https://pubsonline.informs.org/journal/stsy

STOCHASTIC SYSTEMS
Articles in Advance, pp. 1–44
ISSN 1946-5238 (online)

Exploiting Data Locality to Improve Performance of Heterogeneous Server Clusters

Zhisheng Zhao,[a,*] Debankur Mukherjee,[a] Ruoyu Wu[b]

[a] H. Milton Stewart School of Industrial and Systems Engineering, Georgia Institute of Technology, Atlanta, Georgia 30332; [b] Department of Mathematics, Iowa State University, Ames, Iowa 50011

*Corresponding author

Contact: zhisheng@gatech.edu, https://orcid.org/0000-0003-0720-1895 (ZZ); debankur.mukherjee@isye.gatech.edu, https://orcid.org/0000-0003-1678-4893 (DM); ruoyu@iastate.edu, https://orcid.org/0000-0003-2649-6086 (RW)

Received: November 29, 2022
Revised: June 7, 2023
Accepted: September 22, 2023
Published Online in Articles in Advance: February 6, 2024

https://doi.org/10.1287/stsy.2022.0040

Abstract. We consider load balancing in large-scale heterogeneous server systems in the presence of data locality that imposes constraints on which tasks can be assigned to which servers. The constraints are naturally captured by a bipartite graph between the servers and the dispatchers handling assignments of various arrival flows. When a task arrives, the corresponding dispatcher assigns it to a server with the shortest queue among $d \geq 2$ randomly selected servers obeying these constraints. Server processing speeds are heterogeneous, and they depend on the server type. For a broad class of bipartite graphs, we characterize the limit of the appropriately scaled occupancy process, both on the process level and in steady state, as the system size becomes large. Using such a characterization, we show that imposing data locality constraints can significantly improve the performance of heterogeneous systems. This is in stark contrast to either heterogeneous servers in a full flexible system or data locality constraints in systems with homogeneous servers, both of which have been observed to degrade the system performance. Extensive numerical experiments corroborate the theoretical results.

Funding: This work was partially supported by the National Science Foundation [CCF. 07/2021–06/2024].

Keywords: heterogeneous load-balancing system • data locality • compatibility constraint • power of two • mean-field • McKean–Vlasov process

1. Introduction

Over the last two decades, large-scale load balancing has emerged as a fundamental research problem. In simple terms, the goal is to investigate how to efficiently allocate tasks in large-scale service systems, such as data centers and cloud networks. As modern data centers continue to process massive amounts of data with increasingly stringent processing time requirements, the need for more efficient and scalable, dynamic load-balancing algorithms is greater than ever. The study of scalable load-balancing algorithms started with the seminal works of Adler et al. (1995), Mitzenmacher (1996a, b), and Vvedenskaya et al. (1996), in which the popular "power-of-d choices" or the join-shortest-queue (d) (JSQ(d)) algorithm was introduced. Here, a canonical model was considered that consists of N identical parallel servers, each serving a dedicated queue of tasks. Arriving tasks are routed to the shortest of $d \geq 2$ randomly selected queues by a centralized dispatcher, irrevocably and instantaneously, at the time of arrival. Since then, this model has received significant attention from the research community, and we have seen tremendous progress in our understanding of the performance of various algorithms; see van der Boor et al. (2022) for a recent survey.

Despite this phenomenal progress, when it comes to modern large-scale systems, much of the existing wisdom can be observed to be false. This is primarily because of the fact that the classical model fails to capture two of the most significant factors that impact the performance of these systems. The first is data locality constraints. In simple terms, it means that tasks of a particular type can only be routed to a small subset of servers that are equipped with the appropriate resources to execute them (Tsitsiklis and Xu 2017, Tirmazi et al. 2020, Weng et al. 2020, Rutten and Mukherjee 2022). For example, an image classification request must be routed to a server that is trained with

appropriate machine learning models, such as deep convolutional neural network. Also, in online video services like Netflix and YouTube, users' requests may only be routed to servers that are equipped with the required data (e.g., movies, music). The classical model ignores this effect and assumes *full flexibility*: that is, that any task can be assigned to any server in the system. In the presence of data locality constraints, the delay performance of the system may degrade drastically as compared with fully flexible systems. The second is heterogeneity in service rates. Servers in any modern large-scale server clusters do not process tasks at equal speeds. This heterogeneity of the service rates is a major bottleneck in implementing the existing heuristics of the classical model. For example, if there are two groups of servers in the system, one faster and the other slower, then popular dynamic algorithms like JSQ(d), which has a provably excellent delay performance when all server speeds are identical, can be observed to be unstable (i.e., their queue lengths blow up) (Mukhopadhyay and Mazumdar 2016, Mukhopadhyay et al. 2016, Gardner et al. 2021, Hurtado-Lange and Maguluri 2021). In other words, heterogeneity shrinks the stability region as formally established in Hurtado-Lange and Maguluri (2021). This happens simply because if all the servers are treated equally, then the slower server pool may receive a higher flow of arrivals than what it can process.

1.1. Takeaway

In summary, both data locality and heterogeneity of server speeds may significantly degrade the system performance. The main contribution of the current work is to establish that when these two aspects are considered together, then the performance can in fact be drastically improved. That is, if servers are heterogeneous, then efficiently designing the data locality constraints (by appropriately placing the resource files in the server network) can regain the full stability region, which was shrunk for fully flexible systems. Moreover, we also establish that carefully designed data locality constraints can ensure the celebrated double-exponential decay of tail probability of the steady-state queue-length distribution even for the heterogeneous systems.

1.2. Our Contributions

Motivated by this, in the current paper, we consider a bipartite graph model for large-scale load-balancing systems, which has recently gained popularity in the research community. In this model, a bipartite graph between the servers and task types describes the *compatibility* between the two, where an edge represents the server's ability to process the corresponding task type. This encompasses the classical *full-flexibility* models as those having a complete bipartite compatibility graph. An immediate difficulty of the new model is that when the graph is nontrivial (i.e., not a collection of isolated pairs or a complete bipartite graph), the mean-field techniques break down. This is because, the queues no longer remain exchangeable, making the aggregate processes, such as the vector of number of servers with queue length i with $i = 0, 1, 2, \ldots$, non-Markovian. In addition, we also consider that each dispatcher handles the arrival flow of one of K possible task clusters and that there are M server types. The rate of service at a server depends on its type. Throughout the paper, the key quantity of interest will be the *global occupancy process* $\mathbf{q}^N(t) = (q^N_{m,l}(t), m = 1, \ldots, M, l \geq 1)$, where $q^N_{m,l}(t)$ represents the fraction of servers of type m with queue length at least l at time t in the Nth system with N servers, and we will look at the large-system asymptotic regime: $N \to \infty$.

Because of the compatibility constraints, the servers become nonexchangeable, even if they belong to the same type. This causes most of the existing frameworks to break down; see, for example, Mitzenmacher (1996b), Ethier and Kurtz (2009), and Stolyar (2015). To characterize the process-level limit of the queue-length process, we resort to the theory of weakly interacting particle systems and asymptotically couple the evolution of the N-dimensional vector of queue lengths with an appropriately defined infinite system of independent *McKean–Vlasov processes*; see, for example, Sznitman (1991) and Méléard (1996). We also show the asymptotic independence of any finite number of queue-length processes, also known as the *propagation of chaos* property. This convergence of the queue-length processes (in L_2 sense) is then used to establish the transient convergence of the occupancy process. One downside of the convergence is that it depends on the assumption that the initial queue lengths within each set of servers of the same type are independent and identically distributed (i.i.d.) and are independent across the set of servers of different types. Because of this assumption, this convergence result cannot be used to establish the interchange of $t \to \infty$ and $N \to \infty$ limits, which is crucial in studying the limit of steady states.

To overcome this issue, we use the framework of Rutten and Mukherjee (2022), recently introduced in the context of homogeneous systems. Here, a notion called *proportional sparsity* for graph sequences was introduced, which ensures that the empirical queue-length distribution within the set of compatible servers of any dispatcher is close to the empirical queue-length distribution of the entire system. This was used in Rutten and Mukherjee (2022) to construct conditions on graphs that match the performance of a fully flexible system. In the current setup, however, this notion is inadequate because our goal is not to match the performance of the fully flexible system (which is usually poor under heterogeneity). That is why we extend this notion to what we call the *clustered proportional sparsity* for a sequence of graphs with increasing size to accommodate the heterogeneous

systems. The clustered proportional sparsity property allows us to construct a stochastic coupling between the system and another intermediate system whose task allocation is done by a carefully constructed algorithm called global weighted shortest queue (GWSQ(d)) (Algorithm 1). This coupling with the intermediate system, along with clustered proportional sparsity, helps us establish that if the initial occupancy of two systems is close, then the distance (in the ℓ_1-norm) between their global occupancy remains small uniformly over any finite time interval. In turn, it implies that their limits of the global occupancy systems are the same. As a consequence, we can remove the i.i.d. assumption of the initial queue lengths because this guarantees that under clustered proportional sparsity, the convergence of the occupancy process depends only on the initial occupancy and *not* on how the individual queues are distributed.

The process-level limit result shows that the transient limit of the occupancy process can be described as a system of ODEs that depend on various graph parameters. Next, we also show that the interchange of limits holds and that the sequence of occupancy states in stationarity converges weakly to the unique fixed point of the ordinary differential equation (ODE). One celebrated feature of the classical JSQ(d) policy for homogeneous systems under full flexibility is that the steady-state queue length decays doubly exponentially as $\lambda^{(d^i-1)/(d-1)}$, where $\lambda \in (0,1)$ is the load per server (Mitzenmacher 1996b, Vvedenskaya et al. 1996). We establish this double-exponential decay property for the heterogeneous system.

It is worthwhile to note that the strength of the results lies in that they hold for arbitrary deterministic sequences of graphs satisfying certain properties. However, we show that all these properties are satisfied almost surely by a sequence of inhomogeneous random graphs (IRGs) with parameters prescribed by the theorems. This makes it easy to design graphs with the desired favorable properties.

1.3. Related Works

The research on task allocation systems with limited flexibility can be traced back to the works of Foss and Chernova (1998) and Turner (1998). Of particular importance to the current work, Foss and Chernova (1998) considered stability properties of the system using the fluid model. Later, Bramson (2011) generalized some parts of the results in Foss and Chernova (1998) to a broad class of JSQ-type systems, including the JSQ(d) policy, via the Lyapunov function approach. Stolyar (2005) considered optimal routing in an output-queued flexible server system, which is essentially the bipartite graph model for the load-balancing system. Here, the author considered a system with a fixed number of servers and dispatchers in the conventional heavy traffic regime and proposed a routing policy that is optimal in terms of server workload. Recently, Cruise et al. (2020) considered load-balancing problems on hypergraphs and proved their stability conditions. The works, however, did not aim to precisely characterize the system performance in the large-scale scenario.

The analysis in the large-scale scenario became prominent in the last decade, with the emergence of its applications to load balancing in data centers and cloud networks. In the full-flexibility setup, the analysis of heterogeneous server systems gained some attention. In this case, Stolyar (2015, 2017) studied the zero-queueing property of the join-idle-queue policy, Mukhopadhyay and Mazumdar (2016) and Mukhopadhyay et al. (2016) analyzed the JSQ(d) policy in heterogeneous systems with processor-sharing service discipline, Hurtado-Lange and Maguluri (2021) studied the throughput and delay optimality properties of JSQ(d), and Bhambay and Mukhopadhyay (2022) studied a speed-aware JSQ policy. The works on the JSQ(d) policy observe that the stability region shrinks if the dispatcher applies the JSQ(d) policy blindly. One way to mitigate this performance degradation is to take the server speeds into consideration while sampling servers or while assigning tasks to the sampled servers. Such a "hybrid JSQ(d)" scheme is able to recover the stability region. The current work can be contrasted with this approach. First, in the presence of data locality, both the server speeds and the underlying compatibility constraints need to be taken into account during the sampling procedure, and the approach becomes significantly more complicated. Second, we show how exploiting the data locality, the blind JSQ(d) policy can recover the stability region and even achieve the double-exponential decay of tail probabilities of the steady-state queue-length distribution. One advantage of the latter approach is that the dispatchers can be oblivious to the server speeds, which reduces the implementation complexity and also, makes it robust against changes to the servers (e.g., when servers are added/removed).

Recently, Allmeier and Gast (2022) studied the application of (refined) mean-field approximations for heterogeneous systems. Their method is using an ODE to approximate the evolution of each server, and the error vanishes as the system scales. However, this method cannot be directly used in our case. Because of the bipartite compatibility graph structure, it is hard to capture the interactions between two servers, which means that we cannot write the transition rates of the underlying Markov chain as Allmeier and Gast (2022) does. Also, one important assumption in their work is the finite buffer, but we consider the infinite buffer case here.

