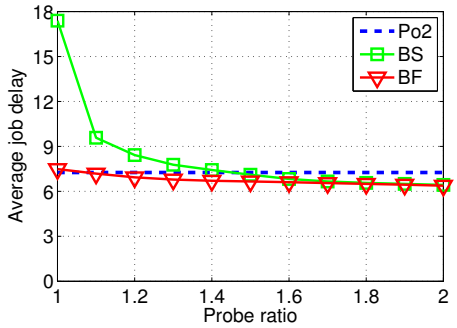


Figure 24: The average job delays for power-of-two-choices (Po2), batch-sampling (BS) and batch-filling (BF) with $\lambda = 0.7$ and deterministic batch sizes

4.6.1 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 23 and 24 show the per-task delays and per-job delays, respectively, when $\lambda = 0.7$. Figures 25 and 26 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

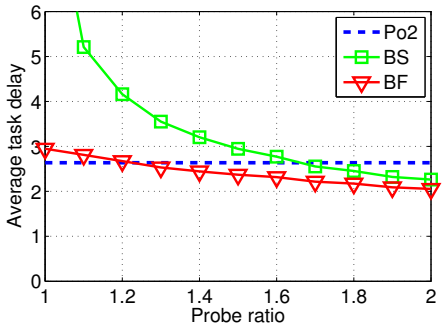


Figure 25: The average task delays for power-of-two-choices (Po2), batch-sampling (BS) and batch-filling (BF) with $\lambda = 0.9$ and deterministic batch sizes

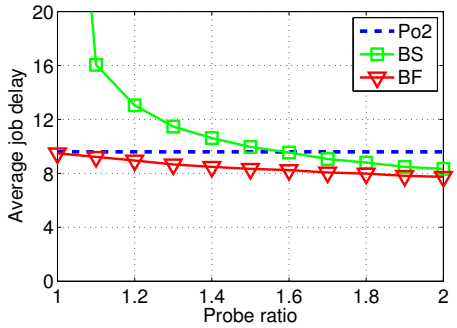


Figure 26: The average job delays for power-of-two-choices (Po2), batch-sampling (BS) and batch-filling (BF) with $\lambda = 0.9$ and deterministic batch sizes

$d = 1$ (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 24 and 26, *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 1.

4.6.2 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

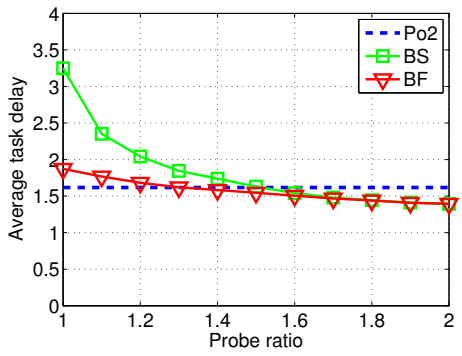


Figure 27: The average task delays for power-of-two-choices (Po2), batch-sampling (BS) and batch-filling (BF) with $\lambda = 0.7$ with random batch sizes

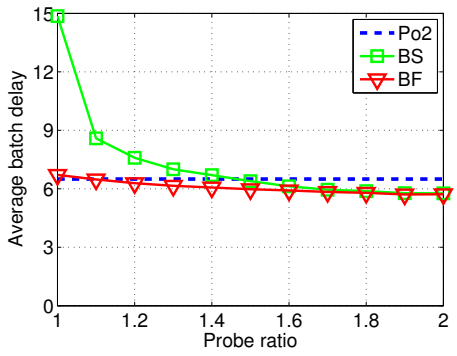


Figure 28: The average job delays for power-of-two-choices (Po2), batch-sampling (BS) and batch-filling (BF) with $\lambda = 0.7$ with random batch sizes

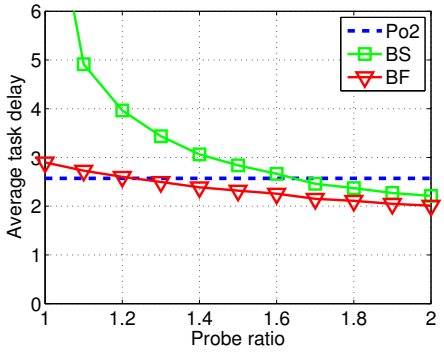


Figure 29: The average task delays for power-of-two-choices (Po2), batch-sampling (BS) and batch-filling (BF) with $\lambda = 0.9$ with random batch sizes

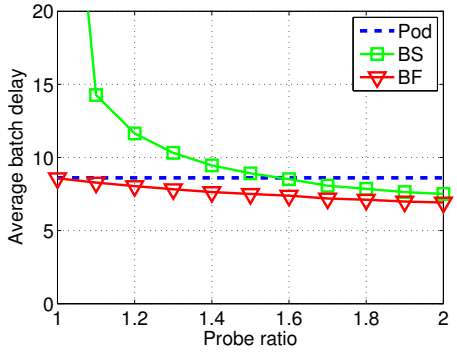


Figure 30: The average job delays for power-of-two-choices (Po2), batch-sampling (BS) and batch-filling (BF) with $\lambda = 0.9$ with random batch sizes

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 27 and Figure 28; and the results for $\lambda = 0.9$ are shown in Figures 29 and 30. We note that the conclusions of our previous simulations do not change with these modifications.

Chapter 5

CONCLUSION

In Chapter 2 we studied the stability of the longest-queue-first scheduling policy in wireless networks with multihop traffic flows and the one-hop interference model. Using fluid techniques, we proved that LQF is throughput optimal in this scenario. The proof itself is an interesting contribution and can be useful when considering similar fluid systems since we focused on state transition instead of an explicit Lyapunov function. The result may also be a first step to understand the stability performance of LQF in general networks with multihop traffic flows.

In Chapter 3 we considered the problem of scheduling real-time traffic in wireless networks under general stochastic arrivals and deadlines and general interference model. The fraction of delivered packets at a link is required to be no less than a certain threshold. We used deficits to inspect the stability of the system, and studied the stability performance of a scheduling policy that we call the largest-deficit-first (LDF) policy. We proved that the efficiency ratio of LDF can be lower bounded by a quantity that we call the real-time local-pooling factor (R-LPF). Furthermore, we showed lower bounds on the R-LPF can be calculated by assigning weights to the links, with a special case lower bound of $1/(\beta+1)$, where β is the interference degree. We also proposed a heuristic consensus algorithm that can be used to estimate the R-LPF for general networks.

In Chapter 4 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 Chapter 4 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 Chapter 4 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 5 can be established by using the same idea that the Lyapunov drift of water-filling is dominated by random routing. Lemma 4 also holds because $\frac{M^{(n)}(0)}{\mathbb{E}[M^{(n)}(0)]}$ converge in probability. The differential equations remain the same under random batch size, so Theorem 9 is still valid. Finally, it is easy to verify that $D_i/(dm)$ converges in mean as $m \to \infty$, where $m = \mathbb{E}[M^{(n)}(0)]$.

REFERENCES

Anantharam, V. and M. Benchekroun, "A technique for computing sojourn times in large networks of interacting queues", Prob. Eng. and Informational Sci. **7**, 441–464 (1993).

Andrews, M., K. Kumaran, K. Ramanan, A. Stolyar, R. Vijayakumar and P. Whiting, "Scheduling in a queuing system with asynchronously varying service rates", Prob. Eng. and Informational Sci. **18**, 2, 191–217 (2004).

Billingsley, P., *Convergence of Probability Measures*, Wiley series in probability and statistics (John Wiley & Sons, Inc., 1999), second edn.

Birand, B., M. Chudnovsky, B. Ries, P. Seymour, G. Zussman and Y. Zwols, "Analyzing the performance of greedy maximal scheduling via local pooling and graph theory", IEEE/ACM Trans. Netw. **20**, 1, 163–176 (2012).

Boyd, S. and L. Vandenberghe, *Convex Optimization* (Cambridge Univ. Press, New York, NY, 2004).

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

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

Brzezinski, A., G. Zussman and E. Modiano, "Local pooling conditions for joint routing and scheduling", in "Proc. Inf. Theory and Appl. Workshop (ITA)", pp. 499–506 (San Diego, CA, 2008).

Bui, L. X., R. Srikant and A. Stolyar, "A novel architecture for reduction of delay and queueing structure complexity in the back-pressure algorithm", IEEE/ACM Trans. Netw. **19**, 6, 1597–1609 (2011).

Chaporkar, P., K. Kar and S. Sarkar, "Throughput guarantees through maximal scheduling in wireless networks", in "Proc. Annu. Allerton Conf. Communication, Control and Computing", pp. 28–30 (Monticello, IL, 2005).

Chen, H. and D. D. Yao, *Fundamentals of Queueing Networks: Performance, Asymptotics, and Optimization*, vol. 46 of *Stochastic Modelling and Applied Probability* (Springer, New York, NY, 2001).

Chung, K. L., *A Course in Probability Theory* (Academic Press, San Diego, CA, 2001), third edn.

Dai, J. G., "On positive Harris recurrence of multiclass queueing networks: A unified approach via fluid limit models", Ann. Appl. Probab. **5**, 1, 49–77 (1995).

Dai, J. G. and B. Prabhakar, "The throughput of data switches with and without speedup", in "Proc. IEEE Int. Conf. Computer Communications (INFOCOM)", vol. 2, pp. 556–564 (Tel Aviv, Israel, 2000).

Dimakis, A. and J. Walrand, "Sufficient conditions for stability of longest queue first scheduling: Second-order properties using fluid limits", Adv. Appl. Probab. **38**, 2, 505–521 (2006).

Draief, M. and L. Massoulié, *Epidemics and Rumours in Complex Networks*, London Mathematical Society Lecture Note Series (Cambridge Univ. Press, 2010), first edn.

Eryilmaz, A. and R. Srikant, "Asymptotically tight steady-state queue length bounds implied by drift conditions", Queueing Syst. **72**, 3, 311–359 (2012).

Ethier, S. N. and T. G. Kurtz, *Markov Processes: Characterization and Convergence* (John Wiley & Sons, Inc., Hoboken, NJ, 2005), second edn.

Hellings, T., S. C. Borst and J. S. van Leeuwaarden, "Tandem queueing networks with neighbor blocking and back-offs", Queueing Syst. **68**, 3–4, 321–331 (2011).

Hoeffding, W., "Probability inequalities for sums of bounded random variables", J. Am. Stat. Assoc. **58**, 301, 13–30 (1963).

Hou, I.-H., V. Borkar and P. R. Kumar, "A theory of QoS for wireless", in "Proc. IEEE Int. Conf. Computer Communications (INFOCOM)", pp. 486–494 (Rio de Janeiro, Brazil, 2009).

Hou, I.-H. and P. R. Kumar, "Admission control and scheduling for QoS guarantees for variable-bit-rate applications on wireless channels", in "Proc. ACM Int. Symp. Mobile Ad Hoc Networking and Computing (MobiHoc)", pp. 175–184 (New Orleans, LA, 2009).

Hou, I.-H. and P. R. Kumar, "Scheduling heterogeneous real-time traffic over fading wireless channels", in "Proc. IEEE Int. Conf. Computer Communications (INFOCOM)", pp. 1–9 (San Diego, CA, 2010a).

Hou, I.-H. and P. R. Kumar, "Utility-optimal scheduling in time-varying wireless networks with delay constraints", in "Proc. ACM Int. Symp. Mobile Ad Hoc Networking and Computing (MobiHoc)", pp. 31–40 (Chicago, IL, 2010b).

Jain, K., J. Padhye, V. N. Padmanabhan and L. Qiu, "Impact of interference on multi-hop wireless network performance", in "Proc. ACM Int. Conf. Mobile Computing and Networking (MobiCom)", pp. 66–80 (San Diego, CA, 2003).

Jaramillo, J. J. and R. Srikant, "Optimal scheduling for fair resource allocation in ad hoc networks with elastic and inelastic traffic", IEEE/ACM Trans. Netw. **19**, 4, 1125–1136 (2011).

Jaramillo, J. J., R. Srikant and L. Ying, "Scheduling for optimal rate allocation in ad hoc networks with heterogeneous delay constraints", IEEE J. Sel. Areas Commun. **29**, 5, 979–987 (2011).

Ji, B., C. Joo and N. Shroff, "Throughput-optimal scheduling in multihop wireless networks without per-flow information", IEEE/ACM Trans. Netw. **21**, 634–647 (2013).

Joo, C., X. Lin and N. B. Shroff, "Performance limits of greedy maximal matching in multi-hop wireless networks", in "Proc. IEEE Conf. Decision and Control (CDC)", pp. 1128–1133 (New Orleans, LA, 2007).

Joo, C., X. Lin and N. B. Shroff, "Greedy maximal matching: Performance limits for arbitrary network graphs under the node-exclusive interference model", IEEE Trans. Autom. Control **54**, 2734–2744 (2009a).