The aspect of task-server compatibility constraints in large-scale load balancing and scheduling gained popularity only recently, as the data locality became prominent in data centers and cloud networks. This led to many works in this area (Tsitsiklis and Xu 2013, 2017; Gast 2015; Mukherjee et al. 2018a; Budhiraja et al. 2019; Weng et al. 2020; Rutten and Mukherjee 2022). All these works consider homogeneous processing speeds at the servers. The initial works of Turner (1998) and Gast (2015) focused on certain fixed-degree graphs and showed that the flexibility to forward tasks to even a few neighbors with possibly shorter queues may significantly improve the waiting time performance as compared with dedicated arrival streams or a collection of independent M/M/1 queues that the system has a Poisson arrival process, an exponential service time distribution, and one server. Tsitsiklis and Xu (2013, 2017) considered asymptotic optimality properties of the bipartite graph topology in an input-queued, dynamic scheduling framework. Later, in the (output-queued) load-balancing setup, Mukherjee et al. (2018a) considered the JSQ policy, and Budhiraja et al. (2019) considered the transient analysis of the JSQ(d) policy on nonbipartite graphs. The goal in these papers was to provide sufficient conditions on the graph sequence to asymptotically match the performance of a complete graph. Here, we should mention that the nonbipartite graph model cannot be used to capture the data locality constraints. In the presence of data locality constraints, the analysis of the JSQ(d) policy for homogeneous systems, including both transient and interchange of limits, was performed by Rutten and Mukherjee (2022). Weng et al. (2020) is the first to consider the large-scale heterogeneous server model under data locality. They showed that the join-the-fastest-shortest-queue and join-the-fastest-idle-queue policies achieve asymptotic optimality for minimizing mean steady-state waiting time when the bipartite graph is sufficiently well connected. However, these results fall in the category of JSQ-type policies, where the asymptotic behavior is degenerate in the sense that the queue lengths at servers can be either zero or one. Naturally, the results and their analysis are very different from the JSQ(d)-type policies where queues of any length are possible.

1.4. Notations

Let $\mathbb{N}_0 = \mathbb{N} \cup \{0\}$. For a set S, its cardinality is denoted as $|S|$. For a polish space $\mathcal{S}$, the space of right continuous functions with left limits from $[0,\infty)$ to $\mathcal{S}$ is denoted as $\mathbb{D}([0,\infty),\mathcal{S})$, endowed with the Skorokhod topology. The distribution of $\mathcal{S}$-valued random variable X will be denoted as $\mathcal{L}(X)$. For a function $f : [0,\infty) \to \mathbb{R}$, let $\|f\|_{*,t} := \sup_{0\le s\le t}|f(s)|$. The distribution of $\mathcal{S}$-valued random variable X will be denoted as $\mathcal{L}(X)$. For $x \in \mathcal{S}$, the Dirac measure at the point x is denoted as δ_x. $\|\cdot\|_p$ represents the ℓ_p-norm. Define $\binom{X}{Y} = \frac{X(X-1)\cdots(X-Y+1)}{Y!}$ if $X \ge Y$ and is zero otherwise. RHS is the acronym of right-hand side.

2. Model Description

The model for large-scale systems with limited flexibility was considered by Tsitsiklis and Xu (2013, 2017) in the context of scheduling algorithms for input-queued systems. Subsequently, it was considered in Mukherjee et al. (2018a), Budhiraja et al. (2019), Weng et al. (2020), and Rutten and Mukherjee (2022) for output-queued load-balancing systems. Let $G^N = (W^N, V^N, E^N)$ be a system with N single servers, each serving its own queue, and $W(N)$ dispatchers, each handling the assignment of tasks of one type, where $W^N = \{1,\dots,W(N)\}$ and $V^N = \{1,\dots,N\}$ denote the sets of dispatchers and servers, respectively. We will interchangeably use the terms task type and dispatcher type throughout the article. Similar to Tsitsiklis and Xu (2013, 2017), we assume that $\lim_{N\to\infty} W(N)/N = \zeta$, where $\zeta > 0$ is a constant. The set $E^N \subseteq W^N \times V^N$ of edges represents hard compatibility between the dispatchers and servers in the Nth system. In other words, tasks of type i can be assigned to a server j if and only if $(i,j) \in E^N$. Tasks arriving at a dispatcher must be assigned instantaneously and irrevocably to one of the *compatible* servers.

- Dispatcher clusters. Each dispatcher belongs to one of K possible clusters labeled in $\mathcal{K} = \{1,\dots,K\}$. Let W_k^N denote the set of all dispatchers in the kth cluster. As $N \to \infty$, assume that $|W_k^N|/W(N) \to w_k \in (0,1)$ for $k \in \mathcal{K}$ with $\sum_{k=1}^K w_k = 1$. Tasks arrive at each dispatcher as an independent Poisson process with rate λ. Note that dispatchers in the same cluster may not have the same set of compatible servers.
- Server types. Based on its processing capability, each server belongs to one of M possible types labeled in $\mathcal{M} = \{1,\dots,M\}$. The processing time at a type-m server is exponentially distributed with mean $1/u_m$, where u_m is a positive constant. Let V_m^N denote the set of type-m servers, and as $N \to \infty$, $|V_m^N|/N \to v_m \in (0,1)$ for $m \in \mathcal{M}$ with $\sum_{m=1}^M v_m = 1$. Throughout, we will assume that asymptotically, the system has sufficient service capacity in the sense that

$$\lambda\zeta < \sum_{m\in\mathcal{M}} u_m v_m. \tag{2.1}$$

Note that the left- and right-hand sides represent the scaled total arrival rate and the scaled maximum departure rate, respectively.

For all the asymptotic results, we consider a general class of systems where the compatibility graph satisfies certain asymptotic criteria as specified in Condition 1. Define

$$\deg_w^N(i,m) = |\{j \in V_m^N : (i,j) \in E^N\}|, \quad i \in W^N, m \in \mathcal{M},$$
$$\deg_v^N(k,j) = |\{i \in W_k^N : (i,j) \in E^N\}|, \quad j \in V^N, k \in \mathcal{K}.$$

Namely, $\deg_w^N(i,m)$ is the number of the dispatcher i's neighboring servers whose type is $m \in \mathcal{M}$. Similarly, $\deg_v^N(k,j)$ is the number of the server j's neighboring dispatchers whose cluster is $k \in \mathcal{K}$.

Condition 1. *The sequence* $\{G^N\}_{N\geq 1}$ *satisfies the following.*

a. *For each* $k \in \mathcal{K}$ *and* $m \in \mathcal{M}$, *let* $E^N(k,m) = \{(i,j) \in W_k^N \times V_m^N : (i,j) \in E^N\}$,

$$\lim_{N\to\infty} \frac{|E^N(k,m)|}{|W_k^N| \times |V_m^N|} = p_{k,m} \in [0,1]. \tag{2.2}$$

We call the matrix $\mathbf{p} = (p_{k,m}, k \in \mathcal{K}, m \in \mathcal{M})$ *the compatibility matrix.*

b. *For each* $k \in \mathcal{K}$ *and* $m \in \mathcal{M}$,

$$\lim_{N\to\infty} \frac{\max_{i\in W_k^N} \deg_w^N(i,m)}{\min_{i\in W_k^N} \deg_w^N(i,m)} = 1, \quad \lim_{N\to\infty} \frac{\max_{j\in V_m^N} \deg_v^N(k,j)}{\min_{j\in V_m^N} \deg_v^N(k,j)} = 1.$$

Intuitively, the condition implies that the "asymptotic density" of edges between cluster-k dispatchers and type-m servers is given by $p_{k,m}$ and that, for each task-cluster-server-type pair, the servers have similar levels of flexibility. The classical, well-studied setup, where any task can be processed by any server, corresponds to the complete bipartite graph with $p_{k,m}^N = 1$, $\forall k \in \mathcal{K}, m \in \mathcal{M}$. In Section 3.5, we show that for any given $\mathbf{p} := (p_{k,m}, k \in \mathcal{K}, m \in \mathcal{M})$, a sequence of graphs satisfying Condition 1 can be obtained simply by putting edges suitably randomly. This is a certain class of *inhomogeneous random graphs*, which we call $\text{IRG}(\mathbf{p})$; see Definition 3 for details. In fact, the $\text{IRG}(\mathbf{p})$ sequence of graphs will be proved to satisfy the required conditions for all the results of this article to hold.

2.1. State Space

In the Nth system, let $X_j^N(t)$ be the number of tasks (including those in service) in the queue of server $j \in V^N$ at time t. Let $q_{m,l}^N(t)$ be the proportion of servers of type m with queue length at least l at time t, namely

$$q_{m,l}^N(t) := \frac{1}{|V_m^N|} \sum_{j\in V_m^N} \mathbb{1}_{(X_j^N(t)\geq l)}, \quad t \geq 0, m \in \mathcal{M}, l \in \mathbb{N}_0. \tag{2.3}$$

Let $\mathbf{q}^N(t) = (q_{m,l}^N(t), m \in \mathcal{M}, l \in \mathbb{N}_0)$. Then, $\mathbf{q}^N := \{\mathbf{q}^N(t)\}_{0\leq t<\infty}$ is a process with sample paths in $\mathbb{D}([0,\infty), \mathcal{S})$, where

$$\mathcal{S} := \left\{ \mathbf{q} \in [0,1]^{M\times\mathbb{N}_0} : q_{m,0} = 1, q_{m,l} \geq q_{m,l+1}, \text{ and } \sum_{l\in\mathbb{N}_0} q_{m,l} < \infty, \ \forall m \in \mathcal{M}, l \in \mathbb{N}_0 \right\}$$

is equipped with the ℓ_1-topology. Note that the space $\mathcal{S}$ is a complete metric space.

2.2. Local JSQ(d) Policy

For any fixed $d \geq 2$, each dispatcher uses the JSQ(d) policy (Mitzenmacher 1996b, Vvedenskaya et al. 1996) to assign the incoming tasks to servers. To describe the policy, define the *neighborhood of dispatcher* $i \in W^N$, $\mathcal{N}_w^N(i) := \{j \in V^N : (i,j) \in E^N\}$ with $\delta_i^N = |\mathcal{N}_w^N(i)|$. When a new task arrives at the dispatcher $i \in W^N$ with $\delta_i^N \geq d$, it is immediately assigned to the server with the shortest queue among d servers selected uniformly at random from $\mathcal{N}_w^N(i)$. Ties are broken uniformly at random. If $\delta_i^N < d$, then the task is assigned to one server selected from $\mathcal{N}_w^N(i)$ uniformly at random. This $\delta_i^N < d$ scenario is asymptotically not relevant for us because all the graphs that we will consider have diverging degrees as $N \to \infty$.

3. Main Results

3.1. Mitigating the Stability Issue

As discussed earlier, when the server speeds are heterogeneous, the fully flexible systems (with the complete bipartite compatibility graph) may not be stable under the JSQ(d) policy, even if we assume that the sufficient service capacity in (2.1) is satisfied. The next lemma provides a necessary and sufficient condition for ergodicity of the queue-length process. Recall $\delta_i^N = |\mathcal{N}_w^N(i)|$. For any fixed N, define

$$\rho^N := \max_{\substack{U\subseteq V^N\\ U\neq\emptyset}} \left\{ \left(\sum_{j\in U}\sum_{m\in\mathcal{M}} \mathbb{1}_{(j\in V_m^N)} u_m \right)^{-1} \sum_{i\in W^N} \left(\mathbb{1}_{(\delta_i^N\geq d)} \sum_{\substack{S\subseteq(U\cap\mathcal{N}_w^N(i)):\\ |S|=d}} \frac{\lambda}{\binom{\delta_i^N}{d}} + \mathbb{1}_{(\delta_i^N<d)} \frac{|U\cap\mathcal{N}_w^N(i)|}{\delta_i^N} \right) \right\}.$$

Lemma 1. *The queue-length process $(X_j^N(t))_{j\in V^N}$ under the local JSQ(d) policy is ergodic if and only if $\rho^N < 1$.*

The lemma is an immediate consequence of Foss and Chernova (1998, theorem 2.5); see also Bramson (2011). We omit its proof. Intuitively, $\rho^N < 1$ means that in the Nth system, for any subset U of servers with possibly long queues (compared with the rest servers), the total rate at which tasks are assigned to some server in this set must be less than the rate of departure from this set.

Because we are interested in large-N behavior, we will assume a certain asymptotic version of the stability criterion. This is fairly standard in the large-system analysis, as one would want to avoid the "heavy traffic" regime when $\rho^N \uparrow 1$ as $N\to\infty$. The behavior in the latter scenario is typically qualitatively different from the so-called "subcritical" regime as defined.