Joo, C., X. Lin and N. B. Shroff, "Understanding the capacity region of the greedy maximal scheduling algorithm in multihop wireless networks", IEEE/ACM Trans. Netw. **17**, 4, 1132–1145 (2009b).

Khalil, H. K., *Nonlinear Systems* (Prentice Hall, 2001), third edn.

Leconte, M., J. Ni and R. Srikant, "Improved bounds on the throughput efficiency of greedy maximal scheduling in wireless networks", IEEE/ACM Trans. Netw. **19**, 709–720 (2011).

Li, B., C. Boyaci and Y. Xia, "A refined performance characterization of longest-queue-first policy in wireless networks", IEEE/ACM Trans. Netw. **19**, 1382–1395 (2011).

Li, B., C. Boyaci and Y. Xia, "Performance guarantee under longest-queue-first schedule in wireless networks", IEEE Trans. Inf. Theory **58**, 5878–5889 (2012).

Lin, X. and N. B. Shroff, "The impact of imperfect scheduling on cross-layer rate control in wireless networks", in "Proc. IEEE Int. Conf. Computer Communications (INFOCOM)", vol. 3, pp. 1804–1814 (Miami, FL, 2005).

Loynes, R. M., "The stability of a queue with non-independent inter-arrival and service times", Math. Proc. Cambridge Philos. Soc. **58**, 3, 497–520 (1962).

Maguluri, S., B. Hajek and R. Srikant, "The stability of longest-queue-first scheduling with variable packet sizes", IEEE Trans. Autom. Control **59**, 2295–2300 (2014).

Maguluri, S. T., B. Hajek and R. Srikant, "The stability of longest-queue-first scheduling with variable packet sizes", in "Proc. IEEE Conf. Decision and Control (CDC)", pp. 3770–3775 (Orlando, FL, 2011).

Mitzenmacher, M., "The power of two choices in randomized load balancing", IEEE Trans. Parallel Distrib. Syst. **12**, 1094–1104 (2001).

Mitzenmacher, M., A. W. Richa and R. Sitaraman, "The power of two random choices: A survey of techniques and results", in "Handbook of randomized computing", edited by S. Rajasekaran, vol. 1, chap. 9, pp. 255–312 (Kluwer Academic Publishers, 2001).

Mitzenmacher, M. D., *The Power of Two Choices in Randomized Load Balancing*, Ph.D. thesis, University of California at Berkeley (1996).

Mukhopadhyay, A. and R. R. Mazumdar, "Analysis of load balancing in large heterogeneous processor sharing systems", arXiv e-prints: 1311.5806 (2013).

Ni, J., B. Tan and R. Srikant, "Q-CSMA: Queue-length-based CSMA/CA algorithms for achieving maximum throughput and low delay in wireless networks", IEEE/ACM Trans. Netw. **20**, 825–836 (2012).

Ousterhout, K., P. Wendell, M. Zaharia and I. Stoica, "Sparrow: Distributed, low latency scheduling", in "Proc. ACM Symp. Oper. Syst. Princ. (SOSP)", pp. 69–84 (Farminton, PA, 2013).

Reddy, A. A., S. Sanghavi and S. Shakkottai, "On the effect of channel fading on greedy scheduling", in "Proc. IEEE Int. Conf. Computer Communications (INFOCOM)", pp. 406–414 (Orlando, FL, 2012).

Royden, H. L. and P. M. Fitzpatrick, *Real Analysis* (Prentice Hall, 2010), fourth edn.

Shah, D. and D. Wischik, "Switched networks with maximum weight policies: Fluid approximation and multiplicative state space collapse", Ann. Appl. Probab. **22**, 1, 70–127 (2012).

Srikant, R. and L. Ying, *Communication Networks: An Optimization, Control and Stochastic Networks Perspective* (Cambridge Univ. Press, New York, 2014).

Stolyar, A. L., "Large number of queues in tandem: Scaling properties under back-pressure algorithm", Queueing Syst. **67**, 2, 111–126 (2011).

Sznitman, A.-S., "Topics in propagation of chaos", in "Ecole d'Eté de Probabilités de Saint-Flour XIX – 1989", edited by P.-L. Hennequin, vol. 1464 of *Lecture Notes in Mathematics*, pp. 165–251 (Springer Berlin Heidelberg, 1991).

Tassiulas, L. and A. Ephremides, "Stability properties of constrained queueing systems and scheduling policies for maximum throughput in multihop radio networks", IEEE Trans. Autom. Control **37**, 1936–1948 (1992).

Tassiulas, L. and A. Ephremides, "Dynamic scheduling for minimum delay in tandem and parallel constrained queueing models", Ann. Oper. Res. **48**, 4, 333–355 (1994).

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

Tsitsiklis, J. N. and K. Xu, "Queueing system topologies with limited flexibility", in "Proc. Annu. ACM SIGMETRICS Conf.", pp. 167–178 (Pittsburgh, PA, 2013).

Vvedenskaya, N. D., 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).

Weller, T. and B. Hajek, "Scheduling nonuniform traffic in a packet-switching system with small propagation delay", IEEE/ACM Trans. Netw. **5**, 813–823 (1997).

Wu, X. and R. Srikant, "Regulated maximal matching: A distributed scheduling algorithm for multi-hop wireless networks with node-exclusive spectrum sharing", in "Proc. IEEE Conf. Decision and Control (CDC)", pp. 5342–5347 (Seville, Spain, 2005).

APPENDIX A

PROOFS FOR CHAPTER 2

A.1 Proof of Proposition 1

First notice $\left|\bar{\mathbb{X}}^{x_k}(0)\right| = \left|\frac{x_k}{|x_k|}\right| = 1$ and $\left\{x \in \mathcal{X} \colon |x| = 1\right\}$ is compact, so there exists a subsequence $\left(k_p^{(1)}\right)$ such that $\bar{\mathbb{X}}^{x_{k_p^{(1)}}}(0) \to \bar{\mathbb{X}}(0)$ as $p \to \infty$.

By (2.4) we know for any j, $T_j^{x_{k_p^{(1)}}}(t)$ is Lipschitz continuous with Lipschitz constant 1 and then $\bar{T}_j^{x_{k_p^{(1)}}}(t)$ is also Lipschitz with Lipschitz constant 1. Then the sequence of functions $\left(\bar{T}_j^{x_{k_p^{(1)}}}(t)\right)$ is uniformly bounded and equicontinuous on the interval $[0,1]$ and by the Arzelà–Ascoli theorem there exists a subsequence $\left(k_{1,p}^{(1)}\right)$ of $\left(k_p^{(1)}\right)$ such that $\left(\bar{T}_j^{x_{k_{1,p}^{(1)}}}(t)\right)$ converges on $[0,1]$ uniformly as $p \to \infty$. Then for the interval $[0,2]$ there exists a subsequence $\left(k_{2,p}^{(1)}\right)$ of $\left(k_{1,p}^{(1)}\right)$ such that $\left(\bar{T}_j^{x_{k_{2,p}^{(1)}}}(t)\right)$ converges on $[0,2]$ uniformly as $p \to \infty$. By induction, for any positive integer q, we can find a subsequence $\left(k_{q,p}^{(1)} \colon p\right)$ such that $\left(\bar{T}_j^{x_{k_{q,p}^{(1)}}}(t) \colon p\right)$ converges on $[0,q]$ uniformly as $p \to \infty$. We take the diagonal subsequence $\left(k_p^{(2)} \colon p\right)$ by $k_p^{(2)} = k_{p,p}^{(1)}$ and then $\left(\bar{T}_j^{x_{k_p^{(2)}}}(t) \colon p\right)$ converges u.o.c. as $p \to \infty$. In the same way we can find a subsequence $\left(k_p\right)$ such that for any $j = 1, 2, \ldots, r$,

$$\bar{T}_j^{x_{k_p}}(t) \to \bar{T}_j(t) \quad \text{u.o.c. as } p \to \infty.$$

Similarly, Y_i and D_i are Lipschitz with constant 1, so we can find $\left(k_p\right)$ such that for all $i = 1, 2, \ldots, N$,

$$\bar{Y}_i^{x_{k_p}}(t) \to \bar{Y}_i(t) \quad \text{u.o.c. as } p \to \infty$$
$$\bar{D}_i^{x_{k_p}}(t) \to \bar{D}_i(t) \quad \text{u.o.c. as } p \to \infty.$$

The exogenous arrivals satisfy SLLN, so we may assume for the sample path ω and all $i = 1, 2, \ldots, N$,

$$\frac{1}{|x_{k_p}|} E_i(\lfloor |x_{k_p}| t \rfloor) \to \alpha_i t \quad \text{u.o.c. as } p \to \infty.$$

Then by (2.2) and (2.3),

$$\bar{A}_i^{x_{k_p}}(t) \to \bar{A}_i(t) \quad \text{u.o.c. as } p \to \infty$$

$$\bar{Z}_i^{x_{k_p}}(t) \to \bar{Z}_i(t) \quad \text{u.o.c. as } p \to \infty.$$

Then (2.9), (2.10), (2.11), (2.12), (2.13) and (2.15) readily come from (2.7), (2.2), (2.3), (2.4), (2.5) and (2.8).

Notice that (2.14) is equivalent to the following: Whenever $\bar{Z}_i(t) > 0$, there exists $\delta > 0$ such that $\bar{Y}_i(t') = \bar{Y}_i(t)$ for any $t' \in [t, t+\delta]$. To show this is true, we use a technique from Dai and Prabhakar (2000). We consider a time $t \geq 0$ and suppose $\bar{Z}_i(t) > 0$. Then by continuity there exists $\delta > 0$ such that $\min_{t' \in [t,t+\delta]} \bar{Z}_i(t') > 0$. Set $a = \min_{t' \in [t,t+\delta]} \bar{Z}_i(t')$. Thus by uniform continuity, there exists $K \geq 0$ such that for any $p \geq K$,

$$Z_i^{x_{k_p}}(t') \geq a/2 \quad \forall t' \in [t, t+\delta].$$

Then

$$Z_i^{x_{k_p}}(|x_{k_p}|t') \geq 1 \quad \forall t' \in [t, t+\delta].$$

That is, all systems in the subsequence $\left(k_p\right)$ have nonempty queue at link l_i during a period of time slots. By the work-conserving property in (2.6), the cumulative idle time of link l_i can increase by at most 1 (possibly because the queue is emptied at the end of the period of time slots); i.e.,

$$0 \leq Y_i^{x_{k_p}}(|x_{k_p}|t') - Y_i^{x_{k_p}}(|x_{k_p}|t) \leq 1 \quad \forall t' \in [t, t+\delta]$$

$$\Rightarrow 0 \leq \bar{Y}_i^{x_{k_p}}(t') - \bar{Y}_i^{x_{k_p}}(t) \leq \frac{1}{|x_{k_p}|} \quad \forall t' \in [t, t+\delta].$$

Then as $p \to \infty$,

$$\bar{Y}_i(t') = \bar{Y}_i(t) \quad \forall t' \in [t, t+\delta]$$

so we have (2.14).

Note that by repeatedly taking subsequences we can find $\left(k_p\right)$ such that all the convergences aforementioned hold at the same time. All components of $\bar{\mathbb{X}}$ are absolutely continuous because they are Lipschitz continuous. The monotonicity of A, D, T, Y implies the monotonicity of $\bar{A}, \bar{D}, \bar{T}, \bar{Y}$.

A.2 Proof of Lemma 1

We first notice that if $\bar{Z} \in B$, then there are no adjacent dominating nodes; i.e., if $\bar{Z} \in B_i$ for some $i \in \{1, 2, \ldots, N\}$, then $\bar{Z} \notin B_{i-1}$ and $\bar{Z} \notin B_{i+1}$. Let the dominating set at time t be

$$I_{\text{dom}}(t) = \left\{ i \in \{1, 2, \ldots, N\} \colon \bar{Z}(t) \in B_i \right\}.$$

Then we can easily check that $I_{\text{dom}}(t) \subseteq \bigcap \text{LQF}(\bar{Z}(t))$; i.e., all dominating links must be scheduled by LQF at time t. Due to the one-hop interference, there is no internal arrival to a scheduled link. Then for regular time t and any $i \in I_{\text{dom}}(t)$,

$$\frac{\mathrm{d}A_i}{\mathrm{d}t}(t) = \frac{\mathrm{d}E_i}{\mathrm{d}t}(t) = \alpha_i$$

and

$$\frac{\mathrm{d}D_i}{\mathrm{d}t}(t) = 1,$$

while

$$\frac{\mathrm{d}D_{i-1}}{\mathrm{d}t}(t) = \frac{\mathrm{d}D_{i+1}}{\mathrm{d}t}(t) = 0.$$

Thus,

$$\frac{\mathrm{d}\bar{Z}_i}{\mathrm{d}t}(t) - \frac{\mathrm{d}\bar{Z}_{i-1}}{\mathrm{d}t}(t) \leq \alpha_i - 1 \tag{A.1}$$

and

$$\frac{\mathrm{d}\bar{Z}_i}{\mathrm{d}t}(t) - \frac{\mathrm{d}\bar{Z}_{i+1}}{\mathrm{d}t}(t) \leq \alpha_i - 1. \tag{A.2}$$

Let $\bar{Z}_{\max} = \max_j \bar{Z}_j(0)$. We now present two propositions to complete the proof of Lemma 1.

Proposition 4. *There exists $t_1 \in (t_0, t_0 + \bar{Z}_{\max}/(1-\alpha_i)]$ such that $\bar{Z}(t_1) \notin B_i$; i.e., link l_i is not dominating anymore at some time before $t_0 + \bar{Z}_{\max}/(1-\alpha_i)$.* ◇

Proof. Indeed, if link l_i remains dominating up to (and including) time $t_0 + \bar{Z}_{\max}/(1-\alpha_i)$, then by (A.1), (A.2) and the absolute continuity of the fluids, we would have for any adjacent link l_j of l_i,

$$\left[\bar{Z}_i\left(t_0 + \frac{\bar{Z}_{\max}}{1-\alpha_i}\right) - \bar{Z}_j\left(t_0 + \frac{\bar{Z}_{\max}}{1-\alpha_i}\right)\right] - \left[\bar{Z}_i(t_0) - \bar{Z}_j(t_0)\right] \leq -\bar{Z}_{\max}.$$

Hence

$$\bar{Z}_i\left(t_0 + \frac{\bar{Z}_{\max}}{1-\alpha_i}\right) - \bar{Z}_j\left(t_0 + \frac{\bar{Z}_{\max}}{1-\alpha_i}\right) \leq \bar{Z}_i(t_0) - \bar{Z}_{\max} \leq 0.$$

Then by continuity, there is some $t_1 \in (t_0, t_0 + \bar{Z}_{\max}/(1-\alpha_i)]$ such that $\bar{Z}_i(t_1) - \bar{Z}_j(t_1) = 0$, which contradicts our assumption that link l_i remains dominating up to $t_0 + \bar{Z}_{\max}/(1-\alpha_i)$. This completes the proof of the proposition. □

Therefore, for any $i \in I_{\mathrm{dom}}(t_0)$, there exists $t_1 \in (t_0, t_0 + \bar{Z}_{\max}/(1-\max_j \alpha_j)]$ such that $\bar{Z}(t_1) \notin B_i$.

Proposition 5. *If $\bar{Z}(t_1) \notin B_i$, then $\bar{Z}(t) \notin B_i$ for any $t \geq t_1$.* ◇

Proof. Indeed, if $\bar{Z}(t_2) \in B_i$ for some $t_2 > t_1$, let $t_3 = \sup\left\{t < t_2 \colon \bar{Z}(t) \notin B_i\right\}$. Then by Lipschitz continuity $\bar{Z}_i(t_3) = \bar{Z}_j(t_3)$ for some neighbor l_j of link l_i and $t_3 < t_2$. Since $\frac{\mathrm{d}}{\mathrm{d}t}(\bar{Z}_i(t) - \bar{Z}_j(t)) \leq \max_k \alpha_k - 1$ for almost all $t \in [t_3, t_2]$, we have

$$\bar{Z}_i(t_2) - \bar{Z}_j(t_2) \leq \bar{Z}_i(t_3) - \bar{Z}_j(t_3) + (t_2 - t_3)(\max_k \alpha_k - 1) \leq 0,$$

which contradicts the assumption that $\bar{Z}(t_2) \in B_i$. Hence, $\bar{Z}(t) \notin B_i$ for any $t \geq t_1$. □

Considering all i, we have $\bar{Z}(t) \notin B$ for any $t \geq t_0 + \bar{Z}_{\max}/(1-\max_j \alpha_j)$.

A.3 Proof of Proposition 2

To show this proposition by contradiction, we suppose $J(t) = \emptyset$.

We first notice that for $j_1 = \min\{j\colon j \in J_0(t)\}$ we must have $\frac{\mathrm{d}}{\mathrm{d}t}\bar{W}_{j_1}(t) > 0$. If this is not the case, we would have $\frac{\mathrm{d}}{\mathrm{d}t}\bar{W}_{j_1}(t) < 0$ and $j_1 \geq 2$. Then it would follow that there exists some $\delta > 0$ such that for any $s \in (t, t+\delta)$ we have $\bar{Z}_1(s) > \bar{Z}_2(s)$ if $j_1 = 2$, and $\bar{Z}_{j_1-1}(s) > \max\{\bar{Z}_{j_1-2}(s), \bar{Z}_{j_1}(s)\}$ if $j_1 > 2$, which implies $\bar{Z}(s) \in B$, a contradiction.

We then conclude that if all $j \in \{1, 2, \ldots, k\} \cap J_0(t)$ satisfy

$$\frac{\mathrm{d}}{\mathrm{d}t}\bar{W}_j(t) > 0,$$

then either $k+1 \notin J_0(t)$ or

$$\frac{\mathrm{d}}{\mathrm{d}t}\bar{W}_{k+1}(t) > 0.$$

If this is not the case, there would exist $\delta > 0$ such that for any $s \in (t, t+\delta)$, we have $\bar{Z}_k(s) > \max\{\bar{Z}_{k-1}(s), \bar{Z}_{k+1}(s)\}$ if $\bar{W}_k(t) \geq 0$, or $\bar{Z}_j(s) > \max\{\bar{Z}_{j-1}(s), \bar{Z}_{j+1}(s)\}$ for some $j < k$ otherwise, either of which leads to a contradiction.

By induction we have $\frac{\mathrm{d}}{\mathrm{d}t}\bar{W}_j(t) > 0$ for all $j \in J_0(t)$, which also leads to contradiction since by letting $j_2 = \max\{j\colon j \in J_0(t)\}$ there exists $\delta > 0$ such that for any $s \in (t, t+\delta)$ we have $\bar{Z}_N(s) > \bar{Z}_{N-1}(s)$ if $j_2 = N$, and $\bar{Z}_{j_2}(s) > \max\{\bar{Z}_{j_2-1}(s), \bar{Z}_{j_2+1}(s)\}$ if $j_2 \neq N$. Then $\bar{Z}(s) \in B$, which is a contradiction. This completes the proof of Proposition 2.

A.4 Proof of Proposition 3

Note that $u \geq 2$. By definition of u, $\bar{W}_u(t_1) = \frac{\mathrm{d}}{\mathrm{d}t}\bar{W}_u(t_1) = 0$, i.e., $\bar{Z}_u(t_1) = \bar{Z}_{u-1}(t_1)$ and $\frac{\mathrm{d}}{\mathrm{d}t}\bar{Z}_u(t_1) = \frac{\mathrm{d}}{\mathrm{d}t}\bar{Z}_{u-1}(t_1)$.

We first claim that there exists $\delta > 0$ such that for any $t \in (t_1, t_1+\delta)$,

$$0 < \bar{Z}_1(t) < \bar{Z}_2(t) < \cdots < \bar{Z}_{u-1}(t).$$

Indeed, if $J_0(t_1) \cap \{1, 2, \ldots, u-1\} = \emptyset$, then $\bar{W}_i(t_1) \neq 0$ for $i = 1, 2, \ldots, u-1$. Otherwise for any $j \in J_0(t_1) \cap \{1, 2, \ldots, u-1\}$, by the definitions of u and $J_0(\cdot)$ we have $\bar{W}_j(t_1) = 0$ and $\frac{\mathrm{d}}{\mathrm{d}t}\bar{W}_j(t_1) \neq 0$, so there exists $\delta_j > 0$ such that $\bar{W}_j(t) \neq 0$ for any $t \in (t_1, t_1+\delta_j)$. In either case, there exists $\delta > 0$ such that $\bar{W}_i(t) \neq 0$ for any $t \in (t_1, t_1+\delta)$ and any $i \in \{1, 2, \ldots, u-1\}$. If $\bar{W}_i(t) < 0$ for some $i \in \{1, 2, \ldots, u-1\}$ and some $t \in (t_1, t_1+\delta)$, then $\bar{Z}_j(t) > \max\{\bar{Z}_{j-1}(t), \bar{Z}_{j+1}(t)\}$ for some $j \leq i$, which contradicts Lemma 1. Hence $\bar{W}_i(t) > 0$ for $i = 1, 2, \ldots, u-1$; i.e., $0 < \bar{Z}_1(t) < \bar{Z}_2(t) < \cdots < \bar{Z}_{u-1}(t)$. The claim then follows.

In the actual system with this strict order of queues, either all odd links up to l_u get scheduled at a time slot, or all even links up to l_u get scheduled. Then in our fluid limits we would have

$$\mu_i(t) = \mu_{i+2}(t) \quad \forall i \in \{1, 2, \ldots, u-2\}$$

for any regular time $t \in (t_1, t_1 + \delta)$. Then by the absolute continuity,

$$\bar{D}_3(t) - \bar{D}_1(t) = \bar{D}_3(t_1) - \bar{D}_1(t_1) + \int_{t_1}^{t} (\mu_3(s) - \mu_1(s))\,\mathrm{d}s = \bar{D}_3(t_1) - \bar{D}_1(t_1).$$

By the definition of derivatives, we have

$$\frac{\mathrm{d}}{\mathrm{d}t}(\bar{D}_3(t_1) - \bar{D}_1(t_1)) = \lim_{t \to t_1^+} \frac{(\bar{D}_3(t) - \bar{D}_1(t)) - (\bar{D}_3(t_1) - \bar{D}_1(t_1))}{t - t_1} = 0.$$

So $\mu_1(t_1) = \mu_3(t_1)$. Similarly, we have $\mu_i(t_1) = \mu_{i+2}(t_1)$ for $i = 1, 2, \ldots, u-2$.

APPENDIX B

PROOFS FOR CHAPTER 3

B.1 Proof of Theorem 2

Lemma 7. *The maximum stability region of the F-framed policies can be characterized by*

$$\overline{\Lambda_{\mathrm{NC}}(F)} = \left\{ \lambda \succeq 0 \colon \lambda F \preceq \sum_{J \in \mathcal{J}(F)} \pi(J)\eta_J, \eta_J \in \mathcal{CH}(M_J) \right\},$$

where $\overline{A}$ denotes the closure of A and $\mathcal{CH}(M_J)$ is the convex hull over the set of columns of M_J.

Proof of Lemma 7. If λ is strictly outside the maximum stability region, it can be proved that the total amount of deficits increase to infinity with probability one using the strictly separating hyperplane theorem (Boyd and Vandenberghe, 2004) and Lyapunov drift arguments. If λ is strictly inside the maximum stability region, then we can find $\eta = \big(\eta_J \colon J \in \mathcal{J}(F)\big)$ that dominates λ and make the long-term-average of the scheduling process be at least $\eta \succ \lambda$, where $\succ$ denotes strict pairwise greater than. So the system can be stabilized. □

We then present the next lemma.

Lemma 8. *For any F,*

$$\Lambda_{\mathrm{NC}}(F) \supseteq \mathrm{int}\left(\Lambda \cap \left(\Lambda - \frac{\tau_{\max}}{F}\mathbf{1}\right)\right),$$

where $\Lambda - \frac{\tau_{\max}}{F}\mathbf{1} = \big\{\lambda - \frac{\tau_{\max}}{F}\mathbf{1} \colon \lambda \in \Lambda\big\}$.

Proof of Lemma 8. Note that given the schedules of any causal policy, we can convert them into valid schedules under the F-framed policy by removing those transmissions that serve those packets whose arrival times and transmission times are not in the same frame. Since at most one packet can be scheduled on a link at each time slot and the maximum delay bound is $\tau_{\max}$, the number of packets across frame (i.e., those arriving in one frame with deadlines in another) scheduled by the causal policy on a link is at most $\tau_{\max}$ at the end of a frame. As a result, we only need to remove at most $\tau_{\max}$ transmissions for each frame on each link, which is equivalent to at most $\tau_{\max}/F$ packets per time slot. So the lemma holds. □

Lemma 9. *If $F > \tau_{\max}$, then*

$$\Lambda_{\mathrm{LDF}} \supseteq \sigma^* \cdot \mathrm{int}(\Lambda_{\mathrm{NC}}(F)).$$

Proof of Lemma 9. Let $\lambda' \in \text{int}(\Lambda_{\text{NC}}(F))$ and let $\lambda = \sigma^*\lambda'$. Then by the definition of interior point and the characterization of $\Lambda_{\text{NC}}(F)$ in Lemma 7, there exist $\big(\xi_J \colon J \in \mathcal{J}(F)\big)$ with $\xi_J \in \mathcal{CH}(M_J)$ for each $J \in \mathcal{J}(F)$ and $\delta > 0$ such that

$$\lambda' + \delta \mathbf{1} \preceq \frac{1}{F} \sum_{J \in \mathcal{J}(F)} \pi(J) \xi_J, \tag{B.1}$$

where $\mathbf{1}$ is a vector with all 1's.