Definition 1 (Subcritical Regime). The sequence $\{G^N\}_N$ of systems defined is said to be in the *subcritical regime* with asymptotic load $\rho < 1$ if $\rho^N \to \rho < 1$, as $N\to\infty$.

Throughout this paper, we will assume that the sequence of systems under consideration is in the subcritical regime. From Lemma 1, it is immediate that if a sequence of systems is in subcritical regime, then its queue-length process is ergodic for all large-enough N. The potential nonergodicity of fully flexible, heterogeneous server clusters brings us to the question of when the sufficient service capacity in (2.1) is satisfied, whether we can design the underlying compatibility structure carefully so that the queue-length process is ergodic. In other words, can we regain the stability region? Proposition 1 shows that this is indeed the case. In some sense, this highlights the first-order improvements (i.e., in terms of stability properties) of a careful compatibility structure design in contrast to a fully flexible system.

The establishment of Proposition 1 relies on first building a simple criteria involving the system parameters, which for the sequence of systems satisfying Condition 1, ensures stability for all large-enough N (Lemma 2). Then, we show that given other parameters, a value of $(p_{k,m})_{k\in\mathcal{K},m\in\mathcal{M}}$ satisfying this criteria can be found by checking the feasibility region defined by M inequalities. Denote $\delta_k := \sum_{m\in\mathcal{M}} p_{k,m} v_m$ for each $k\in\mathcal{K}$.

Lemma 2. *Let $\{G^N\}_N$ be a sequence satisfying Condition 1. The sequence of systems is in subcritical regime if*

$$\frac{\lambda\zeta}{u_m}\sum_{k\in\mathcal{K}}\frac{w_k p_{k,m}}{\delta_k} < 1, \quad \text{for all } m\in\mathcal{M}. \tag{3.1}$$

Proposition 1. *Let the parameters λ,ζ,d and w_k,v_m,u_m, $k\in\mathcal{K}$, $m\in\mathcal{M}$, be such that (2.1) is satisfied. Then, there exists $(p_{k,m})_{k\in\mathcal{K},m\in\mathcal{M}}\in[0,1]^{K\times M}$ such that for any sequence of systems $\{G^N\}_{N\geq 1}$ satisfying Condition 1, the queue-length process $(X_j^N(t))_{j\in V^N}$ is ergodic for all N large enough. Moreover, such a $(p_{k,m})_{k\in\mathcal{K},m\in\mathcal{M}}$ can be obtained explicitly by solving a set of inequalities.*

The proof of Proposition 1 is provided in Appendix A.

In the following sections, we will demonstrate, in addition to the first-order improvements, how asymptotic queue-length distribution can be improved as well, for example, in terms of having a double-exponential decay of tail probabilities.

3.2. Process-Level Limit: i.i.d. Case

Our first main result characterizes the process-level limit of the queue-length process $(X_j^N, j\in V)$, as $N\to\infty$, when the starting states $\{X_j^N(0): j\in V_m^N\}$ are i.i.d. for all $m\in\mathcal{M}$ and independent across different m-values. When the

sequence of graphs $\{G^N\}_N$ satisfies a stronger condition, called *clustered proportional sparsity* (Definition 2), the i.i.d. condition can be removed. This is the content of Section 3.3.

Now, note that for a fixed $N \geq 1$, $\{X_j^N : j \in V_N\}$ is a system of N interacting stochastic processes, where interactions enter the dynamics through the local empirical measures of neighboring states (the precise dynamics are given in (4.3) and (4.4)). Exploiting tools from the theory of weakly interacting particles, we show in Theorem 1 that as the system size becomes large, queue-length processes converge weakly to those of an infinite system of independent *McKean–Vlasov processes* $\{X_j : j \in \mathbb{N}\}$ (see, e.g., Sznitman 1991, Méléard 1996). In fact, using a suitable coupling to be described in more detail in Section 4.1, the convergence holds in L_2. For ease of describing such processes and coupling, although we only assumed that certain fractions of servers are of certain types in the model description, it will be convenient to fix the type of each server $j \in \mathbb{N}$ in this subsection by defining a *membership map* $\mathbf{M} : \mathbb{N} \to \mathcal{M}$, so that $V_m^N = \{j \in V^N : \mathbf{M}(j) = m\}$ with $\lim_{N\to\infty} \frac{|V_m^N|}{N} = v_m$ and $V_m = \lim_{N\to\infty} V_m^N$ for each $m \in \mathcal{M}$. With such fixed server types and $X_j^N(0) \equiv X_j(0)$, let

$$X_j(t) = X_j(0) - \int_0^t \mathbb{1}_{(X_j(s-)>0)} D_j(ds) + \int_{[0,t]\times\mathbb{R}_+} \mathbb{1}_{(0\leq y\leq C_j(s-))} A_j(dsdy), \tag{3.2}$$

$$C_j(t) = d\zeta \sum_{k\in\mathcal{K}} \frac{p_{k,m} w_k}{\delta_k} \sum_{(M_2,\ldots,M_d)\in\mathcal{M}^{d-1}} h_t(j, M_2, \ldots, M_d), \tag{3.3}$$

where $\mathbf{M}(j) = m$ and

$$h_t(j, M_2, \ldots, M_d) = \prod_{h=2}^{d} \frac{v_{M_h} p_{k,M_h}}{\delta_k} \int_{\mathbb{N}^{d-1}} b(X_j(t), x_{j_2}, \ldots, x_{j_d}) \mu_t^{M_2}(dx_{j_2}) \cdots \mu_t^{M_d}(dx_{j_d}),$$

$$b(\mathbf{x}) = b(x_1, \ldots, x_d) := \sum_{r=1}^{d} \frac{1}{r} \mathbb{1}_{(x_1 = \min_{j\in[d]\mathbf{x}},\, |\arg\min \mathbf{x}| = r)}, \quad \mathbf{x} = (x_1, \ldots, x_d) \in \mathbb{N}_0^d,$$

$$\mu_t^m = \mathcal{L}(X_i(t)), \quad \forall\, i \in V_m, m \in \mathcal{M}, t \geq 0. \tag{3.4}$$

Here, $\{D_j : j \in V_m\}$ are i.i.d. Poisson processes with rate u_m for each $m \in \mathcal{M}$, $\{A_j : j \in \mathbb{N}\}$ are i.i.d. Poisson random measures on $[0,\infty) \times \mathbb{R}_+$ with intensity $\lambda dsdy$, and all D_j's and A_j's are independent. Loosely speaking, A_j corresponds to the arrival processes, and D_j corresponds to the departure processes at servers. $h_t(j, \cdots)$ is the probability that at time t, the server j will receive the new task given the event that the server j is among the d selected servers and the new task is of cluster k. Neglecting d and ζ, $C_j(t)$ can be understood as the probability that the server j will receive the new task. We note that the existence and uniqueness of solutions to (3.2) and (3.3) can be proved by standard arguments (see, e.g., Sznitman 1991, Méléard 1996) using the boundedness and Lipschitz property of the functions b and $x \longmapsto \mathbb{1}_{(x>0)}$ on $\mathbb{N}_0$.

Theorem 1 (Convergence to the McKean–Vlasov Process and Propagation of Chaos). *Consider any fixed* $\mathbf{q}^\infty = (q_{m,l}^\infty, m \in \mathcal{M}, l \in \mathbb{N}_0) \in \mathcal{S}$. *Assume that all* $X_j^N(0)$*'s are independent, and for each* $m \in \mathcal{M}$, $\{X_j^N(0) : j \in V_m^N\}$ *is i.i.d. with* $\mathbb{P}(X_j^N(0) \geq l) = q_{m,l}^\infty$, $l \in \mathbb{N}_0$. *On any finite time interval* $[0,T]$, $T > 0$, *for any* $m \in \mathcal{M}$ *and* $j \in V_m$, *the queue-length process* $X_j^N(\cdot)$ *at server* j *weakly converges to the process* $X_j(\cdot)$ *in* (3.2). *In fact, one can suitably couple* X_j^N *with* X_j *such that*

$$\max_{j\in V^N} \mathbb{E}\|X_j^N - X_j\|_{*,T}^2 \xrightarrow{N\to\infty} 0, \tag{3.5}$$

and hence, the propagation of chaos property holds; that is, for any $n \in \mathbb{N}$ *and distinct* $j_h \in V_{M_h}$, $h = 1, \ldots, n$,

$$\mathcal{L}(X_{j_1}^N, \ldots, X_{j_n}^N) \xrightarrow{N\to\infty} \mathcal{L}(X_{j_1}, \ldots, X_{j_n}) = \mu^{M_1} \otimes \cdots \otimes \mu^{M_n}. \tag{3.6}$$

Theorem 1 gives us the limit law of all individual queues. Next, in Theorem 2, we will show how such a server-level convergence can be used to obtain a convergence result for the global occupancy process $\mathbf{q}^N(\cdot)$ to a deterministic dynamical system, which was our primary goal. The proofs of Theorem 1 and Theorem 2 are provided in Section 4.

Theorem 2 (Process-Level Convergence for i.i.d. Starting State). *Assume that all* $X_j^N(0)$*'s are independent, and for each* $m \in \mathcal{M}$, $\{X_j^N(0) : j \in V_m^N\}$ *is i.i.d. with* $\mathbb{P}(X_j^N(0) \geq l) = q_{m,l}^\infty$, $l \in \mathbb{N}_0$ *for some* $\mathbf{q}^\infty = (q_{m,l}^\infty, m \in \mathcal{M}, l \in \mathbb{N}_0) \in \mathcal{S}$. *Then, on any finite time interval, the occupancy process* $\mathbf{q}^N(\cdot)$ *converges weakly with respect to Skorokhod* J_1 *topology to the deterministic*

limit $\mathbf{q}(\cdot) := (q_{m,l}(\cdot), m \in \mathcal{M}, l \in \mathbb{N}_0)$ *given by the unique solution to the following system of ODEs. For all* $m \in \mathcal{M}$, $q_{m,0}(t) = 1$, $q_{m,l}(0) = q^{\infty}_{m,l}$, *and*

$$\frac{dq_{m,l}(t)}{dt} = -u_m(q_{m,l}(t) - q_{m,l+1}(t)) + \lambda\zeta(q_{m,l-1}(t) - q_{m,l}(t))\sum_{k\in\mathcal{K}} \frac{p_{k,m}w_k}{\delta_k} \frac{(\tilde{q}_{k,l-1}(t))^d - (\tilde{q}_{k,l}(t))^d}{\tilde{q}_{k,l-1}(t) - \tilde{q}_{k,l}(t)}, \quad \forall l \in \mathbb{N}. \tag{3.7}$$

Here, $\tilde{q}_{k,l}(t) = \sum_{m\in\mathcal{M}} \frac{v_m p_{k,m}}{\delta_k} q_{m,l}(t)$ *for all* $k \in \mathcal{K}$.

Remark 1. Using the propagation of chaos property (3.6) and the fact that $\{X_j(t) : j \in \mathbb{N}\}$ is independent and $\{X_j(t) : j \in V_m\}$ is i.i.d. for each $m \in \mathcal{M}$, it follows that the limit of the global occupancy process at any time instant t, in fact, corresponds to the laws of $X_j(t)$ for each type of servers j in (3.2): that is,

$$\mu^m_t[l, \infty) = \mathbb{P}(X_j(t) \geq l) = q_{m,l}(t), \quad j \in V_m, m \in \mathcal{M}, l \in \mathbb{N}_0, t \geq 0.$$

3.3. Process-Level Limit: General Case

Theorem 2 requires the strong assumption that for each $m \in \mathcal{M}$, $X^N_j(0)$, $j \in V^N_m$, are i.i.d. In order to argue the interchange of limits, we need to relax this assumption on initial states. This is because the arguments for the interchange of limits involve initiating the prelimit system at the steady state and then showing that as $N \to \infty$, the system must converge to the unique fixed point of the limiting ODE. This requires us to characterize the (process-level) limiting trajectory of the system starting from the arbitrary occupancy state. We achieve this in this section.

Intuitively, the assumption of i.i.d. in Theorems 1 and 2 ensures that the local occupancy observed by any dispatcher $i \in W^N_k$, $k \in \mathcal{K}$ is "close," in suitable sense, to the average occupancy at the entire system. This phenomenon can be ensured asymptotically, even without the i.i.d. assumption, if the graph sequence satisfies a property we call the *clustered proportional sparsity*. This notion was first introduced for the homogeneous systems in Rutten and Mukherjee (2022). The definition is a modified notion that is suitable for the current heterogeneous setting.