We now establish the fluid limits for the system sampled every F time slots. Let $\big(D(t) \colon t \in \mathbb{N}\big)$ and $\big(S(t) \colon t \in \mathbb{N}\big)$ be the cumulative deficit and service processes under LDF (without frame). Let $\Psi_{l,\tau}(t)$ be the number of packets with deadline $t + \tau - 1$ on link l at time slot t, and let $\Psi(t) = \big(\Psi_{l,\tau}(t) \colon l \in \mathcal{K}, 1 \le \tau \le \tau_{\max}\big)$. Then under LDF $\big((\Psi(t), D(t)) \colon t \in \mathbb{N}\big)$ is a Markov chain. Let $\Big((\Psi^{(n)}(t), D^{(n)}(t)) \colon t \in \mathbb{N}\Big)$ be the system with arbitrary initial state $(\Psi^{(n)}(0), D^{(n)}(0))$ associated with the requirement that $\|(\Psi^{(n)}(0), D^{(n)}(0))\| \triangleq \sum_{l \in \mathcal{K}} \sum_{\tau=1}^{\tau_{\max}} \Psi_{l,\tau}^{(n)}(0) + \sum_{l \in \mathcal{K}} D_l^{(n)}(0) = n$ for any $n \in \mathbb{N}$, and let $S^{(n)}(t)$ be the corresponding cumulative service process. We sample $\Psi^{(n)}$, $D^{(n)}$ and $S^{(n)}$ every F time slots to get $\Psi^{(n)(F)}(t) = \Psi^{(n)}(Ft)$, $D^{(n)(F)}(t) = D^{(n)}(Ft)$ and $S^{(n)(F)}(t) = S^{(n)}(Ft)$ for $t \in \mathbb{N}$. Define the *scaled* deficit and service processes to be

$$\bar{D}^{(n)(F)}(t) = \frac{1}{n} D^{(n)(F)}(\lfloor nt \rfloor)$$

and

$$\bar{S}^{(n)(F)}(t) = \frac{1}{n} S^{(n)(F)}(\lfloor nt \rfloor).$$

Note that the scaled processes are defined for any nonnegative real number t rather than just integers, and can take values in vectors of multiples of $\frac{1}{n}$ rather than vectors of integers. Following Lemma 1 in Andrews *et al.* (2004), for almost all sample paths and any sequence of initial states there exists a subsequence $\big(n_j\big)$ such that for any $l \in \mathcal{K}$

$$\bar{D}_l^{(n_j)(F)} \to \bar{D}_l^{(F)} \quad \text{u.o.c.} \tag{B.2}$$

and

$$\bar{S}_l^{(n_j)(F)} \to \bar{S}_l^{(F)} \quad \text{u.o.c.}$$

as $j \to \infty$, where u.o.c. denotes uniform convergence over compact sets, and $\bar{D}_l^{(F)}$ and $\bar{S}_l^{(F)}$ are nonnegative nondecreasing Lipschitz-continuous functions with domain $\mathbb{R}_+$. The limiting functions are called the fluid limits.

Let $L_0^{(F)}(t)$ be the set of links with the largest deficit fluids at time t, and let $L^{(F)}(t) \subseteq L_0^{(F)}(t)$ be the set of links in $L_0^{(F)}(t)$ with largest derivatives at time t; i.e.,

$$L_0^{(F)}(t) = \left\{ l \in \mathcal{K} \colon \bar{D}_l^{(F)}(t) = \max_{i \in \mathcal{K}} \bar{D}_i^{(F)}(t) \right\}$$

and

$$L^{(F)}(t) = \left\{ l \in L_0^{(F)}(t) \colon \frac{\mathrm{d}}{\mathrm{d}t} \bar{D}_l^{(F)}(t) = \max_{i \in L_0(t)} \frac{\mathrm{d}}{\mathrm{d}t} \bar{D}_i^{(F)}(t) \right\},$$

where we assume t is a regular point; i.e., the derivatives of the fluid limits exist at t. Then we can construct $\left(\eta_J \in \mathcal{CH}(M_J) \colon J \in \mathcal{J}(F)\right)$ such that for any $l \in L^{(F)}(t)$, the service fluids satisfy

$$\frac{\mathrm{d}}{\mathrm{d}t} \bar{S}_l^{(F)}(t) \geq \sum_{J \in \mathcal{J}(F)} \pi(J) \eta_{J,l} - \tau_{\max}, \tag{B.3}$$

where $\eta_{J,l}$ is the lth entry of the vector η_J.

To understand (B.3), note that $\bar{S}^{(F)}(t)$ is the fluid limit of the service process sampled every F time slots, so the derivative of $\bar{S}^{(F)}(t)$ is the average service over F time slots under LDF. Now consider a frame of F time slots with arrival and maximum delay pattern J, and denote by s_J^{LDF} the schedule under LDF during the F time slots. We next construct another schedule s_J^F, which is a maximal schedule under the F-framed policy. The construction is by removing those transmissions in s_J^{LDF} which serve packets that arrived before the frame started, and then add more transmissions to make it a maximal schedule. So for link l,

$$W(s_J^{\mathrm{LDF}})_l - o_{J,l} + n_{J,l} = W(s_J^F)_l,$$

where $o_{J,l}$ is the number of removed transmissions on link l, $n_{J,l}$ is the number of added transmissions on link l, $W(s_J^F) \in \mathcal{M}_J$, and $(\cdot)_l$ denotes the lth component of the vector. Note that those removed transmissions must occur at the first $\tau_{\max}$ time slots because the maximum delay is $\tau_{\max}$ so none of the packets that arrived before the frame can be transmitted after the first $\tau_{\max}$ time slots. This also implies that the added transmissions must be in the first $\tau_{\max}$ time slots as well. Therefore, $n_{J,l} \leq \tau_{\max}$, and

$$W(s_J^{\mathrm{LDF}})_l \geq W(s_J^F)_l - \tau_{\max}$$

holds for any J.

Now assuming $\bar{D}_l^{(F)}(t) > 0$ for $l \in L^{(F)}(t)$, the derivative of $\bar{D}_l^{(F)}(t)$ is

$$\frac{\mathrm{d}}{\mathrm{d}t}\bar{D}_l^{(F)}(t) = \lambda_l F - \frac{\mathrm{d}}{\mathrm{d}t}\bar{S}_l^{(F)}(t) \tag{B.4}$$

$$\leq \lambda_l F - \sum_{J \in \mathcal{J}(F)} \pi(J)\eta_{J,l} + \tau_{\max} \tag{B.5}$$

$$\leq \sigma^* \left(\sum_{J \in \mathcal{J}(F)} \pi(J)\xi_{J,l} - \delta F \right) - \sum_{J \in \mathcal{J}(F)} \pi(J)\eta_{J,l} + \tau_{\max} \tag{B.6}$$

$$= \left[\sigma^* \left(\sum_{J \in \mathcal{J}(F)} \pi(J)\xi_{J,l} \right) \right. \tag{B.7}$$

$$\left. - \left(\sum_{J \in \mathcal{J}(F)} \pi(J)\eta_{J,l} \right) \right] \tag{B.8}$$

$$- \sigma^* \delta F + \tau_{\max}, \tag{B.9}$$

where (B.5) comes from (B.3), and (B.6) holds because $\lambda = \sigma^* \lambda'$ and inequality (B.1). By the definition of the R-LPF and the fact that $L^{(F)}(t)$ has higher scheduling priority over $\mathcal{K} \backslash L^{(F)}(t)$, there exists $i \in L^{(F)}(t)$ such that

$$\sigma^* \left(\sum_{J \in \mathcal{J}(F)} \pi(J)\xi_{J,i} \right) \leq \left(\sum_{J \in \mathcal{J}(F)} \pi(J)\eta_{J,i} \right).$$

Thus by definition of $L^{(F)}(t)$,

$$\frac{\mathrm{d}}{\mathrm{d}t}\bar{D}_l^{(F)}(t) = \frac{\mathrm{d}}{\mathrm{d}t}\bar{D}_i^{(F)}(t) \leq \tau_{\max} - \sigma^* \delta F.$$

We note that for any positive integer k,

$$\Lambda_{\mathrm{NC}}(F) \subseteq \Lambda_{\mathrm{NC}}(kF),$$

since any F-framed policy is a valid but more restrictive kF-framed policy. Then (B.1) holds with the same δ for any frame size kF. Thus for large enough integer k, the deficit fluid limits associated with the frame size kF satisfy

$$\frac{\mathrm{d}}{\mathrm{d}t}\bar{D}_l^{(kF)}(t) \leq \tau_{\max} - \sigma^* \delta k F \leq -\epsilon < 0$$

for some $\epsilon > 0$, as long as $\max_l \bar{D}_l^{(kF)}(t) > 0$ and t is regular. Since $\|\bar{D}^{(kF)}(0)\| = 1$, we have $\|\bar{D}^{(kF)}(t)\| = 0$ for any $t \geq 1/\epsilon$. By the convergence in (B.2) and the arbitrary choice of initial states of the systems with the prescribed requirements, we have that $\|\bar{D}^{(n)(kF)}(t)\| \to 0$ almost surely as $n \to \infty$ for any $t \geq 1/\epsilon$. Since $\left(\bar{D}^{(n)(kF)}(t) \colon n \in \mathbb{N}\right)$ is uniformly integrable (see, e.g., Dai (1995)), we get that $\mathbb{E}\,\|\bar{D}^{(n)(kF)}(t)\| \to 0$ as $n \to \infty$ for $t \geq 1/\epsilon$. Note that $\sum_{l\in\mathcal{K}} \sum_{\tau=1}^{\tau_{\max}} \Psi_{l,\tau}^{(n)(kF)}(nt) \leq K a_{\max} \tau_{\max}^2$. We then have

$$\lim_{n\to\infty} \mathbb{E}\left[\frac{1}{n}\|(\Psi^{(n)(kF)}(nt), D^{(n)(kF)}(nt))\|\right] = 0$$

for any $t \geq 1/\epsilon$. Then by Theorem 4 in Andrews *et al.* (2004) we get that the sampled deficit process of the original system $\left(D^{(kF)}(t) \colon t \in \mathbb{N}\right)$ is stable as defined in Andrews *et al.* (2004), which implies the existence of a stationary distribution of $\left(D^{(kF)}(t) \colon t \in \mathbb{N}\right)$ and in turn implies the stability of $\left(D^{(kF)}(t) \colon t \in \mathbb{N}\right)$ as defined in Definition 1. Finally,

$$\begin{aligned}
&\lim_{C\to\infty} \limsup_{t\to\infty} \mathbb{P}\left(\sum_{l\in\mathcal{K}} D_l(t) \geq C\right) \\
&\leq \lim_{C\to\infty} \limsup_{t\to\infty} \mathbb{P}\left(\sum_{l\in\mathcal{K}} D_l^{(kF)}\left(\left\lfloor \frac{t}{kF} \right\rfloor\right) + kKFa_{\max} \geq C\right) \\
&= \lim_{C\to\infty} \limsup_{t\to\infty} \mathbb{P}\left(\sum_{l\in\mathcal{K}} D_l^{(kF)}(t) \geq C - kKFa_{\max}\right) \\
&= 0.
\end{aligned}$$

So the original unsampled deficit process $\left(D(t) \colon t \in \mathbb{N}\right)$ is stable, and therefore the deficit arrival rate $\lambda = \sigma^* \lambda' \in \Lambda_{\mathrm{LDF}}$. □

We can now proceed to prove Theorem 2.

Proof of Theorem 2. By Lemma 9 and Lemma 8, we have

$$\Lambda_{\mathrm{LDF}} \supseteq \sigma^* \cdot \mathrm{int}\left(\Lambda \cap \left(\Lambda - \frac{\tau_{\max}}{F}\mathbf{1}\right)\right).$$

The theorem is obtained by letting $F \to \infty$. □

B.2 Proof of Theorem 3

For given $\alpha \in \mathbb{R}_+^K$, let α_L denote the vector α restricted to the subset $L \subseteq \mathcal{K}$. We define the pressure of link i in the subset $L \subseteq \mathcal{K}$ for vector α to be

$$\kappa_{i,L}(\alpha) = \frac{\alpha_i}{\alpha_i + \max_{I \in \mathcal{I}(i,L)} \sum_{j \in I} \alpha_j},$$

where $\mathcal{I}(i, L) \subseteq \mathcal{P}(N_L(i))$ is the collection of subsets of the neighbors of i in L that can be scheduled simultaneously. We also define the minimum pressure in the subset $L \subseteq \mathcal{K}$ for α to be

$$\psi_L(\alpha) = \min_{i \in L} \kappa_{i,L}(\alpha).$$

We can check with (3.3) and (3.4) that

$$\kappa_{i,\mathcal{K}}(\alpha) = \kappa_i(\alpha), \psi_{\mathcal{K}}(\alpha) = \psi(\alpha).$$

We note the following lemma.

Lemma 10. *For any $\alpha \in \mathbb{R}_+^K$ and any $L \subseteq \mathcal{K}$,*

$$\psi_L(\alpha) \geq \psi(\alpha).$$

Proof.

$$\begin{aligned}
\psi_L(\alpha) &= \min_{i \in L} \kappa_{i,L}(\alpha) \\
&= \kappa_{i^*,L}(\alpha) \\
&= \frac{\alpha_{i^*}}{\alpha_{i^*} + \sum_{j \in I^*} \alpha_j} \\
&\geq \frac{\alpha_{i^*}}{\alpha_{i^*} + \sum_{j \in I^{**}} \alpha_j} \\
&\geq \min_{i \in \mathcal{K}} \kappa_i(\alpha) \\
&= \psi(\alpha),
\end{aligned}$$

for some $i^* \in L$, some $I^* \in \mathcal{I}(i^*, L)$ and some $I^{**} \in \mathcal{I}(i^*, \mathcal{K})$. □

The following lemma is the key to the proof of Theorem 3.

Lemma 11. *For any $\alpha \in \mathbb{R}_+^K$, any F, any $J \in \mathcal{J}(F)$ and any $L \subseteq \mathcal{K}$,*

$$\frac{\min_{\phi \in M_{J,L}} \alpha_L^{\mathrm{T}} \phi}{\max_{\phi \in M_{J,L}} \alpha_L^{\mathrm{T}} \phi} \geq \psi(\alpha).$$

Proof of Lemma 11. We fix $J \in \mathcal{J}(F)$ and $L \subseteq \mathcal{K}$ and focus on the arrival and maximum delay pattern given by J restricted to the subset of links L. If $\psi(\alpha) = 0$ then the result is trivial, so we may assume $\psi(\alpha) > 0$; i.e., $\alpha \succ 0$. For each link $l \in L$, replace it with n links (each of which has a single packet arrival in the frame) if the total number of packets arriving on l in the frame is $n \geq 2$, leave it alone if the total number of packets arriving on l in the frame is 1, and remove it from our consideration if no packet arrives in this frame according to J. We then get a multigraph whose set of links is denoted by $\mathcal{K}'$, where $K' = |\mathcal{K}'|$ equals the total number of packets arriving on L in the original conflict graph according to J, and each link in $\mathcal{K}'$ represents a packet in the original conflict graph with arriving time and deadline given by J. The interference model of $\mathcal{K}'$ inherits from the interference model of $\mathcal{K}$, plus that two links in $\mathcal{K}'$ that correspond to the same link in $\mathcal{K}$ interfere each other. Let $I'(l)$ denote the set of links that interfere with link l in $\mathcal{K}'$, and by convention assume $l \in I'(l)$. Also let $\mathcal{I}'(l)$ be the collection of subsets of $I'(l)\backslash\{l\}$ that can be scheduled simultaneously according to the interference model.

A multi-schedule over the multigraph $\mathcal{K}'$ in the frame is represented by a function (we overload the symbol s for convenience in this proof)

$$\begin{aligned} s\colon \mathcal{K}' \times \{1, 2, \ldots, F\} &\to \{0, 1\} \\ \langle i, t\rangle &\mapsto s_i(t) \end{aligned}$$

with $s_i(t) = 1$ if link $i \in \mathcal{K}'$ is scheduled by s at time slot t, and $s_i(t) = 0$ otherwise. A multi-schedule s is feasible if no two interfering links are scheduled at the same time slot, no link is scheduled before its arriving time or after its deadline, and each link is scheduled at most once during the entire frame. A feasible multi-schedule s is maximal if no more links can be scheduled without breaking the feasibility. We note that a feasible (or maximal, respectively) schedule for J over the original set of links $\mathcal{K}$ corresponds to a feasible (or maximal, respectively) multi-schedule over the multigraph $\mathcal{K}'$ given by J. Let $\mathrm{supp}(s)$ be the support of s, i.e., the set of $\langle$link, time slot$\rangle$ pairs of scheduled links by s. Let $\|s\| = \sum_i \left(\bar{\alpha}_i \sum_t s_i(t)\right)$, where $\bar{\alpha}_i = \alpha_j$ if link i in $\mathcal{K}'$ corresponds to link j in $\mathcal{K}$. Then we say $\|s\|$ is the weight of the multi-schedule s.

Define the interference neighborhood of the $\langle$link, time slot$\rangle$ pair $\langle i, t\rangle$ to be the interfering links of link i at time slot t; i.e.,

$$I'(i, t) = I'(i) \times \{t\} \subseteq X,$$

where $X = \mathcal{K}' \times \{1, 2, \ldots, F\}$.

We now consider another maximal multi-schedule u, and the set of $\langle$link, time slot$\rangle$ pairs that are in $\mathrm{supp}(u)$ but not in the union of the interference neighborhoods of $\langle$link, time slot$\rangle$ pairs in $\mathrm{supp}(s)$, i.e., the set

$$P = \mathrm{supp}(u)\backslash \bigcup_{\langle i,t\rangle \in \mathrm{supp}(s)} I'(i, t).$$

We note that for any ⟨link, time slot⟩ pair $\langle j, t' \rangle \in P$, we must have $\langle j, \tilde{t} \rangle \in \text{supp}(s)$ for some $\tilde{t} \neq t'$; in other words, link j must be scheduled in s at some time slot other than t'. This holds because otherwise, link j can be added to s at time t without interfering any links in s (note that $\langle j, t' \rangle$ is not in any interference neighborhoods of ⟨link, time slot⟩ pairs in $\text{supp}(s)$). We use an example in Figure 31 to illustrate this point.

Let $u|_P$ be the multi-schedule u supported in P; i.e., $u|_P(i,t) = u(i,t)\mathbb{1}_P(\langle i, t \rangle)$, where

$$\mathbb{1}_P(\langle i, t \rangle) = \begin{cases} 1 & \text{if } \langle i, t \rangle \in P \\ 0 & \text{otherwise} \end{cases}$$

is the indicator function. Then from the analysis above, we know that

$$\|u|_P\| \leq \|s\|.$$

Furthermore, the multi-schedule u can have at most some set of links $I' \in \mathcal{I}'(i)$ active in $I(i,t)$ for any $\langle i, t \rangle \in \text{supp}(s)$ because of the interference. Therefore, we have

$$\begin{aligned} \|u\| &= \|u|_P\| + \|u|_{X \setminus P}\| \\ &\leq \|s\| + \sum_{\langle i,t \rangle \in \text{supp}(s)} \|u|_{I(i,t)}\| \\ &\leq \|s\| + \sum_{\langle i,t \rangle \in \text{supp}(s)} \max_{I \in \mathcal{I}'(i)} \|\mathbb{1}_{I \times \{t\}}\| \\ &= \sum_{\langle i,t \rangle \in \text{supp}(s)} \left(\bar{\alpha}_i + \max_{I \in \mathcal{I}'(i)} \|\mathbb{1}_{I \times \{t\}}\| \right). \end{aligned}$$

Then

$$\begin{aligned}
\frac{\|s\|}{\|u\|} &\geq \frac{\sum_{\langle i,t\rangle\in\mathrm{supp}(s)} \bar{\alpha}_i}{\sum_{\langle i,t\rangle\in\mathrm{supp}(s)} \left(\bar{\alpha}_i + \max_{I\in\mathcal{I}'(i)} \|\mathbb{1}_{I\times\{t\}}\|\right)} \\
&\geq \min_{\langle i,t\rangle\in\mathrm{supp}(s)} \frac{\bar{\alpha}_i}{\bar{\alpha}_i + \max_{I\in\mathcal{I}'(i)} \|\mathbb{1}_{I\times\{t\}}\|} \\
&\geq \min_{j\in L} \frac{\alpha_j}{\alpha_j + \max_{I\in\mathcal{I}(j,L)} \sum_{k\in I} \alpha_k} \\
&= \psi_L(\alpha) \\
&\geq \psi(\alpha),
\end{aligned}$$

where the last inequality comes from Lemma 10. Since s and u are any arbitrary maximal multi-schedules, we have

$$\frac{\min_{\phi\in M_{J,L}} \alpha_L^{\mathrm{T}}\phi}{\max_{\phi\in M_{J,L}} \alpha_L^{\mathrm{T}}\phi} \geq \psi(\alpha).$$

□

Combining (3.2) and Lemma 10, we have

$$\begin{aligned}
\sigma_L^*(F) &\geq \frac{\min_{\phi\in M_L(F)} \alpha_L^{\mathrm{T}}\phi}{\max_{\phi\in M_L(F)} \alpha_L^{\mathrm{T}}\phi} \\
&= \frac{\sum_{J\in\mathcal{J}(F)} \pi(J)\alpha_L^{\mathrm{T}}\phi_J^{(1)}}{\sum_{J\in\mathcal{J}(F)} \pi(J)\alpha_L^{\mathrm{T}}\phi_J^{(2)}} \\
&\geq \min_{J\in\mathcal{J}(F)} \frac{\alpha_L^{\mathrm{T}}\phi_J^{(1)}}{\alpha_L^{\mathrm{T}}\phi_J^{(2)}} \\
&\geq \psi(\alpha).
\end{aligned}$$

By taking supremum over α we get Theorem 3.

B.3 Proof of Corollary 3

Choose an arbitrary leaf l_0 in the tree and set it to be the root. Define the depth of each link to be the number of hops it is away from the root. Then the root has depth 0 and any other link has depth greater than 0. For any link l, set the weight of

it to be $\alpha_l = (\beta-1)^{-\operatorname{dep}(l)/2}$, where $\operatorname{dep}(l)$ is the depth of link l. Note that the root has the largest weight 1. Then each link has at most one parent and at most $\beta-1$ children, where the parent has a weight $\sqrt{\beta-1}$ times the weight of that link and the children have weights $\frac{1}{\sqrt{\beta-1}}$ times the weight of that link. Hence the pressure of link l for α is

$$\kappa_l(\alpha) \geq \frac{1}{1+\sqrt{\beta-1}+(\beta-1)\frac{1}{\sqrt{\beta-1}}} = \frac{1}{2\sqrt{\beta-1}+1}.$$

Then by the definition of minimum pressure for α and Theorem 3 we get Corollary 3.

B.4 Proof of Theorem 4

By the definition of the interference degree β, there exists a link l_0 together with β of its neighbors $\{l_1, l_2, \ldots, l_\beta\} \subseteq N(l_0)$ such that no two links in $\{l_1, l_2, \ldots, l_\beta\}$ interfere each other. Now let $L = \{l_0, l_1, l_2, \ldots, l_\beta\}$, then L induces a star network with the center link l_0 and β leaves in the conflict graph. In this proof we construct periodic traffic processes with packets only arriving on L, and show that the efficiency ratio of LDF is $\gamma^*_{\text{LDF}} \leq \frac{1}{\sqrt{\beta}+1}$.

Let the traffic be such that packets only arrive on $L = \{l_0, l_1, \ldots, l_\beta\}$ at odd time slots with the following 2-patterns (a_i is the number of arriving packets on link l_i, and the vector τ_i are the associated maximum delays)

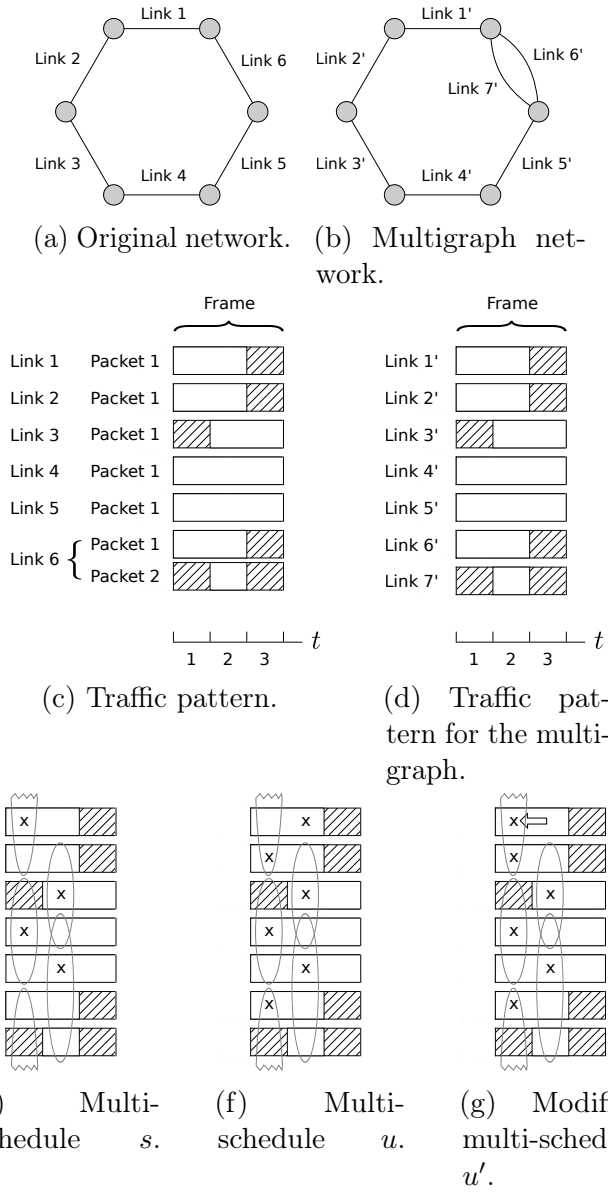
- Pattern 0 (J_0): $a_i = 1$ for all $i = 0, 1, 2, \ldots, \beta$, $\tau_0 = 2$, and $\tau_i = 1$ for $i = 1, 2, \ldots, \beta$.
- Pattern j ($J_j, j = 1, 2, \ldots, \beta$): $a_0 = 1$, $a_i = 1$, $a_k = 0$ for any other k, $\tau_0 = 1$, $\tau_j = 2$.