Definition 2 (Clustered Proportional Sparsity). Recall $\mathcal{N}^N_w(i) = \{j \in V^N : (i,j) \in E^N\}$. The sequence $\{G^N\}_N$ is called clustered proportionally sparse if for any $\varepsilon > 0$,

$$\sup_{k\in\mathcal{K}} \sup_{U\subseteq V^N} \left| \left\{ i \in W^N_k : \left| \frac{|\mathcal{N}^N_w(i) \cap U|}{|\mathcal{N}^N_w(i)|} - \frac{|E^N_k(U)|}{|E^N_k(V^N)|} \right| \geq \varepsilon \right\} \right| / |W^N_k| \xrightarrow{N\to\infty} 0, \tag{3.8}$$

where $E^N_k(U) := \{(i,j) \in W^N_k \times U : (i,j) \in E^N\}$.

Remark 2. We can view the subset U in the definition as a test set, say $U = \mathcal{Q}^N_{m,l}(t)$, where $\mathcal{Q}^N_{m,l}(t)$ is the set of type $m \in \mathcal{M}$ servers with queue length at least $l \in \mathbb{N}_0$ at time t. Hence, Definition 2 ensures that for all but $o(N)$ dispatchers, the observed empirical queue-length distribution within its neighborhood is close to the global weighted empirical queue-length distribution (Definition 4) of its corresponding type. Then, the global occupancy process evolves similarly to (and converges to the same limit as) the case when the initial states are i.i.d.

Theorem 3 (Process-Level Convergence). *Let* $\{G^N\}_N$ *be a clustered proportionally sparse sequence of graphs. Assume that* $\mathbf{q}^N(0)$ *weakly converges to* $\mathbf{q}^{\infty} \in \mathcal{S}$. *Then, on any finite time interval, the occupancy process* $\mathbf{q}^N(\cdot)$ *converges weakly with respect to the Skorokhod* J_1 *topology to the deterministic limit* $\mathbf{q}(\cdot) := (q_{m,l}(\cdot), m \in \mathcal{M}, l \in \mathbb{N}_0)$ *given by the unique solution to the system of ODEs defined by* (3.7) *with initial state* $\mathbf{q}(0) = (q^{\infty}_{m,l}, m \in \mathcal{M}, l \in \mathbb{N}_0)$.

The proof of Theorem 3 is given in Section 4.4.

3.4. Convergence of Steady States

In the last section, we showed the process-level convergence of global occupancy process $\mathbf{q}^N(\cdot)$ to a mean-field limit $\mathbf{q}(\cdot)$. In this section, we will establish the convergence of the sequence of stationary distributions to the unique fixed point of the mean-field limit by establishing the interchange of large-N and large-t limits: $\lim_{t\to\infty}\lim_{N\to\infty}\mathbf{q}^N(t) = \lim_{N\to\infty}\lim_{t\to\infty}\mathbf{q}^N(t)$. Throughout this section, we will assume that the sequence of systems is in the subcritical

regime (recall Definition 1). The first result states that the limiting system of ODEs has a unique fixed point $\mathbf{q}^*$ and that it satisfies the global stability property (i.e., for any initial point $\mathbf{q}(0) \in \mathcal{S}$, $\lim_{t\to\infty} \mathbf{q}(t) = \mathbf{q}^*$).

Theorem 4 (Global Stability). *Let $\overline{\mathbf{q}}(t, \mathbf{q}_0)$ be the solution to the system of ODEs in (3.7) with the initial point $\mathbf{q}(0) = \mathbf{q}_0 \in \mathcal{S}$. Then, there exists a unique fixed point $\mathbf{q}^* = (q^*_{m,l}, m \in \mathcal{M}, l \in \mathbb{N}_0) \in \mathcal{S}$ such that $\lim_{t\to\infty} \overline{\mathbf{q}}(t, \mathbf{q}_0) = \mathbf{q}^*$.*

The proof of Theorem 4 is given in Section 5. It relies on a monotonicity property of the system, which ensures that for two processes $\mathbf{q}^1(\cdot)$ and $\mathbf{q}^2(\cdot)$, if $\mathbf{q}^1(0) \le \mathbf{q}^2(0)$, then $\mathbf{q}^1(t) \le \mathbf{q}^2(t)$ for all $t \ge 0$ (see Martin and Suhov 1999, Stolyar 2015).

The last ingredient that we need in order to prove the interchange of limits is to establish tightness of the sequence of random variables $\{\mathbf{q}^N(\infty)\}_{N\ge1}$ under a suitable metric, where $\mathbf{q}^N(\infty) := \lim_{t\to\infty} \mathbf{q}^N(t)$. Here, as before, we should note that the process $(\mathbf{q}^N(t))_{t\ge0}$ is not Markovian. That is why the random variable $\mathbf{q}^N(\infty)$ should be interpreted as the functional applied to the steady-state system. The tightness result is stated in the next theorem.

Theorem 5 (Tightness). *For any $\varepsilon > 0$, there exists a compact subset $\overline{K}(\varepsilon) \subseteq \mathcal{S}$, when $\mathcal{S}$ is equipped with the ℓ_1-topology, such that $\mathbb{P}(\mathbf{q}^N(\infty) \notin \overline{K}(\varepsilon)) < \varepsilon, \quad \forall N \ge 1$.*

Theorem 5 is proved in Section 5. The key idea is to use the Lyapunov function approach to bound the expected sum of tails $q^N_{m,l}(\infty)$. Combining Theorems 3, 4, and 5, we can prove the following interchange of limits result.

Theorem 6 (Convergence of Steady States). *Let $\{G^N\}_{N\ge1}$ be a clustered proportionally sparse sequence of graphs satisfying Condition 1. Then, the sequence of random variables $\{\mathbf{q}^N(\infty)\}_{N\ge1}$ converges weakly to $\mathbf{q}^*$, the unique fixed point of the system of ODEs in (3.7).*

One major discovery about the JSQ(d) policy for the classical, homogeneous, fully flexible system is that the limit of the stationary distribution (which in our case, is given by $\mathbf{q}^*$) has a double-exponential decay of tail (Mitzenmacher 1996b, Vvedenskaya et al. 1996) for any $d \ge 2$. This is in sharp contrast with the (single) exponential decay of the corresponding tail for random routing or $d = 1$. In fact, in this case, for any $d \ge 2$, $\mathbf{q}^*$ can be characterized explicitly as $q^*_l = \lambda^{\frac{d^l-1}{d-1}}$, where q^*_l is the (limiting) steady-state fraction of servers with queue length at least $l = 1, 2, \ldots$. In the current case of heterogeneous systems, it is intractable to characterize the fixed point $\mathbf{q}^*$ explicitly. However, as stated in the next theorem, we can still prove that the doubly exponential decay of the tails $q^*_{m,l}$ for each $m \in \mathcal{M}$ holds.

Theorem 7 (Double-Exponential Tail Decay). *Let $\mathbf{q}^* = (q^*_{m,l}, m \in \mathcal{M}, l \in \mathbb{N}_0)$ be the unique fixed point of the system of ODEs in (3.7). Then, for all $m \in \mathcal{M}$, the sequence $\{q^*_{m,l}, l \in \mathbb{N}_0\}$ decreases doubly exponentially; that is, there exist positive constant $l_m \in \mathbb{N}_0$, $a_m \in (0,1)$, and $b_m > 0$ such that for all $l \ge l_m$, $q^*_{m,l} \le b_m a_m^{d^l}$.*

3.5. Simple Data Locality Design Using Randomization

Sections 3.1–3.4 characterize the performance of the occupancy process for arbitrary deterministic sequence of systems where the underlying graph sequence satisfies certain properties. In particular, Condition 1 and Definition 2 provide sufficient criteria under which both the process-level convergence (Theorem 3) and the interchange of limits (Theorem 6) hold. In this section, we show that graphs satisfying the required criteria can be obtained easily if the compatibility graph is designed suitably randomly. Given the asymptotic edge-density parameters in Condition 1, we define a certain sequence of *inhomogeneous random graphs* or IRG as follows.

Definition 3 (IRG($\mathbf{p}$)). Given $\mathbf{p} := (p_{k,m}, k \in \mathcal{K}, m \in \mathcal{M})$, the Nth system of IRG($\mathbf{p}$) is constructed as follows. For any $k \in \mathcal{K}$ and $m \in \mathcal{M}$, dispatcher i and server j share an edge with probability $p_{k,m}$ for all $i \in W^N_k$ and $j \in V^N_m$, independently of each other.

For any $\mathbf{p}$ for which the asymptotic stability criterion holds, we have the following result for the sequence of IRG($\mathbf{p}$).

Theorem 8. *Let $\mathbf{p} = (p_{k,m}, k \in \mathcal{K}, m \in \mathcal{M})$ be such that the stability criterion in (3.1) holds and $\{G_N\}_{N\ge1}$ be a sequence of IRG($\boldsymbol{p}$) with increasing N. Then, the conclusions of Theorems 3 and 6 hold for $\{G_N\}_{N\ge1}$.*

The proof of Theorem 8 is provided in Appendix I. It relies on verifying that the sequence of IRG($\mathbf{p}$) graphs satisfies Condition 1 and the property of clustered proportional sparsity almost surely. The verification involves using the concentration of measure arguments to establish structural properties of the compatibility graphs.

4. Proof of Transient Limit Results

In this section, we will prove the results of transient limit results (Theorems 1–3 in Sections 4.2, 4.3, and 4.4, respectively). We start by proving a few auxiliary results in Section 4.1.

4.1. Auxiliary Results

First, we will need a characterization of the evolution of the queue-length process at each server. To describe this evolution, let us introduce the following notations:

$$\mathtt{set}^N(j) := \{(j_2,\dots,j_d) \in [N]^{d-1} : (j, j_2,\dots,j_d) \text{ are distinct}\}, \tag{4.1}$$

$$\mathtt{sett}^N(j) := \{(j_2,\dots,j_d,j_2',\dots,j_d') \in [N]^{2d-2} : (j_2,\dots,j_d) \in \mathtt{set}^N(j), (j_2',\dots,j_d') \in \mathtt{set}^N(j), (j_2,\dots,j_d) \cap (j_2',\dots,j_d') \neq \emptyset\}. \tag{4.2}$$

To represent the graph, define the edge occupancy $\xi_{i,j}^N$ to be the binary variable:

$$\xi_{i,j}^N = \begin{cases} 1, & \text{if } (i,j) \in E^N, \\ 0, & \text{otherwise,} \end{cases} \quad \text{for all } i \in W^N, j \in V^N.$$

Recall the function b, Poisson processes $\{D_j\}$, and Poisson random measures $\{A_j\}$ in and after (3.4). By Condition 1, for all large-enough N, all dispatchers in the Nth system have at least d neighbors. Hence, without loss of generality, in the rest of this section, we will only consider the case $\delta_i^N \geq d, \ \forall i \in W^N$. In that case, because of the Poisson thinning property, note that we can write $X_j^N(t)$ as follows:

$$X_j^N(t) = X_j^N(0) - \int_0^t \mathbb{1}_{(X_j^N(s-)>0)} D_j(ds) + \int_{[0,\infty)\times\mathbb{R}_+} \mathbb{1}_{(0\leq y\leq C_j^N(s-))} A_j(dsdy), \tag{4.3}$$

where

$$\begin{aligned} C_j^N(s) &= \sum_{i\in W^N} \xi_{i,j}^N \sum_{(j_2,\dots,j_d)\in \mathtt{set}^N(j)} \frac{\xi_{i,j_2}^N \times\cdots\times \xi_{i,j_d}^N}{\binom{\delta_i^N}{d}(d-1)!} b(X_j^N(s), X_{j_2}^N(s),\dots,X_{j_d}^N(s)) \\ &= \sum_{k\in\mathcal{K}} \sum_{i\in W_k^N} \xi_{i,j}^N \sum_{(j_2,\dots,j_d)\in \mathtt{set}^N(j)} \frac{\xi_{i,j_2}^N \times\cdots\times \xi_{i,j_d}^N}{\binom{\delta_i^N}{d}(d-1)!} b(X_j^N(s), X_{j_2}^N(s),\dots,X_{j_d}^N(s)). \end{aligned} \tag{4.4}$$

The RHS of the first summation in (4.4) represents the probability that a job arriving at the dispatcher $i \in W^N$ will be assigned to the server $j \in V^N$ given the state $(X_j^N, j \in V^N)$. Moreover, by Condition 1, the term C_j^N for all $j \in V^N$ can be upper bounded, uniformly for all t, by a constant for all large-enough N, which is stated in Lemma 4.

When we do some estimation, like bounding the term C_j^N, we need to uniformly bound the number of the neighbors of servers or dispatchers. Such uniformity is stated in Lemma 3 and is a direct result of Condition 1. Recall $\delta_i^N = |\mathcal{N}_w^N(i)|$ and $\delta_k = \sum_{m\in\mathcal{M}} p_{k,m} v_m$.