Note that since no packets arrive at the even time slots and the maximum delay is 2, all packets expire at the end of the even time slots. Also note that each 2-pattern takes two time slots.

Now consider periodic traffic with a period of $2n$ time slots, where each period consists of n_0 consecutive J_0's, followed by n_1 consecutive J_1's, and then n_2 consecutive J_2's, and so on, till n_β consecutive J_β's, with $n = \sum_{i=0}^{\beta} n_i$. We then notice that for each traffic pattern J_j, all of the packets can be scheduled if the packets with maximum delay 1 get scheduled in the odd time slots. Hence the optimal scheduler can always achieve the QoS vector $p_{\text{OPT}} = \mathbf{1}$. For LDF, we assume each time the pattern J_j arrives, only deficit j increases by one, and LDF chooses to schedule link j, reducing deficit j by one. Then with all the deficits unchanged, LDF achieves the QoS vector of

$$p_{\text{LDF}} = \left(\frac{n_0}{n}, \frac{n_1}{n_0+n_1}, \frac{n_2}{n_0+n_2}, \ldots, \frac{n_\beta}{n_0+n_\beta}\right).$$

Then we can easily see that LDF cannot achieve QoS vector $p = p_{\text{LDF}} + \epsilon\mathbf{1}$ for any $\epsilon > 0$ since adding an extra ϵ/n deficit for each pattern on each link would make the deficits grow without bound. Now let n_0/n approximate $\frac{1}{\sqrt{\beta}+1}$ and let n_j approximate $\frac{1}{\beta+\sqrt{\beta}}$ for $j = 1, 2, \ldots, \beta$. Then we will have that p_{LDF} approximates $\frac{1}{\sqrt{\beta}+1}\mathbf{1}$. Therefore the efficiency ratio of LDF under the given traffic and deficit arrival is at most $\frac{1}{\sqrt{\beta}+1}$.



(a) Original network. (b) Multigraph network.

(c) Traffic pattern. (d) Traffic pattern for the multigraph.

(e) Multi-schedule s. (f) Multi-schedule u. (g) Modified multi-schedule u'.

Figure 31: Illustration of the proof of Theorem 3

Note: Consider a six-cycle network with one-hop interference as in (a) and traffic pattern for a frame of 3 time slots as in(c). The network is converted into a multigraph in (b) by dividing the two packets on link 6 into two links with the same end nodes, and the corresponding traffic pattern J is shown in (d). Two maximal multi-schedules, s and u, are given in (e) and (f) with x's denoting the scheduled links, and the one-hop neighborhoods of the scheduled links of s are illustrated by circles. We see that the one-hop neighborhoods of s cover all scheduled links by s, but miss one link scheduled by u. However, as shown in (g), the missed link can be inserted into the one-hop neighborhood in the first time slot so that all scheduled links by the modified multi-schedule u are now covered by the one-hop neighborhoods of s.

APPENDIX C

PROOFS FOR CHAPTER 4

C.1 Proof of Theorem 5

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

$$V(q) = \sum_{k=1}^{n} q_k^2.$$

Let $x, y \in \mathbb{N}^n$ denote the state of the Markov chains, and $q_{x,y}$ denote the transition rate from state x to state y. According to the Foster–Lyapunov theorem for continuous-time Markov chain (see, for example, Theorem 9.1.8 in Srikant and Ying (2014)), we consider

$$\sum_{y \neq x} q_{x,y} \left(V(y) - V(x) \right). \tag{C.1}$$

Define $1 \times n$ vector e_k such that $e_k[k] = 1$ and $e_k[l] = 0$ for any $l \neq k$. Then

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

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

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

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:

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

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

Therefore, we have

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

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

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

which implies that

$$\mathbb{E}\left[\frac{1}{n} \sum_k \hat{Q}_k\right] \leq \frac{1 + \lambda}{2 - 2\lambda}.$$

Therefore, the theorem holds by choosing $c = \frac{1+\lambda}{2-2\lambda}$.

C.2 Proof of Lemma 4

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 $\mu_j = \mathbb{E}[X_j] = (dm - 1)\pi_j$.

Consider any i such that $i \geq \bar{Q}_\pi$. The probability that server 1 receives a task in

water filling is upper bounded by

$$\begin{aligned}
&\mathbb{E}\left[\left(\frac{m-\sum_{j=0}^{i-1}(i-j)X_j}{1+\sum_{j=0}^{i}X_j}\right)^+\right]\\
\leq&\mathbb{E}\left[\left(\frac{m-\sum_{j=0}^{\bar{Q}_\pi-1}(\bar{Q}_\pi-j)X_j}{1+\sum_{j=0}^{\bar{Q}_\pi}X_j}\right)^+\right]\\
=&\mathbb{E}\left[\left(\frac{\frac{m}{dm-1}-\sum_{j=0}^{\bar{Q}_\pi-1}(\bar{Q}_\pi-j)\frac{X_j}{dm-1}}{\frac{1}{dm-1}+\sum_{j=0}^{\bar{Q}_\pi}\frac{X_j}{dm-1}}\right)^+\right] \qquad \text{(C.2)}
\end{aligned}$$

which converges to

$$\left(\frac{\frac{1}{d}-\sum_{j=0}^{\bar{Q}_\pi-1}(\bar{Q}_\pi-j)\pi_j}{\sum_{j=0}^{\bar{Q}_\pi}\pi_j}\right)^+ \qquad \text{(C.3)}$$

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 $\bar{Q}_\pi$ (4.3), we know that

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

so $(C.2)\to 0$ and,

$$q_{i,j}=0 \quad i\geq\bar{Q}_\pi \text{ and } j\notin\{i,i-1\}. \qquad \text{(C.4)}$$

Now we assume $i<\bar{Q}_\pi$. 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 $Q \geq \bar{Q}_\pi + 1$, according to the definition of $\bar{Q}_\pi$, we have

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

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

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

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

$$\begin{aligned}
&\frac{\frac{1}{d} - \sum_{j=0}^{Q-2}(Q-1-j)\pi_j}{\sum_{j=0}^{Q-1}\pi_j} \\
\geq &\frac{\frac{1}{d} - \sum_{j=0}^{\bar{Q}_\pi - 3}(\bar{Q}_\pi - 2 - j)\pi_j}{\sum_{j=0}^{\bar{Q}_\pi - 2}\pi_j} \\
\geq &\frac{\sum_{j=0}^{\bar{Q}_\pi - 2}(\bar{Q}_\pi - 1 - j)\pi_j - \sum_{j=0}^{\bar{Q}_\pi - 3}(\bar{Q}_\pi - 2 - j)\pi_j}{\sum_{j=0}^{\bar{Q}_\pi - 2}\pi_j} \\
= &1. \qquad \text{(C.5)}
\end{aligned}$$

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

$$q_{i,j} = \begin{cases} \lambda d \alpha_\pi, & \text{if } j = \bar{Q}_\pi \\ \lambda d(1-\alpha_\pi), & \text{if } j = \bar{Q}_\pi - 1 \\ 0, & \text{otherwise.} \end{cases} \qquad \text{(C.6)}$$

Hence, the lemma holds.

C.3 Proof of Theorem 9

Motivated by the proof in Mitzenmacher (1996), 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

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

in three different cases.

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

$$\frac{\mathrm{d}\epsilon_i}{\mathrm{d}t} = \begin{cases} -(1+\lambda d)\epsilon_i + \epsilon_{i+1} & i \leq \bar{Q}_{BF} - 1, \\ \lambda d \sum_{j=0}^{i-1} \epsilon_j - \epsilon_i + \epsilon_{i+1}, & i = \bar{Q}_{BF} \\ -\epsilon_i + \epsilon_{i+1} & \text{otherwise.} \end{cases} \tag{C.7}$$

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

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

  which implies that

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

  Similarly, we can obtain that

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

  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,

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

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

$$\frac{\mathrm{d}\epsilon_i}{\mathrm{d}t} \leq -(1+\lambda d)|\epsilon_i| + |\epsilon_{i+1}| \quad \forall\ i \leq \bar{Q}_{BF} - 1. \tag{C.8}$$

  We next consider two subcases.

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

$$\begin{aligned}
\sum_{i=\bar{Q}_{BF}}^{\infty} \frac{\mathrm{d}|\epsilon_i|}{\mathrm{d}t} = \sum_{i=\bar{Q}_{BF}}^{\infty} \frac{\mathrm{d}\epsilon_i}{\mathrm{d}t} &= \sum_{i=\bar{Q}_{BF}}^{\infty} \frac{\mathrm{d}s_i}{\mathrm{d}t} \\
&= \lambda - \lambda d \sum_{j=0}^{\bar{Q}_{BF}-1} (1 - s_j) - s_{\bar{Q}_{BF}} \\
&= \lambda d \sum_{j=0}^{\bar{Q}_{BF}-1} \epsilon_j - \epsilon_{\bar{Q}_{BF}} \\
&\leq -|\epsilon_{\bar{Q}_{BF}}| + \lambda d \sum_{j=0}^{\bar{Q}_{BF}-1} |\epsilon_j|.
\end{aligned} \tag{C.9}$$

  Combining (C.8) and (C.9), we obtain

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

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

$$\begin{aligned}
&\sum_{i=\bar{Q}_{BF}+1}^{\infty} \frac{\mathrm{d}|\epsilon_i|}{\mathrm{d}t} \\
&= \sum_{i=\bar{Q}_{BF}+1}^{\infty} \frac{\mathrm{d}\epsilon_i}{\mathrm{d}t} = \sum_{i=\bar{Q}_{BF}+1}^{\infty} \frac{\mathrm{d}s_i}{\mathrm{d}t} \\
&= \lambda - \lambda d \sum_{j=0}^{\bar{Q}_{BF}} (1 - s_j) - s_{\bar{Q}_{BF}+1} \\
&= \lambda - \lambda d \sum_{j=0}^{\bar{Q}_{BF}} (1 - \hat{s}_j) \\
&\quad + \lambda d \sum_{j=0}^{\bar{Q}_{BF}} \epsilon_j - \epsilon_{\bar{Q}_{BF}+1} \\
&\leq \lambda d \sum_{j=0}^{\bar{Q}_{BF}} |\epsilon_j| - |\epsilon_{\bar{Q}_{BF}+1}|, \qquad \text{(C.10)}
\end{aligned}$$

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

$$\begin{aligned}
&\frac{\mathrm{d}|\epsilon_{\bar{Q}_{BF}}|}{\mathrm{d}t} \\
=& - \frac{\mathrm{d}s_{\bar{Q}_{BF}}}{\mathrm{d}t} \\
=& - \lambda d + (1 + \lambda d) s_{\bar{Q}_{BF}} - s_{\bar{Q}_{BF}+1} \\
=& - \lambda d + (1 + \lambda d) \hat{s}_{\bar{Q}_{BF}} \\
&\quad + (1 + \lambda d) \epsilon_{\bar{Q}_{BF}} - \epsilon_{\bar{Q}_{BF}+1} \\
\leq& - (1 + \lambda d) |\epsilon_{\bar{Q}_{BF}}| + |\epsilon_{\bar{Q}_{BF}+1}|, \qquad \text{(C.11)}
\end{aligned}$$

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

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

Combining inequalities (C.8), (C.10) and (C.11), we obtain

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

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

$$\sum_{i=\bar{Q}_{BF}+1}^{\infty} \frac{\mathrm{d}|\epsilon_i|}{\mathrm{d}t} = \sum_{i=\bar{Q}_{BF}+1}^{\infty} \frac{\mathrm{d}s_i}{\mathrm{d}t} = -|\epsilon_{\bar{Q}_{BF}+1}|, \tag{C.12}$$

and

$$\frac{\mathrm{d}\epsilon_i}{\mathrm{d}t} \leq -(1+\lambda d)|\epsilon_i| + |\epsilon_{i+1}| \quad \forall\ i < \bar{X}_s. \tag{C.13}$$

We next further consider the following subcases.