Lemma 3. *For each $k \in \mathcal{K}$,*

$$\lim_{N\to\infty} \max_{i\in W_k^N} \frac{\deg_w^N(i,m)}{|V_m^N|} = \lim_{N\to\infty} \min_{i\in W_k^N} \frac{\deg_w^N(i,m)}{|V_m^N|} = p_{k,m}, \quad m \in \mathcal{M}, \tag{4.5}$$

and

$$\lim_{N\to\infty} \max_{i\in W_k^N} \frac{\delta_i^N}{N} = \lim_{N\to\infty} \min_{i\in W_k^N} \frac{\delta_i^N}{N} = \delta_k. \tag{4.6}$$

Also, for each $m \in \mathcal{M}$,

$$\lim_{N\to\infty} \max_{j\in V_m^N} \frac{\deg_v^N(k,j)}{|W_k^N|} = \lim_{N\to\infty} \min_{j\in V_m^N} \frac{\deg_v^N(k,j)}{|W_k^N|} = p_{k,m}, \quad k \in \mathcal{K}. \tag{4.7}$$

Lemma 4. *For all large-enough* N, *we have that for any* $m \in \mathcal{M}$, $j \in V_m^N$, *and* $t \geq 0$,

$$C_j^N(t) \leq 2\zeta d \sum_{k\in\mathcal{K}} \frac{p_{k,m} w_k}{\delta_k}. \tag{4.8}$$

Proof. By the definition of $C_j^N(t)$, for any $t \geq 0$ and large-enough N,

$$C_j^N(t) \leq \sum_{k\in\mathcal{K}} \sum_{i\in W_k^N} \xi_{i,j}^N \sum_{(j_2,\ldots,j_d)\in \mathtt{set}^N(j)} \frac{\xi_{i,j_2}^N \times \cdots \times \xi_{i,j_d}^N}{\binom{\delta_i^N}{d}(d-1)!} = \sum_{k\in\mathcal{K}} \sum_{i\in W_k^N} \xi_{i,j}^N \frac{\binom{\delta_i^N - 1}{d-1}}{\binom{\delta_i^N}{d}} \leq 2\zeta d \sum_{k\in\mathcal{K}} \frac{p_{k,m} w_k}{\delta_k},$$

where the first inequality is because of $b(\cdot) \leq 1$ and the last inequality comes from Lemma 3. □

By Lemma 3, we know that the neighborhoods of dispatchers of the same type are almost the same. With the scale of the system size, the local graph structure for each dispatcher of the same type will converge to the average one. The following two lemmas give necessary approximation of the graph structures for large-N systems. Their proofs are combinatorial and are based on Condition 1 and Lemma 3. They are provided in Appendix B.

Lemma 5. *Consider a sequence* $\{G^N\}_N$ *satisfying Condition 1. For each* $m \in \mathcal{M}$,

$$\max_{j\in V_m^N} \max_{k\in\mathcal{K}} \max_{(M_2,\ldots,M_d)\in\mathcal{M}^{d-1}} \left| \sum_{i\in W_k^N} \xi_{i,j}^N \sum_{\substack{(j_2,\ldots,j_d)\in \mathtt{set}^N(j) \\ \text{s.t. } j_2\in V_{M_2}^N,\ldots,j_d\in V_{M_d}^N}} \frac{\xi_{i,j_2}^N \times \cdots \times \xi_{i,j_d}^N}{\binom{\delta_i^N}{d}(d-1)!} - \zeta d \frac{p_{k,m} w_k}{\delta_k} \prod_{h=2}^{d} \frac{v_{M_h} p_{k,M_h}}{\delta_k} \right| \xrightarrow{N\to\infty} 0. \tag{4.9}$$

Lemma 5 states that the probability that the server j in the Nth system will be among the d selected servers when a new task arrives converges to the corresponding probability in the limit system. The argument is mainly based on the law of large numbers (L.L.N.).

Lemma 6. *Consider any* $m \in \mathcal{M}$ *and* $j \in V_m$. *For large-enough* N,

$$\sum_{i\in W^N} \sum_{\mathtt{sett}^N(j)} \frac{\xi_{i,j}^N \times \xi_{i,j_2}^N \times \cdots \times \xi_{i,j_d}^N}{\binom{\delta_i^N}{d}(d-1)!} \frac{\xi_{i,j}^N \times \xi_{i,j_2'}^N \times \cdots \times \xi_{i,j_d'}^N}{\binom{\delta_i^N}{d}(d-1)!} \leq \frac{C_1}{N^2}, \tag{4.10}$$

where C_1 *is a positive constant. Similarly,*

$$\sum_{\substack{i_1,i_2\in W^N, \\ i_1\neq i_2}} \sum_{\mathtt{sett}^N(j)} \frac{\xi_{i_1,j}^N \times \xi_{i_1,j_2}^N \times \cdots \times \xi_{i_1,j_d}^N}{\binom{\delta_{i_1}^N}{d}(d-1)!} \frac{\xi_{i_2,j}^N \times \xi_{i_2,j_2'}^N \times \cdots \times \xi_{i_2,j_d'}^N}{\binom{\delta_{i_2}^N}{d}(d-1)!} \leq \frac{C_2}{N}, \tag{4.11}$$

where C_2 *is a positive constant.*

Lemma 6 implies that if we select two elements, say $(j_2,\ldots,j_d),(j_2',\ldots,j_d')$ independently from $\mathtt{set}^N(j)$, then the probability of $(j_2,\ldots,j_d,j_2',\ldots,j_d') \in \mathtt{sett}^N(j)$ is small. (4.10) and (4.11) are used in (4.16).

4.2. Convergence to the McKean–Vlasov Process: i.i.d. Case

Proof of Theorem 1. It suffices to prove (3.5). Fix any $m \in \mathcal{M}$, $j \in V_m$, and $T > 0$. We have that for any fixed $t \in [0,T]$ and any N such that $j \in V^N$,

$$\begin{aligned}
\mathbb{E}\|X_j^N - X_j\|_{*,t}^2 &\le c_0 \mathbb{E}\|X_j^N(t) - X_j(t)\|^2 \\
&\le c_1 \mathbb{E}\left(\int_0^t |\mathbb{1}_{(X_j^N(s)>0)} - \mathbb{1}_{(X_j(s)>0)}|^2 ds\right) + c_1 \mathbb{E}\left(\int_0^t |\mathbb{1}_{(X_j^N(s)>0)} - \mathbb{1}_{(X_j(s)>0)}| ds\right)^2 \\
&\quad + c_1 \mathbb{E}\left(\int_{[0,t]\times\mathbb{R}_+} |\mathbb{1}_{(0\le y\le C_j^N(s))} - \mathbb{1}_{(0\le y\le C_j(s))}|^2 ds dy\right) \\
&\quad + c_1 \mathbb{E}\left(\int_{[0,t]\times\mathbb{R}_+} |\mathbb{1}_{(0\le y\le C_j^N(s))} - \mathbb{1}_{(0\le y\le C_j(s))}| ds dy\right)^2 \\
&\le c_1 \mathbb{E}\left(\int_0^t |X_j^N(s) - X_j(s)|^2 ds\right) + c_1 \mathbb{E}\left(\int_0^t |X_j^N(s) - X_j(s)| ds\right)^2 \\
&\quad + c_1 \mathbb{E}\left(\int_0^t |C_j^N(s) - C_j(s)|^2 ds\right) + c_1 \mathbb{E}\left(\int_0^t |C_j^N(s) - C_j(s)| ds\right)^2 \\
&\le c_2 \int_0^t \mathbb{E}|X_j^N(s) - X_j(s)|^2 ds + c_2 \int_0^t \mathbb{E}|C_j^N(s) - C_j(s)| ds, \qquad (4.12)
\end{aligned}$$

where c_0, c_1, and c_2 are positive constants. The first two inequalities are by Doob's inequalities and Cauchy–Schwarz, respectively. The last inequality comes from the uniform boundedness of $C_j^N(t)$ proved in Lemma 4 and $C_j(t) \le d\zeta$ by the definition. By adding and subtracting terms, we have

$$|C_j^N(s) - C_j(s)| \le |C_j^N(s) - C_j^{N,1}(s)| + |C_j^{N,1}(s) - C_j^{N,2}(s)| + |C_j^{N,2}(s) - C_j(s)|, \qquad (4.13)$$

where

$$C_j^{N,1} = \sum_{k\in\mathcal{K}} \sum_{i\in W_k^N} \left[\xi_{i,j}^N \sum_{(j_2,\dots,j_d)\in \mathtt{set}^N(j)} \frac{\xi_{i,j_2}^N \times \cdots \times \xi_{i,j_d}^N}{\binom{\delta_i^N}{d}(d-1)!} b(X_j(s), X_{j_2}(s), \dots, X_{j_d}(s)) \right],$$

$$C_j^{N,2} = \sum_{k\in\mathcal{K}} \sum_{i\in W_k^N} \left[\xi_{i,j}^N \sum_{(j_2,\dots,j_d)\in \mathtt{set}^N(j)} \frac{\xi_{i,j_2}^N \times \cdots \times \xi_{i,j_d}^N}{\binom{\delta_i^N}{d}(d-1)!} \int_{\mathbb{N}^{d-1}} b(X_j(t), x_{j_2}, \dots, x_{j_d}) \mu_t^{\mathbf{M}(j_2)}(dx_{j_2}) \cdots \mu_t^{\mathbf{M}(j_d)}(dx_{j_d}) \right].$$

First, consider $|C_j^N(s) - C_j^{N,1}(s)|$. For large-enough N,

$$\begin{aligned}
&\mathbb{E}|C_j^N(s) - C_j^{N,1}(s)| \\
&= \mathbb{E}\left|\sum_{k\in\mathcal{K}}\sum_{i\in W_k^N}\left[\xi_{i,j}^N \sum_{(j_2,\ldots,j_d)\in \mathtt{set}^N(j)} \frac{\xi_{i,j_2}^N \times\cdots\times \xi_{i,j_d}^N}{\binom{\delta_i^N}{d}(d-1)!}(b(X_j^N(s), X_{j_2}^N(s),\ldots,X_{j_d}^N(s))\right.\right. \\
&\qquad \left.\left. - b(X_j(s), X_{j_2}(s),\ldots,X_{j_d}(s))\right]\right| \\
&\le \mathbb{E}\sum_{k\in\mathcal{K}}\sum_{i\in W_k^N}\left[\xi_{i,j}^N \sum_{(j_2,\ldots,j_d)\in \mathtt{set}^N(j)} \frac{\xi_{i,j_2}^N \times\cdots\times \xi_{i,j_d}^N}{\binom{\delta_i^N}{d}(d-1)!}(|X_j^N(s) - X_j(s)| + \cdots + |X_{j_d}^N(s) - X_{j_d}(s)|)\right] \\
&\le d \times \max_{j\in V^N} \mathbb{E}|X_j^N(s) - X_j(s)| \times \sum_{k\in\mathcal{K}}\sum_{i\in W_k^N} \xi_{i,j}^N \sum_{(j_2,\ldots,j_d)\in \mathtt{set}^N(j)} \frac{\xi_{i,j_2}^N \times\cdots\times \xi_{i,j_d}^N}{\binom{\delta_i^N}{d}(d-1)!} \\
&\le c_3 \max_{j\in V^N} \mathbb{E}|X_j^N(s) - X_j(s)|, \qquad (4.14)
\end{aligned}$$

where c_3 is constant. The first inequality is from that $b(\cdot)$ is Lipschitz continuous with Lipschitz constant 1 and that the last inequality is from (4.9).

Second, consider $|C_j^{N,1}(s) - C_j^{N,2}(s)|$. By Jensen's inequality, we have $[\mathbb{E}|C_j^{N,1}(s) - C_j^{N,2}(s)|]^2 \le \mathbb{E}|C_j^{N,1}(s) - C_j^{N,2}(s)|^2$. Hence, it is sufficient to bound $\mathbb{E}|C_j^{N,1}(s) - C_j^{N,2}(s)|^2$:

$$\begin{aligned}
&\mathbb{E}|C_j^{N,1}(s) - C_j^{N,2}(s)|^2 \\
&= \mathbb{E}\left|\sum_{k\in\mathcal{K}}\sum_{i\in W_k^N}\left[\xi_{i,j}^N \sum_{(j_2,\ldots,j_d)\in \mathtt{set}^N(j)} \frac{\xi_{i,j_2}^N \times\cdots\times \xi_{i,j_d}^N}{\binom{\delta_i^N}{d}(d-1)!} b(X_j(s), X_{j_2}(s),\ldots,X_{j_d}(s))\right]\right. \\
&\qquad \left. - \sum_{k\in\mathcal{K}}\sum_{i\in W_k^N}\left[\xi_{i,j}^N \sum_{(j_2,\ldots,j_d)\in \mathtt{set}^N(j)} \frac{\xi_{i,j_2}^N \times\cdots\times \xi_{i,j_d}^N}{\binom{\delta_i^N}{d}(d-1)!} \int_{\mathbb{N}_0^{d-1}} b(X_j(s), x_{j_2},\ldots,x_{j_d})\mu_s^{\mathbf{M}(j_2)}(dx_{j_2})\cdots\mu_s^{\mathbf{M}(j_d)}(dx_{j_d})\right]\right|^2 \qquad (4.15) \\
&\le \mathbb{E}\left[\sum_{i_1,i_2\in W^N}\sum_{\mathtt{sett}^N(j)} \frac{\xi_{i_1,j}^N \times \xi_{i_1,j_2}^N \times\cdots\times \xi_{i_1,j_d}^N}{\binom{\delta_{i_1}^N}{d}(d-1)!} \frac{\xi_{i_2,j}^N \times \xi_{i_2,j_2'}^N \times\cdots\times \xi_{i_2,j_d'}^N}{\binom{\delta_{i_2}^N}{d}(d-1)!}\right] \\
&\overset{(a)}{\le} \mathbb{E}\left[\sum_{i\in W^N}\sum_{\mathtt{sett}^N(j)} \frac{\xi_{i,j}^N \times \xi_{i,j_2}^N \times\cdots\times \xi_{i,j_d}^N}{\binom{\delta_i^N}{d}(d-1)!} \frac{\xi_{i,j}^N \times \xi_{i,j_2'}^N \times\cdots\times \xi_{i,j_d'}^N}{\binom{\delta_i^N}{d}(d-1)!}\right.
\end{aligned}$$

$$+\sum_{i_1,i_2\in W^N,i_1\neq i_2}\sum_{\mathtt{sett}^N(j)}\left.\frac{\xi^N_{i_1,j}\times\xi^N_{i_1,j_2}\times\cdots\times\xi^N_{i_1,j_d}}{\binom{\delta^N_{i_1}}{d}(d-1)!}\frac{\xi^N_{i_2,j}\times\xi^N_{i_2,j'_2}\times\cdots\times\xi^N_{i_2,j'_d}}{\binom{\delta^N_{i_2}}{d}(d-1)!}\right]$$

$$\le c_4N^{-2}+c_5N^{-1}, \tag{4.16}$$

where the first inequality is because of the fact that $X_j(0)$ is i.i.d. for $j\in V_m$ and independent for different m; so, for each $m\in\mathcal{M}$, $\{X_j(s), j\in V_m\}$ are also i.i.d., and the independence across the server pools holds for any fixed $s>0$. Hence, if $(j,j_2,\ldots,j_d,j'_2,\ldots,j'_d)$ are distinct, then

$$\mathbb{E}\left[\left(b(X_j(t),X_{j_2}(t),\ldots,X_{j_d}(t))-\int_{\mathbb{N}^{d-1}}b(X_j(t),x_{j_2},\ldots,x_{j_d})\mu_t^{\mathbf{M}(j_2)}(dx_{j_2})\cdots\mu_t^{\mathbf{M}(j_d)}(dx_{j_d})\right)\times\right.$$

$$\left.\left(b(X_j(t),X_{j'_2}(t),\ldots,X_{j'_d}(t))-\int_{\mathbb{N}^{d-1}}b(X_j(t),x_{j'_2},\ldots,x_{j'_d})\mu_t^{\mathbf{M}(j'_2)}(dx_{j'_2})\cdots\mu_t^{\mathbf{M}(j'_d)}(dx_{j'_d})\right)\right]=0,$$

and $b(\cdot)$ and $\int b(\cdot)\mu(d\cdot)$ are both in $[0,1]$. The last inequality of (4.16) is by (4.10) and (4.11).

Third, consider $|C_j^{N,2}(s)-C_j(s)|$:

$$\mathbb{E}|C_j^{N,2}(s)-C_j(s)|$$

$$=\mathbb{E}\left|\sum_{k\in\mathcal{K}}\sum_{i\in W_k^N}\left[\xi^N_{i,j}\sum_{(j_2,\ldots,j_d)\in\mathtt{set}^N(j)}\frac{\xi^N_{i,j_2}\times\cdots\times\xi^N_{i,j_d}}{\binom{\delta_i^N}{d}(d-1)!}\int_{\mathbb{N}^{d-1}}b(X_j(t),x_{j_2},\ldots,x_{j_d})\mu_t^{\mathbf{M}(j_2)}(dx_{j_2})\cdots\mu_t^{\mathbf{M}(j_d)}(dx_{j_d})\right]\right.$$

$$\left.-d\zeta\sum_{k\in\mathcal{K}}\frac{p_{k,m}w_k}{\delta_k}\sum_{(M_2,\ldots,M_d)\in\mathcal{M}^{d-1}}\prod_{h=2}^{d}\frac{v_{M_h}p_{k,M_h}}{\delta_k}\int_{\mathbb{N}^{d-1}}b(X_j(t),x_{j_2},\ldots,x_{j_d})\mu_t^{\mathbf{M}(j_2)}(dx_{j_2})\cdots\mu_t^{\mathbf{M}(j_d)}(dx_{j_d})\right|$$

$$\le c_6(N), \tag{4.17}$$

where $c_6(N)$ only depends on N and goes to zero as $N\to\infty$ and where the inequality comes from (4.9) and the fact that $\int b(\cdot)\mu(d\cdot)\in[0,1]$. Now, by (4.12), (4.13), (4.14), (4.16), and (4.17), we have that for large-enough N,

$$\max_{j\in V^N}\mathbb{E}\|X_j^N-X_j\|^2_{*,t}\le c_{10}\int_0^t\max_{j\in V^N}\mathbb{E}\|X_j^N-X_j\|^2_{*,t}ds+f(N),$$

where c_{10} is a constant and $f(N)$ is a function, which goes to zero as $N\to\infty$. Last, by Gronwall's inequality, we have (3.5), and this completes the proof. □

4.3. Convergence of the Occupancy Process: i.i.d. Case

In this section, we want to show the convergence of the occupancy process $\mathbf{q}^N(\cdot)$ to the limit process $\mathbf{q}$ represented by the ODE (3.7). The first step is to investigate the existence and uniqueness of the solution of the ODE (3.7). Define

$$\overline{\mathcal{S}}:=\{\mathbf{q}\in[0,1]^{M\times\mathbb{N}_0}: q_{m,0}=1, q_{m,l}\ge q_{m,l+1},\ \forall m\in\mathcal{M}, l\in\mathbb{N}_0\},$$

and clearly, $\mathcal{S}\subseteq\overline{\mathcal{S}}$.

Lemma 7. *If $\mathbf{q}(0)=\mathbf{q}_0\in\overline{\mathcal{S}}$, then the ODE system (3.7) has a unique solution denoted as $\overline{\mathbf{q}}(t,\mathbf{q}_0)$, $t\ge0$ in $\overline{\mathcal{S}}$.*

The proof of Lemma 7 is based on the Picard successive approximation method (Martin and Suhov 1999, theorem 1(i)) and is provided in Appendix C.

Proof of Theorem 2. Fix any $T\in(0,\infty)$. For each $m\in\mathcal{M}$, consider random measures $\mu_m^N=\frac{1}{|V_m^N|}\sum_{j\in V_m^N}\delta_{X_j^N(\cdot)}$ and $\overline{\mu}_m^N=\frac{1}{|V_m^N|}\sum_{j\in V_m^N}\delta_{X_j(\cdot)}$ on $\mathbb{S}:=\mathbb{D}([0,T],\mathbb{N}_0)$, where $X_j(\cdot)$ is defined in (3.2). Denote the joint measures $\mu^N=(\mu_1^N,\ldots,\mu_M^N)$

and $\overline{\mu}^N = (\overline{\mu}_1^N, \ldots, \overline{\mu}_M^N)$. Denote by $d_{BL}(\cdot,\cdot)$ the bounded Lipschitz metric for probability measures on $\mathbb{S}$:

$$d_{BL}(\mu_1,\mu_2) := \sup_{\|f\|_{BL}\le 1}\left|\int_{\mathbb{S}} f d\mu_1 - \int_{\mathbb{S}} f d\mu_2\right|, \quad \|f\|_{BL} := \max\left\{\|f\|_\infty, \sup_{x\neq y}\frac{f(x)-f(y)}{d(x,y)}\right\}.$$

From (3.5), we have

$$\mathbb{E}d_{BL}(\mu_m^N,\overline{\mu}_m^N) \le \mathbb{E}\sup_{\|f\|_{BL}\le 1}\frac{1}{|V_m^N|}\sum_{j\in V_m^N}|f(X_j^N)-f(X_j)| \le \frac{1}{|V_m^N|}\sum_{j\in V^N}\mathbb{E}\|X_j^N - X_j\|_{*,T} \xrightarrow{N\to\infty} 0,$$

which implies that $d_{BL}(\mu_m^N,\overline{\mu}_m^N)\xrightarrow{\mathbb{P}}0$ for each $m\in\mathcal{M}$. Because $\overline{\mu}_m^N\xrightarrow{\mathbb{P}}\mu_m$ by the L.L.N., we have $\mu^N = (\mu_1^N,\ldots,\mu_M^N)\xrightarrow{\mathbb{P}}(\mu_1,\ldots,\mu_M)$ by Slutsky's theorem. Also, it is easy to check that $\sup_N \mathbb{E}[\sup_{0\le t\le T}\|\mathbf{q}^N(t)\|_{\ell_1}^2] < \infty$. Thus, we have $\mathbf{q}^N\xrightarrow{\mathbb{P}}\mathbf{q}$. Next, we need to show that $\mathbf{q}$ satisfies (3.7). Define $f_l(x) = \mathbb{1}_{\{x\ge l\}}$, $l\in\mathbb{N}_0$. By (3.2), we have that for any $m\in\mathcal{M}$ and $j\in V_m$,

$$\begin{aligned}
\mathbb{E}f_l(X_j(t)) &= \mathbb{E}f_l(X_j(0)) + \int_0^t u_m\mathbb{E}\,\mathbb{1}_{\{X_j(s)>0\}}(f_l(X_j(s)-1)-f_l(X_j(s)))ds\\
&\quad + \int_0^t\int_{\mathbb{N}^{d-1}}\lambda\zeta_d\sum_{k\in\mathcal{K}}\frac{p_{k,m}w_k}{\delta_k}\sum_{(M_2,\ldots,M_d)\in\mathcal{M}^{d-1}}\prod_{h=2}^d\frac{v_{M_h}p_{k,M_h}}{\delta_k}\\
&\qquad\times\mathbb{E}[b(X_j(s),x_{j_2},\ldots,x_{j_d})(f_l(X_j(s)+1)-f_l(X_j(s)))]\mu_s^{M_2}(dx_{j_2})\cdots\mu_s^{M_d}(dx_{j_d})ds\\
&= \mathbb{E}f_l(X_j(0)) + \int_0^t u_m\mathbb{E}\,\mathbb{1}_{\{X_j(s)>0\}}(f_{l+1}(X_j(s))-f_l(X_j(s)))ds\\
&\quad + \int_0^t\int_{\mathbb{N}^{d-1}}\lambda\zeta_d\sum_{k\in\mathcal{K}}\frac{p_{k,m}w_k}{\delta_k}\sum_{(M_2,\ldots,M_d)\in\mathcal{M}^{d-1}}\prod_{h=2}^d\frac{v_{M_h}p_{k,M_h}}{\delta_k}\\
&\qquad\times\mathbb{E}[b(l-1,x_{j_2},\ldots,x_{j_d})(f_{l-1}(X_j(s))-f_l(X_j(s)))]\mu_s^{M_2}(dx_{j_2})\cdots\mu_s^{M_d}(dx_{j_d})ds.
\end{aligned}$$

For any $m\in\mathcal{M}$, if $j\in V_m$, then $\mathbb{E}f_l(X_j(t)) = q_{m,l}(t) = \mu_t^m[l,\infty)$ for $l = 1,2,\ldots$. Hence,

$$\begin{aligned}
q_{m,l}(t) &= q_{m,l}(0) - \int_0^t u_m(q_{m,l}(s)-q_{m,l+1}(s))ds + \int_0^t\lambda\zeta_d\sum_{k\in\mathcal{K}}\frac{p_{k,m}w_k}{\delta_k}(q_{m,l-1}(s)-q_{m,l}(s))\\
&\quad\times\sum_{(M_2,\ldots,M_d)\in\mathcal{M}^{d-1}}\prod_{h=2}^d\frac{v_{M_h}p_{k,M_h}}{\delta_k}\int_{\mathbb{N}^{d-1}}b(l-1,x_{j_2},\ldots,x_{j_d})\mu_s^{M_2}(dx_{j_2})\cdots\mu_s^{M_d}(dx_{j_d})ds.
\end{aligned}\tag{4.18}$$