  - Assume $s_{\bar{X}_s} < \hat{s}_{\bar{X}_s}$, so

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

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

$$\begin{aligned}
\frac{\mathrm{d}|\epsilon_{\bar{X}_s}|}{\mathrm{d}t} =& -\lambda + \lambda d \sum_{j=0}^{\bar{X}_s}(1-\hat{s}_j) \\
& - \lambda d \sum_{j=0}^{\bar{X}_s-1} \epsilon_j + \epsilon_{\bar{X}_s} - \epsilon_{\bar{X}_s+1} \\
\leq& -\lambda + \lambda d \sum_{j=0}^{\bar{X}_s}(1-\hat{s}_j) \\
& + \lambda d \sum_{j=0}^{\bar{X}_s-1} |\epsilon_j| - |\epsilon_{\bar{X}_s}| + |\epsilon_{\bar{X}_s+1}|.
\end{aligned} \tag{C.14}$$

Next for $\bar{X}_s < i < \bar{Q}_{BF}$, 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

$$\frac{\mathrm{d}|\epsilon_i|}{\mathrm{d}t} \leq \lambda d(1-\hat{s}_i) - |\epsilon_i| + |\epsilon_{i+1}| \quad \forall\ \bar{X}_s < i < \bar{Q}_{BF}. \tag{C.15}$$

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

$$\frac{\mathrm{d}|\epsilon_{\bar{Q}_{BF}}|}{\mathrm{d}t} \leq \lambda - \lambda d \sum_{j=0}^{\bar{Q}_{BF}-1} (1-\hat{s}_j) - |\epsilon_{\bar{Q}_{BF}}| + |\epsilon_{\bar{Q}_{BF}+1}|. \tag{C.16}$$

Summing inequalities (C.12) - (C.16), we

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

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

$$\begin{aligned}\frac{\mathrm{d}|\epsilon_{\bar{X}_s}|}{\mathrm{d}t} =& \lambda - \lambda d \sum_{j=0}^{\bar{X}_s-1} (1-s_j) - s_{\bar{X}_s} + s_{\bar{X}_s+1} \\ =& \lambda - \lambda d \sum_{j=0}^{\bar{X}_s} (1-s_j) + \lambda d(1-s_{\bar{X}_s}) - s_{\bar{X}_s} + s_{\bar{X}_s+1} \\ \leq_{(a)}& \lambda d - (1+\lambda d) s_{\bar{X}_s} + s_{\bar{X}_s+1} \\ \leq& -(1+\lambda d)|\epsilon_{\bar{X}_s}| + |\epsilon_{\bar{X}_s+1}|,\end{aligned} \tag{C.17}$$

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

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

when $\bar{X}_s < \bar{Q}_{BF}$.

The summation of (C.15) and (C.16) yields that

$$\begin{aligned}\sum_{i=\bar{X}_s+1}^{\bar{Q}_{BF}} \frac{\mathrm{d}|\epsilon_i|}{\mathrm{d}t} \leq & \lambda - \lambda d \sum_{j=0}^{\bar{X}_s}(1-\hat{s}_j) - |\epsilon_{\bar{X}_s+1}| + |\epsilon_{\bar{Q}_{BF}+1}| \\ \leq & \lambda - \lambda d \sum_{j=0}^{\bar{X}_s}(1-s_j) \\ & - \lambda d \sum_{j=0}^{\bar{X}_s} \epsilon_j - |\epsilon_{\bar{X}_s+1}| + |\epsilon_{\bar{Q}_{BF}+1}| \\ \leq & \lambda d \sum_{j=0}^{\bar{X}_s} |\epsilon_j| - |\epsilon_{\bar{X}_s+1}| + |\epsilon_{\bar{Q}_{BF}+1}|, \end{aligned} \tag{C.18}$$

where the last inequality holds due to the definition of $\bar{X}_s$. The summation of (C.12), (C.13), (C.17) and (C.18) yields

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

In a summary, we have shown that

$$\frac{\mathrm{d}V(t)}{\mathrm{d}t} \begin{cases} \leq 0, & \text{if } s(t) \neq \hat{s} \\ = 0, & \text{otherwise.} \end{cases} \tag{C.19}$$

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^* exists, we can further obtain

$$\frac{\mathrm{d}V(t)}{\mathrm{d}t} \leq -|\epsilon_1| - |\epsilon_{i^*}|, \tag{C.20}$$

which implies

$$\frac{\mathrm{d}V(t)}{\mathrm{d}t} \begin{cases} = 0, & \text{if } s(t) = \hat{s}. \\ < 0, & \text{if } s_i(t) < \hat{s}_i \text{ for some } i \\ < 0, & \text{if } s_1(t) > \hat{s}_1 \\ \leq 0, & \text{if } s_i(t) \geq \hat{s}_i \ \forall i \text{ and } s_1(t) = \hat{s}_1. \end{cases} \tag{C.21}$$

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

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

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

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

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

we have

$$|s(t)| \leq r + |\hat{s}| \quad \forall\, t.$$

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

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

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

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

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

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

Furthermore, given $s(t)$ such that $s_1(t) = \hat{s}_1$ and $s_i(t) \geq \hat{s}_i (i \geq 2)$, it can be easily shown that $s_1(t + \delta t) > \hat{s}_1$ for a sufficiently small δt unless $s(t) = \hat{s}$. The result can be proved by following the idea of LaSalle's invariance principle (see Khalil (2001)).

C.4 Proof of Theorem 10

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

$$\Gamma^{(n)}(t) = \Gamma^{(n)}(0) + \sum_{L \in \mathbb{N}^\infty} L N_L \left(\int_0^t R_L^{(n)}(\Gamma^{(n)}(u))\, \mathrm{d}u \right), \tag{C.22}$$

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

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

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

$$R_L^{(n)}(\Gamma^{(n)}) = \Gamma_1^{(n)} - \Gamma_2^{(n)}$$

because there are $\Gamma_1^{(n)} - \Gamma_2^{(n)}$ servers with queue size 1. Dividing by n on both sides of equation (C.22), we get

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

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 $\gamma^{(n)}(0)$ converges to $s(0)$ 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

$$\gamma_i^{(n)}(u) = \sum_{j=i}^{\infty} \pi_j^{(n)}(u) \leq \frac{C_\alpha}{i} \quad \forall\ 0 \leq u \leq t, \quad \forall i, \tag{C.23}$$

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

$$\gamma_i^{(n)}(u) \leq \frac{1}{2}\left(1 - \frac{1}{d}\right) \quad \forall\ 0 \leq u \leq t, \quad \forall i \geq k. \tag{C.24}$$

Next we define the following four sets:

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

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

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

Define $s(t)$ to be the solution of the differential equations (4.7) with initial condition $s(0)$, and $F(s)$ such that the nonlinear differential equations in (4.7) are given by

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

Following the idea behind the proof of Kurtz's theorem (see Draief and Massoulié

(2010) for an easy exposition), we have

$$\sup_{0\le u\le t}\left|\gamma^{(n)}(u)-s(u)\right| \tag{C.25}$$

$$\le\left|\gamma^{(n)}(0)-s(0)\right| \tag{C.26}$$

$$+\sup_{0\le u\le t}\left|\sum_{L\notin(\mathcal{L}_n^+\cup\mathcal{L}_n^-)}\frac{L}{n}N_L\left(\int_0^u R_L^{(n)}(n\gamma^{(n)}(\tau))\,\mathrm{d}\tau\right)\right| \tag{C.27}$$

$$+\sup_{0\le u\le t}\left|\sum_{L\in\mathcal{L}_n^+\cup\mathcal{L}_n^-}\frac{L}{n}\bar{N}_L\left(\int_0^u R_L^{(n)}(n\gamma^{(n)}(\tau))\,\mathrm{d}\tau\right)\right| \tag{C.28}$$

$$+\sup_{0\le u\le t}\left|\sum_{L\in\mathcal{L}_n^+\cup\mathcal{L}_n^-}\frac{L}{n}\int_0^u R_L^{(n)}(n\gamma^{(n)}(\tau))\,\mathrm{d}\tau - \int_0^u F(\gamma^{(n)}(\tau))\,\mathrm{d}\tau\right| \tag{C.29}$$

$$+\sup_{0\le u\le t}\left|\int_0^u F(\gamma^{(n)}(\tau))\,\mathrm{d}\tau-\int_0^u F(s(\tau))\,\mathrm{d}\tau\right|. \tag{C.30}$$

According to Lemmas 12-14, we obtain that there exists $\bar{n}$ such that for any $n\ge\bar{n}$,

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

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

Let

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

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

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

where the last inequality holds because $F(s)$ is Lipschitz as shown in Lemma 15. By Gronwall's inequality we have $\left|\gamma^{(n)}(u) - s(u)\right| \le 4\delta e^{Mu}$ for any $u \in [0,t]$. Thus

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

as $n \to \infty$.

Lemma 12.

$$\mathbb{P}\left((C.27) > \delta\right) \le \frac{\lambda t}{\delta} e^{-\frac{(d-1)^2}{2(d+1)}m} + \frac{\lambda t C_\alpha}{\delta m} + 2(1 - \mathbb{P}\left(\mathcal{B}_{n,\alpha}\right)).$$

Proof. Note that $L \in \mathcal{T}_n^+ \setminus \mathcal{L}_n^+$ occurs when a task is dispatched to a queue with size at least k. Under condition (C.24), when a batch arrival occurs,

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

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

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

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

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

□

Lemma 13.

$$\mathbb{P}\left((C.28)>\delta\right)\le 4m^k\max\left\{e^{-\frac{n}{m}\lambda th(\frac{\delta}{(m+1)^k\lambda t})},e^{-nth(\frac{\delta}{2m^k t})}\right\}.$$

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

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

where the last inequality follows from Proposition 5.2 in Draief and Massoulié (2010). Similarly, for $L \in \mathcal{L}_n^-$,

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

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

$$\mathbb{P}\left((C.28) > \delta\right) \le 4m^k \max\left\{e^{-\frac{n}{m}\lambda th(\frac{\delta}{(m+1)^k\lambda t})}, e^{-nth(\frac{\delta}{2m^k t})}\right\}.$$

□

Lemma 14. *There exists $\bar{n}$ such that for any $n \ge \bar{n}$,*

$$\mathbb{P}\left((C.29) > \delta\right) \le 1 - \mathbb{P}\left(\mathcal{B}_{n,\alpha}\right).$$

Proof. To study (C.29) under condition (C.23), we define

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

and consider

$$\left|F^{(n)}(\gamma) - F(\gamma)\right| = \sum_i \left|F_i^{(n)}(\gamma) - F_i(\gamma)\right|. \tag{C.31}$$

We divide the analysis into the following cases:

- For $i > m$, $L_i = 0$ for any $L \in \mathcal{L}_n^+ \cup \mathcal{L}_n^-$, which implies $F_i^{(n)}(\gamma) = 0$ and

$$\sum_{i>m}\left|F_i^{(n)}(\gamma) - F_i(\gamma)\right| = \sum_{i>m} F_i(\gamma) = \gamma_{m+1} \le \frac{C_\alpha}{m}.$$

- For $m \ge i > k$,

$$F_i^{(n)}(\gamma) \;=\; -\gamma_i + \gamma_{i+1},$$

  which implies that

$$\left|F_i^{(n)}(\gamma) - F_i(\gamma)\right| = 0.$$

- For $k \geq i$,

$$\begin{aligned} F_i^{(n)}(\gamma) &= \frac{1}{n}\left(\frac{\lambda n}{m}\,\mathbb{E}[D_i|\gamma] - n\gamma_i + n\gamma_{i+1}\right) \\ &= \lambda d\,\mathbb{E}\left[\frac{D_i}{dm}\middle|\,\gamma\right] - \gamma_i + \gamma_{i+1}, \end{aligned}$$

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,

$$\left|F_i^{(n)}(\gamma) - F_i(\gamma)\right| = \lambda d\,\mathbb{E}\left[\frac{D_i}{dm}\middle|\,\gamma\right].$$

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 $(j \leq i)$. Specifically,

$$D_i = \min\left\{dm - Z_i, \left(m - \sum_{j=0}^{i-1}(dm - Z_j)\right)^+\right\}. \tag{C.32}$$

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

$$\mathbb{P}\left(|Z_i - \gamma_i dm| \geq \sqrt{m\log m}\right) \leq 2e^{-2\frac{\log m}{d}} = \frac{2}{m^{2/d}},$$

which implies that

$$\mathbb{P}\left(|Z_i - \gamma_i dm| \leq \sqrt{m\log m} \quad \forall i \leq k\right) \geq 1 - \frac{2k}{m^{2/d}}.$$

Given $|Z_i - \gamma_i dm| \leq \sqrt{m\log m}$ for all $i \leq k$, we can obtain

$$\begin{aligned} &\left|\mathbb{E}\left[\frac{D_i}{dm}\middle|\,\gamma\right] - \min\left\{1 - \gamma_i, \left(\frac{1}{d} - \sum_{j=0}^{i-1}(1 - \gamma_j)\right)^+\right\}\right| \\ \leq &\frac{k\sqrt{\log m}}{d\sqrt{m}}. \end{aligned}$$

By summarizing the cases above, we obtain that under condition (C.23)

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

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

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

So for sufficient large n,

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

□

Lemma 15. *$F(s)$ is Lipschitz.*

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

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

Then

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

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

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

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

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

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

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

where the first inequality holds because

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

according to the definition of $\bar{X}_s$.