Also,

$$\begin{aligned}
&\sum_{(M_2,\ldots,M_d)\in\mathcal{M}^{d-1}}\prod_{h=2}^d\frac{v_{M_h}p_{k,M_h}}{\delta_k}\int_{\mathbb{N}^{d-1}}b(l-1,x_{j_2},\ldots,x_{j_d})\mu_s^{M_2}(dx_{j_2})\cdots\mu_s^{M_d}(dx_{j_d})\\
&= \sum_{\overline{r}\in\overline{\mathcal{R}}}\sum_{\overline{r}'\in\overline{\mathcal{R}}'(\overline{r})}\frac{1}{1+|\overline{r}'|}\prod_{m\in\mathcal{M}}\binom{r_m}{r_m'}\left(\frac{v_m p_{k,m}}{\delta_k}\right)^{r_m}(q_{m,l-1}(s)-q_{m,l}(s))^{r_m'}(q_{m,l}(s))^{r_m-r_m'}\\
&= \sum_{r=0}^{d-1}\frac{1}{1+r}\binom{d-1}{r}\left(\sum_{m\in\mathcal{M}}\frac{v_m p_{k,m}}{\delta_k}q_{m,l-1}(s)-\sum_{m\in\mathcal{M}}\frac{v_m p_{k,m}}{\delta_k}q_{m,l}(s)\right)^r\left(\sum_{m\in\mathcal{M}}\frac{v_m p_{k,m}}{\delta_k}q_{m,l}(s)\right)^{d-1-r}\\
&= \sum_{r=1}^{d}\frac{1}{r}\binom{d-1}{r-1}(\tilde{q}_{k,l-1}(s)-\tilde{q}_{k,l}(s))^{r-1}(\tilde{q}_{k,l}(s))^{d-r}\quad\left(\text{Let }\tilde{q}_{k,l}(s) = \sum_{m\in\mathcal{M}}\frac{v_m p_{k,m}}{\delta_k}q_{m,l}(s)\right)\\
&= \frac{(\tilde{q}_{k,l-1}(s))^d-(\tilde{q}_{k,l}(s))^d}{d(\tilde{q}_{k,l-1}(s)-\tilde{q}_{k,l}(s))},
\end{aligned}\tag{4.19}$$

where $\overline{\mathcal{R}} = \{\overline{r} = (r_1, \ldots, r_M) \in \mathbb{N}_0^M : \sum_{m\in\mathcal{M}} r_m = d-1\}$ and $\overline{\mathcal{R}}'(\overline{r}) = \{\overline{r}' = (r'_1, \ldots, r'_M) \in \mathbb{N}_0^M : r'_m \le r_m,\ \forall m \in \mathcal{M}\}$ given $\overline{r} \in \overline{\mathcal{R}}$. Plugging (4.19) into (4.18), we get the desired result. □

4.4. Convergence of the Occupancy Process: General Case

In this section, we will discuss the case in which the sequence $\{G^N\}_N$ is clustered proportionally sparse, which helps us remove the i.i.d. assumption in Theorem 3. Intuitively, if $\{G^N\}_N$ is clustered proportionally sparse, then for each $k \in \mathcal{K}$ and each dispatcher $i \in W_k^N$, the queue-length distribution of its neighborhood will always be close (in an appropriate sense) to the corresponding global weighted queue-length distribution (GWQD). Clustered proportional sparsity ensures that this statement holds uniformly for all occupancy states. Loosely speaking, this statement enables us to make sure that the evolution of the occupancy process happens in the same way for any initial state as in the case of the i.i.d. initial state. For the case of homogeneous systems, the notion of proportional sparsity was introduced in Rutten and Mukherjee (2022). Here, proportional sparsity was defined in a way that for most dispatcher i, the fraction of its neighbors within any subset U of servers is proportional to the size of the subset U. However, because of the heterogeneous compatibility between dispatchers and servers, such a fraction, in the current setup, depends on the corresponding type of the dispatcher as well $\left(\text{see the term } \frac{E_k^N(U)}{E_k^N(V^N)} \text{ in Definition 2}\right)$. Thus, unlike the homogeneous case where the local queue-length distribution (LQD) is directly compared with the global queue-length distribution of the system, for the heterogeneous case, we need to define K types of global weighted queue-length distribution (see Definition 4), where the weights are determined by the asymptotic properties of the graph structure: $(v_m, m \in \mathcal{M})$ and $(p_{k,m}, k \in \mathcal{K}, m \in \mathcal{M})$. Then, we compare the local queue-length distribution of dispatcher i with the global *weighted* queue-length distribution of the corresponding type as defined.

Definition 4. Consider any fixed $N \in \mathbb{N}$ and $k \in \mathcal{K}$. Given the global occupancy $\mathbf{q}^N = (q_{m,l}^N, m \in \mathcal{M}, l \in \mathbb{N}_0)$ of the Nth system, the GWQD of cluster k is defined as $(x_{k,m,l}^N, m \in \mathcal{M}, l \in \mathbb{N}_0)$, where $x_{k,m,l}^N = \frac{v_m p_{k,m}}{\delta_k}(q_{m,l+1}^N - q_{m,l}^N)$.

Also, the local queue-length distribution is defined as follows.

Definition 5. Consider any fixed $N \in \mathbb{N}$ and $k \in \mathcal{K}$. Given the state $(X_j^N, j \in V^N)$ of the Nth system, the LQD of dispatcher $i \in W_k^N$ is defined as $(\hat{x}_{i,m,l}^N, m \in \mathcal{M}, l \in \mathbb{N}_0)$, where $\hat{x}_{i,m,l}^N = \frac{|\{j \in V_m^N : \xi_{i,j}^N = 1 \text{ and } X_j^N = l\}|}{|\mathcal{N}_w^N(i)|}$.

Although the dispatcher following the JSQ(d) policy selects a target server based on its LQD, if its LQD is close (in a suitable sense) to its corresponding GWQD, then the selection can be viewed as if the decision was based on the GWQD. The latter case is easier to analyze. Hence, if a dispatcher's LQD is close to its corresponding GWQD, we call it a *good dispatcher*

Definition 6 (ε-Good Dispatcher). Consider any fixed $N \in \mathbb{N}$ and an $\varepsilon > 0$. Given the state $(X_j^N, j \in V^N)$ of the Nth system, a dispatcher $i \in W_k^N$, $k \in \mathcal{K}$, is ε-good if

$$\sum_{m\in\mathcal{M}} \sum_{l\in\mathbb{N}_0} |\hat{x}_{i,m,l}^N - x_{k,m,l}^N| \le \varepsilon. \tag{4.20}$$

Also, a dispatcher is ε-bad if it is not ε-good.

4.4.1. Consequences of Clustered Proportional Sparsity.

The proof of Theorem 3 relies on the idea that if the local occupancy of each dispatcher within a particular type evolves similar to the global occupancy of that type, then the process-level limiting behavior should not depend on any specific initial state. That is, it will enable us to go beyond the i.i.d. assumption. The first step for this approach to work is to show that almost all dispatchers are ε-good for any $\varepsilon > 0$. Here is where we need the property of clustered proportional sparsity. This is stated in the next proposition.

Proposition 2. *Let $\{G^N\}_N$ be a sequence of clustered proportionally sparse graphs. For any $T \ge 0$ and $\varepsilon_1, \varepsilon_2 > 0$,*

$$\mathbb{P}\left(\sup_{t\in[0,T]} \mathscr{B}_N^{\varepsilon_1}(t) \ge \varepsilon_2 |W^N|\right) \xrightarrow{N\to\infty} 0, \tag{4.21}$$

where $\mathscr{B}_N^{\varepsilon_1}(t)$ is the number of ε_1-bad dispatchers at time t.

The intuition behind Proposition 2 is that the servers of type $m \in \mathcal{M}$ with queue length $l \in \mathbb{N}_0$ form a subset $U_{m,l}^N$ of the server set V^N. If this set is large, then by the clustered proportional sparsity, for any fixed $k \in \mathcal{K}$ and almost all $i \in W_k^N$, the fraction of the dispatcher i's neighbors within $U_{m,l}^N$ is close to $\frac{|E_k^N(U_{m,l}^N)|}{|E_k^N(V^N)|}$, which is close to $x_{k,m,l}^N$ for large-

enough N by Condition 1. Also, in order to deal with the sum over $l \in \mathbb{N}_0$, we will need to establish uniform bounds of the tail of the occupancy process on any finite time interval. The complete proof is given in Appendix D.

4.4.2. Coupling with an Intermediate System. The main methodology for the proof of Theorem 3 is a stochastic coupling with a sequence $\{G'^N\}_{N\geq 1}$ of carefully constructed systems where the evolution of each system G'^N can be coupled with that of the system G^N. For each N, the system G'^N has the same sets of dispatchers and servers as G^N (i.e., $W'N = W^N$ and $V'^N = V^N$). However, the task assignment in G'^N happens differently. To describe the task assignment policy, let us introduce the following notations. Let $X_j'^N(t)$ be the number of tasks (including those in service) in the queue of server $j \in V'^N$ at time t. Let $\mathbf{q}'^N(t) = (q_{m,l}'^N(t), m \in \mathcal{M}, l \in \mathbb{N}_0)$ be the corresponding global occupancy at time t, which is defined in the same way as $\mathbf{q}^N$ for the system G^N. Then, the system G'^N assigns tasks under the GWSQ(d) policy as described in Algorithm 1. The GWSQ(d) policy is essentially a variant of the JSQ(d) policy because for each new task, the dispatcher selects a target set of servers of size d according to the global weighted queue-length distribution.

Algorithm 1 (GWSQ(d))

while *A new task arrives at dispatcher* $i \in W_k^N$, $k \in \mathcal{K}$ **do**

Get the current global occupancy $\mathbf{q}^N = (q_{m,l}^N, m \in \mathcal{M}, l \in \mathbb{N}_0)$;

Calculate the *global weighted queue-length distribution* $\mathbf{x}_k^N = (x_{k,m,l}^N, m \in \mathcal{M}, l \in \mathbb{N}_0)$ of cluster k,

$$x_{k,m,l}^N = \frac{v_m p_{k,m}}{\delta_k}(q_{m,l+1}^N - q_{m,l}^N);$$

Randomly select a set $\texttt{select}^N$ with size d as the following.

- Let $Y_{k,m,l}^N(t) \in \mathbb{N}_0$ be the number of servers of type $m \in \mathcal{M}$ with queue length $l \in \mathbb{N}_0$ in the set $\texttt{select}^N$;
- $(Y_{k,m,l}^N(t), m \in \mathcal{M}, l \in N_0)$ satisfies

$$\sum_{m\in\mathcal{M}, l\in\mathbb{N}_0} Y_{k,m,l}^N(t) = d;$$

- the probability of selecting $(Y_{k,m,l}^N(t), m \in \mathcal{M}, l \in N_0)$ is

$$\mathbb{P}(Y_{k,m,l}^N(t), m \in \mathcal{M}, l \in N_0) = \prod_{m\in\mathcal{M}, l\in\mathbb{N}_0} \binom{X_{k,m,l}^N(t)}{Y_{k,m,l}^N(t)} \Big/ \binom{N}{d};$$

where $X_{k,m,l}^N = N \times x_{k,m,l}^N$.

Get $l^* = \min(l \in \mathbb{N}_0 : \exists k \in \mathcal{K}, m \in \mathcal{M}$ such that $Y_{k,m,l}^N > 0)$;

Assign the task to a type $m \in \mathcal{M}$ server with queue length l^* with probability

$$\frac{Y_{k,m,l^*}^N}{\sum_{m\in\mathcal{M}} Y_{k,m,l^*}^N}.$$

end

Next, we couple the evolution of the system G'^N with that of the system G^N by the *optimal coupling* method. The optimal coupling for two stochastic processes is similar to the maximal coupling for two discrete random variables (say, X and Y), maximizing the probability $\mathbb{P}(X = Y)$.

4.4.2.1. Optimal Coupling. Fix any N. In both systems, within the pool of servers of each type, arrange the servers in the nondecreasing order of their queue lengths (ties are broken arbitrarily). Now, couple the evolution of the system G^N with the system G'^N in the following way.