Combining the results above, we obtain that

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

Therefore, the lemma holds. □

C.5 Proof of Theorem 11

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

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

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

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

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

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

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

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

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

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

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

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

C.6 Proof of Lemma 6

Since the series $\sum_{i=1}^{\infty} \mathbb{E}\left[\tilde{\gamma}_i\right]$ is increasing and bounded above, in order to get uniform convergence of $\sum_{i=1}^{\infty} \mathbb{E}\left|\hat{\gamma}_i^{(n_k)} - \tilde{\gamma}_i\right|$ for all k it suffices to show the series $\sum_{i=1}^{\infty} \mathbb{E}\,\hat{\gamma}_i^{(n_k)}$ are uniformly convergent for all k.

Now consider the n_kth system and define a Lyapunov function on the queue-length space to be

$$V_b(q) = \sum_{j=1}^{n_k} ((q_j - b + 1)^+)^2,$$

where $b > 0$. Let $x, y \in \mathbb{N}^{n_k}$ denote the state of the Markov chains, and $q_{x,y}$ denote the transition rate from state x to state y. According to the Foster–Lyapunov theorem for continuous-time Markov chain (see, e.g., Theorem 9.1.8 in Srikant and Ying (2014)), we consider the drift

$$\mathcal{G}V_b(x) = \sum_{y \neq x} q_{x,y}(V_b(y) - V_b(x)),$$

where $\mathcal{G}$ is the generator of the CTMC. We first note that the drift for the departure at jth queue satisfies

$$q_{x,x-e_j}(V_b(x - e_j) - V_b(x)) \leq -2(x_j - b)^+.$$

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

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

where $(x_k \colon 1 \leq k \leq dm)$ are the sizes of the probed dm queues. In other words, given x and the set of dm probed servers, batch-filling minimizes $V_b(y)$. 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 x such that

$$x \in \Omega_b \triangleq \left\{ x \colon \sum_{j=1}^{n_k} \mathbb{1}_{x_j \le b-2} \ge n_k \frac{d+1}{2d} \right\}.$$

In other words, at least $(d+1)/2d$ fraction of servers have queue size at most $b-2$. Let $\tilde{q}_{x,y}$ be the transition rate under batch-sampling from state x to state y. For $\phi \subseteq \{1, 2, \ldots, n_k\}$ with $|\phi| = dm$, let $y_{x,\phi}$ be one possible state after probing the servers with indices ϕ at state x under batch-filling, and let $\tilde{y}_{x,\phi}$ be one possible state after probing the servers with indices ϕ at state x under batch-sampling. Then we have

$$\begin{aligned}
&\sum_{y \ge x} q_{x,y}(V_b(y) - V_b(x)) \\
&= \sum_{y \ge x} \frac{\lambda n}{m} \sum_{\phi} \mathbb{P}(\Phi = \phi)\, \mathbb{P}(Y = y \mid X = x, \Phi = \phi)(V_b(y) - V_b(x)) \\
&= \frac{\lambda n}{m} \sum_{\phi} \mathbb{P}(\Phi = \phi) \sum_{y \ge x} \mathbb{P}(Y = y \mid X = x, \Phi = \phi)(V_b(y) - V_b(x)) \\
&= \frac{\lambda n}{m} \sum_{\phi} \frac{1}{\binom{n_k}{dm}} \sum_{y \ge x} \mathbb{P}(Y = y \mid X = x, \Phi = \phi)(V_b(y_{x,\phi}) - V_b(x)) \\
&= \frac{\lambda n}{m\binom{n_k}{dm}} \sum_{\phi} (V_b(y_{x,\phi}) - V_b(x)) \\
&\le \frac{\lambda n}{m\binom{n_k}{dm}} \sum_{\phi} (V_b(\tilde{y}_{x,\phi}) - V_b(x)) \\
&= \frac{\lambda n}{m} \sum_{\phi} \mathbb{P}(\Phi = \phi) \sum_{y \ge x} \mathbb{P}(\tilde{Y} = y \mid X = x, \Phi = \phi)(V_b(y) - V_b(x)) \\
&= \sum_{y \ge x} \tilde{q}_{x,y}(V_b(y) - V_b(x)),
\end{aligned}$$

where Φ is the random element of the uniformly probed indices, X is the state before the arrival, Y and $\tilde{Y}$ are the states after the arrival under batch-filling and batch-sampling, respectively.

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 (Hoeffding, 1963) for sampling without replacement, we get

$$\mathbb{P}(G_{b-2} < m) \leq c_1 e^{-\frac{(d-1)^2}{2d}m}.$$

Therefore, we conclude that

$$\begin{aligned}
&\sum_{y \geq x} \tilde{q}_{x,y} \left(V_b(y) - V_b(x)\right) \\
&\leq \sum_j \frac{\lambda n_k}{m} \frac{dm}{n_k} c_1 e^{-\frac{(d-1)^2}{2d}m} \left(2(x_j - b + 1) + 1\right)^+ \\
&\leq \lambda d c_1 e^{-\frac{(d-1)^2}{2d}m} \sum_j \left(2(x_j - b)^+ + 3\right).
\end{aligned}$$

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 x such that $x \neq \Omega_b$; i.e.,

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

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 5, we have

$$\begin{aligned}
&\sum_{y \geq x} q_{x,y} \left(V_b(y) - V_b(x)\right) \\
&\leq \frac{\lambda n_k}{m} \frac{m}{n_k} \sum_{j=1}^{n_k} \left(((x_j - b + 2)^+)^2 - ((x_j - b + 1)^+)^2\right) \\
&\leq \lambda \sum_{j=1}^{n_k} \left(2(x_j - b)^+ + 3\right) \\
&= 2\lambda \sum_{j=1}^{n_k} (x_j - b)^+ + 3\lambda n_k.
\end{aligned}$$

Combining the results above, we have that for any x,

$$\begin{aligned}\mathcal{G}V_b(x) &= \sum_{y\neq x} q_{x,y}\left(V_b(y) - V_b(x)\right)\\ &\leq \sum_j -2\left(1-\lambda\max\{1, c_1 d e^{-\frac{(d-1)^2}{2d}m}\}\right)(x_j - b)^+\\ &\quad + 3n_k\lambda c_1 d e^{-\frac{(d-1)^2}{2d}m}\mathbb{1}_{x\in\Omega_b} + 3\lambda n_k \mathbb{1}_{x\notin\Omega_b}.\end{aligned}$$

Recall that the Markov chain is positive recurrent according to Theorem 5. We have

$$\mathbb{E}[\mathcal{G}V_b(\hat{Q}^{(n_k)})] = 0,$$

which implies that

$$\mathbb{E}\left(\frac{1}{n_k}\sum_{j=1}^{n_k}(\hat{Q}_j^{(n_k)} - b)^+\right) \leq \frac{3\lambda c_1 d e^{-\frac{(d-1)^2}{2d}m}\,\mathbb{P}(\hat{Q}^{(n_k)}\in\Omega_b) + 3\lambda\,\mathbb{P}(\hat{Q}^{(n_k)}\notin\Omega_b)}{2\left(1-\lambda\max\{1, c_1 d e^{-\frac{(d-1)^2}{2d}m}\}\right)}.$$

Since $m = \Theta(\log n)$, for sufficiently large n we have $1 \geq c_1 d e^{-\frac{(d-1)^2}{2d}m}$. So for any $\epsilon > 0$ there exists n_ϵ such that for any $n \geq n_\epsilon$,

$$\frac{3\lambda c_1 d e^{-\frac{(d-1)^2}{2d}m}\,\mathbb{P}(\hat{Q}^{(n_k)}\in\Omega_b)}{2\left(1-\lambda\max\{1, c_1 d e^{-\frac{(d-1)^2}{2d}m}\}\right)} \leq \frac{\epsilon}{2}.$$

Also note by Theorem 5 we have $\sum_{i=1}^{\infty}\mathbb{E}\,\hat{\gamma}_i^{n_k} \leq c$, which implies $\mathbb{E}\,\hat{\gamma}_i^{n_k} \leq \frac{c}{i}$ for any i and k. Then by Markov's inequality,

$$\begin{aligned}\mathbb{P}(\hat{Q}^{(n_k)}\notin\Omega_b) &= \mathbb{P}\left(\hat{\gamma}_{b-1}^{(n_k)} \geq \frac{d-1}{2d}\right)\\ &\leq \frac{c}{b-1}\frac{2d}{d-1}.\end{aligned}$$

So there exists b_ϵ such that for any $b \geq b_\epsilon$ and any $n \geq n_\epsilon$,

$$\frac{3\lambda\,\mathbb{P}(\hat{Q}^{(n_k)}\notin\Omega_b)}{2\left(1-\lambda\max\{1, c_1 d e^{-\frac{(d-1)^2}{2d}m}\}\right)} \leq \frac{\epsilon}{2}.$$

So

$$\sum_{i=b+1}^{\infty}\mathbb{E}\,\hat{\gamma}_i^{(n_k)} = \mathbb{E}\left(\frac{1}{n_k}\sum_{j=1}^{n_k}(\hat{Q}_j^{(n_k)} - b)^+\right) \leq \epsilon$$

for $n \geq n_\epsilon$ and $b \geq b_\epsilon$. Hence the series $\sum_{i=1}^{\infty}\mathbb{E}\,\hat{\gamma}_i^{(n_k)}$ are uniformly convergent for all k.

C.7 Proof of Corollary 5

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 $(Q_1^{(n)}(t), Q_2^{(n)}(t), \eta^{(n)}(t))$, where

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

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

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

i.e., the stationary distribution of the Markov chain. For the nth system, the global balance equation for a given state (x, y, η) is

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

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

$$\begin{aligned}&\sum_{\eta} \sum_{(\tilde{x},\tilde{y},\tilde{\eta})\neq(x,y,\eta)} \pi^{(n)}(x, y, \eta)\, r^{(n)}_{(x,y,\eta)}(\tilde{x}, \tilde{y}, \tilde{\eta}) \\ =&\sum_{\eta} \sum_{(\tilde{x},\tilde{y},\tilde{\eta})\neq(x,y,\eta)} \pi^{(n)}(\tilde{x}, \tilde{y}, \tilde{\eta})\, r^{(n)}_{(\tilde{x},\tilde{y},\tilde{\eta})}(x, y, \eta). \qquad \text{(C.33)}\end{aligned}$$

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

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

by exchanging the symbols η and $\tilde{\eta}$. 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

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

which implies that

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

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

$$\sum_{\eta} \sum_{\tilde{x}<x,\tilde{y}<y} \pi^{(n)}(\tilde{x},\tilde{y},\tilde{\eta})\, r^{(n)}_{(\tilde{x},\tilde{y},\tilde{\eta})}(x,y,\eta) = O\left(\frac{m}{n}\right). \tag{C.34}$$

Note that

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

so

$$\begin{aligned} &\sum_{\eta} \pi^{(n)}(x,y,\eta)\, r^{(n)}_{(x,y,\eta)}(x-1,y,\eta) \\ =&\sum_{\eta} \pi^{(n)}(x,y,\eta)\, r^{(n)}_{(x,y,\eta)}(x,y-1,\eta) \\ =&\pi^{(n)}(x,y), \end{aligned} \tag{C.35}$$

$$\sum_{\eta} \pi^{(n)}(x+1,y,\eta)\, r^{(n)}_{(x+1,y,\eta)}(x,y,\eta) = \pi^{(n)}(x+1,y), \tag{C.36}$$

and

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

Now we consider

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

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

Note that

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

which implies that

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

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

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

and

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

where $q_{x,\tilde{x}}(\hat{\gamma})$ and $\bar{Q}_{BF}$ are defined in Lemma 4. Since $0\le\pi^{(n)}(x,y)\le 1$ and

$0 \leq \mathbb{E}_\eta \left[r^{(n)}_{(x,y,\eta)}(\tilde{x}, y) \middle| x, y \right] \leq d\lambda$, we can conclude that

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

Similarly, we have

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

Summarizing the results above, (C.33) implies that

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

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 4, and the unique solution therefore is $\pi(x,y) = \hat{\pi}_x \hat{\pi}_y$ for $\hat{\pi}$ defined in (4.4). This means that queue 1 and queue 2 are independent in the large-system limit.