- Departure. For any $m \in \mathcal{M}$ and $n = 1, \ldots, |V_m^N|$, synchronize the departure epochs of the nth ordered servers of type m in the two systems.
- Arrival. The coupling of arrivals is the tricky part. For this, first synchronize the arrival epochs at each dispatcher i in both systems G'^N and G^N. At an arrival epoch of dispatcher $i \in W_k^N$, let $(\hat{x}_{i,m,l}^N, m \in \mathcal{M}, l \in \mathbb{N}_0)$ be the *local* empirical distribution of dispatcher i in the system G^N and $(x_{k,m,l}'^N, m \in \mathcal{M}, l \in \mathbb{N}_0)$ be the weighted *global* empirical distribution of cluster-k dispatchers in the system G'^N. Then, in the system G^N, the probability that the task will be

assigned to a server of type $m \in \mathcal{M}$ with queue length $l \in \mathbb{N}_0$ is given by

$$p^N_{m,l}(i) := \frac{\sum_{r=1}^{d}\sum_{r_1=1}^{r}\frac{r_1}{r}\binom{|\mathcal{N}^N_w(i)|\hat{x}^N_{i,m,l}}{r_1}\binom{|\mathcal{N}^N_w(i)|\sum_{\mathcal{M}\setminus\{m\}}\hat{x}^N_{i,m,l}}{r-r_1}\binom{|\mathcal{N}^N_w(i)|\sum_{\mathcal{M}}\sum_{l'\geq l+1}\hat{x}^N_{i,m,l'}}{d-r}}{\binom{|\mathcal{N}^N_w(i)|}{d}}. \tag{4.22}$$

In the system G'^N, the probability that the task will be assigned to a server of type $m \in \mathcal{M}$ with queue length $l \in \mathbb{N}_0$ is given by

$$p'^N_{m,l}(k) := \frac{\sum_{r=1}^{d}\sum_{r_1=1}^{r}\frac{r_1}{r}\binom{X'^N_{k,m,l}}{r_1}\binom{\sum_{\mathcal{M}\setminus\{m\}}X'^N_{k,m,l}}{r-r_1}\binom{\sum_{\mathcal{M}}\sum_{l'\geq l+1}X'^N_{k,m,l'}}{d-r}}{\binom{N}{d}}. \tag{4.23}$$

For convenience, we denote $p^N_{m,l}(i)$ and $p'^N_{m,l}(k)$ as $p^N_{m,l}$ and $p'^N_{m,l}$, respectively. Denote $\overline{p}^N_{m,l} = \min(p^N_{m,l}, p'^N_{m,l})$ for $m \in \mathcal{M}$ and $l \in \mathbb{N}_0$.

Now, to couple the task assignment, let us draw a Uniform[0,1] random variable U, independently of any other processes and across various arrival epochs. U is used to generate the random variables $(M^N, L^N) \in \mathcal{M} \times \mathbb{N}_0$ and $(M'^N, L'^N) \in \mathcal{M} \times \mathbb{N}_0$ for the system G^N and the system G'^N, respectively. In the system G^N, set $(M^N, L^N) = (m,l) \in \mathcal{M} \times \mathbb{N}_0$ if

$$\begin{aligned} U \in &\left[\sum_{m'=1}^{m-1}\sum_{l'=0}^{\infty}\overline{p}^N_{m',l'} + \sum_{l'=0}^{l-1}\overline{p}^N_{m,l'}, \sum_{m'=1}^{m-1}\sum_{l'=0}^{\infty}\overline{p}^N_{m',l'} + \sum_{l'=0}^{l}\overline{p}^N_{m,l'}\right) \\ &\bigcup \left[\overline{p}^N + \sum_{m'=1}^{m-1}\sum_{l'=0}^{\infty}(p^N_{m',l'} - \overline{p}^N_{m',l'}) + \sum_{l'=0}^{l-1}(p^N_{m,l'} - \overline{p}^N_{m,l'}),\right. \\ &\left.\overline{p}^N + \sum_{m'=1}^{m-1}\sum_{l'=0}^{\infty}(p^N_{m',l'} - \overline{p}^N_{m',l'}) + \sum_{l'=0}^{l}(p^N_{m,l'} - \overline{p}^N_{m,l'})\right), \end{aligned} \tag{4.24}$$

where $\overline{p}^N = \sum_{m'=1}^{M}\sum_{l'=0}^{\infty}\overline{p}^N_{m',l'}$, and assign the task to a server of type m with queue length l. Similarly, in the system G'^N, set $(M'^N, L'^N) = (m,l) \in \mathcal{M} \times \mathbb{N}_0$, if

$$\begin{aligned} U \in &\left[\sum_{m'=1}^{m-1}\sum_{l'=0}^{\infty}\overline{p}^N_{m',l'} + \sum_{l'=0}^{l-1}\overline{p}^N_{m,l'}, \sum_{m'=1}^{m-1}\sum_{l'=0}^{\infty}\overline{p}^N_{m',l'} + \sum_{l'=0}^{l}\overline{p}^N_{m,l'}\right) \\ &\bigcup \left[\overline{p}^N + \sum_{m'=1}^{m-1}\sum_{l'=0}^{\infty}(p'^N_{m',l'} - \overline{p}^N_{m',l'}) + \sum_{l'=0}^{l-1}(p'^N_{m,l'} - \overline{p}^N_{m,l'}),\right. \\ &\left.\overline{p}^N + \sum_{m'=1}^{m-1}\sum_{l'=0}^{\infty}(p'^N_{m',l'} - \overline{p}^N_{m',l'}) + \sum_{l'=0}^{l}(p'^N_{m,l'} - \overline{p}^N_{m,l'})\right), \end{aligned} \tag{4.25}$$

and assign the task to a server of type m with queue length l.

As alluded to before, the coupling is constructed in a way that maximizes the probability of the two systems to assign an arriving task to some server with the same queue length. Next, the difference in the occupancy processes of the two systems, on any finite time interval, can be upper bounded by the number of times the two systems assign to two different queue lengths. This is formalized by the notion of *mismatch*, which was originally introduced in Mukherjee et al. (2018b).

Definition 7 (Mismatch). At an arrival epoch, the system G^N and the system G'^N are said to mismatch if $(M^N, L^N) \neq (M'^N, L'^N)$; that is, the arriving task is not assigned to servers of the same type with the same queue length in the two systems. Denote by $\Delta^N(t)$ the cumulative number of times the systems mismatch in queue length up to time t.

The next proposition provides a deterministic bound on the difference between the occupancy processes of the two systems in terms of the number of mismatches.

Proposition 3. *For any $N \geq 1$, consider the system G^N and the system G'^N coupled. Then, the following holds almost surely on the coupled probability space: for $t \geq 0$,*

$$\sum_{m\in\mathcal{M}}\sum_{l\in\mathbb{N}_0} |Q^N_{m,l}(t) - Q'^N_{m,l}(t)| \leq 2\Delta^N(t), \tag{4.26}$$

provided the inequality holds at $t = 0$. $Q^N_{m,l}(t)$ and $Q'^N_{m,l}(t)$ represent the number of servers of type $m \in \mathcal{M}$ with queue length at least $l \in \mathbb{N}_0$ in the system G^N and the system G'^N at time t, respectively.

Bounds of the form as given in (4.26) were originally established in Mukherjee et al. (2018b, proposition 4), and they were later used in various contexts (Mukherjee et al. 2018a, Rutten and Mukherjee 2022). The proof does not depend on any specific assignment policy and relies on showing inductively that if the inequality in (4.26) holds before an event time epoch, then it is preserved after the event time epoch as well. The proof of Proposition 3 can be obtained following the similar arguments. We omit the details.

Lemma 8. *Given $\sum_{m\in\mathcal{M}}\sum_{l\in\mathbb{N}_0}|Q^N_{m,l} - Q'^N_{m,l}| \leq 2\Delta^N$, then there exist $N_0 \in \mathbb{N}_0$ and a positive constant L such that for any $k \in \mathcal{K}$,*

$$\sum_{m\in\mathcal{M}}\sum_{l\in\mathbb{N}_0} |x^N_{k,m,l} - x'^N_{k,m,l}| \leq L\Delta^N/N, \quad \forall N \geq N_0. \tag{4.27}$$

Proof. By the model assumption, there exists $N_0 \in \mathbb{N}_0$ such that for all $N \geq N_0$, $|V^N_m| \geq \frac{1}{2}Nv_m$, $\forall m \in \mathcal{M}$, which gives us that

$$\begin{aligned}\sum_{m\in\mathcal{M}}\sum_{l\in\mathbb{N}_0} |x^N_{k,m,l} - x'^N_{k,m,l}| &= \sum_{m\in\mathcal{M}}\sum_{l\in\mathbb{N}_0} \frac{v_m p_{k,m}}{\delta_k}|Q^N_{m,l} - Q'^N_{m,l}|/|V^N_m| \\ &\leq \sum_{m\in\mathcal{M}}\sum_{l\in\mathbb{N}_0} \frac{2p_{k,m}}{\delta_k}|Q^N_{m,l} - Q'^N_{m,l}|/N \leq L\Delta^N/N,\end{aligned} \tag{4.28}$$

where $L = 4\max_{k\in\mathcal{K},m\in\mathcal{M}}\frac{p_{k,m}}{\delta_k}$. □

The final ingredient that we need is the probability of mismatch in a particular epoch under the optimal coupling method. The next lemma bounds this probability in terms of the ℓ_1-distance between the LQD of the G_N system and the GWQD of the G'^N system.

Lemma 9. *Consider an arrival epoch at dispatcher i, and assume that in this epoch, the LQD in the system G^N is given by $(\hat{x}^N_{i,m,l}, m \in \mathcal{M}, l \in \mathbb{N}_0)$ and the GWQD of cluster-k servers in the system G'^N is given by $(x'^N_{k,m,l}, m \in \mathcal{M}, l \in \mathbb{N}_0)$. Then, there exists a finite positive constant L_1 such that for all large-enough N,*

$$\mathbb{P}(\mathit{Mismatch}) \leq L_1 \sum_{m\in\mathcal{M}}\sum_{l\in\mathbb{N}_0} |\hat{x}^N_{i,m,l} - x'^N_{k,m,l}|. \tag{4.29}$$

The key step in the proof of Lemma 9 is that given the queue-length distribution $\mathbf{x} = (x_{m,l}, m \in \mathcal{M}, l \in \mathbb{N}_0)$, the probability $p_{m,l}$ that a task will be assigned to a server of type $m \in \mathcal{M}$ with queue length $l \in \mathbb{N}_0$ can be approximated by

$$p_{m,l} \approx \sum_{r=1}^{d}\sum_{r_1=1}^{r}\frac{r_1}{r}\frac{d!}{r_1!(r-r_1)!(d-r)!}(x_{m,l})^{r_1}\left(\sum_{\mathcal{M}\setminus\{m\}} x_{m,l}\right)^{r-r_1}\left(\sum_{\mathcal{M}}\sum_{l'\geq l+1} x_{m,l'}\right)^{d-r}$$

and that the function x^k is Lipschitz for $x \in [0,1]$. The complete proof is given in Appendix E.

4.4.3. Proof of Theorem 3. Now, we have all the ingredients to prove Theorem 3. Let us explain the high-level proof scheme first.

Step 1. Using the optimal coupling, we will show that the global occupancy processes $\{\mathbf{q}^N(\cdot)\}_N$ and $\{\mathbf{q}'^N(\cdot)\}_N$ must converge to the same limit process as $N \to \infty$ if their initial states are the same, $X^N_j(0) = X'^N_j(0)$ for all j. In

other words, with the same initial states,

$$\lim_{N\to\infty} \mathbf{q}^N(\cdot) = \lim_{N\to\infty} \mathbf{q}'^N(\cdot).$$

Step 2. Because there is no graph structure in the system G'^N, all servers of the same type in the system G'^N are exchangeable. Hence, $\mathbf{q}'^N(\cdot)$ is Markovian, which implies that given $\mathbf{q}'^N(0)$, its evolution does not depend on how individual $X_j'^N(0)$'s are distributed. Denote the system G'^N with the i.i.d. assumption as $G_1'^N$, where the i.i.d. assumption refers to that for any $m \in \mathcal{M}$, $X_j^N(0)$, $j \in V_m^N$, are i.i.d. Also, denote the system G'^N without the i.i.d. assumption as $G_2'^N$. Their occupancy processes are $\mathbf{q}_1'^N(\cdot)$ and $\mathbf{q}_2'^N(\cdot)$, respectively. Because task assignment policy in G'^N does not distinguish between two servers having the same type and queue lengths, by a natural coupling, $\mathbf{q}_1'^N(t) = \mathbf{q}_2'^N(t)$ holds for all $t \geq 0$, implying that

$$\lim_{N\to\infty} \mathbf{q}_1'^N(\cdot) = \lim_{N\to\infty} \mathbf{q}_2'^N(\cdot).$$

Step 3. Denote the system G^N with the i.i.d. assumption as G_1^N and the system G'^N without the i.i.d. assumption as G_2^N, and denote their occupancy processes by $\mathbf{q}_1^N(\cdot)$ and $\mathbf{q}_2^N(\cdot)$, respectively. Combining Step 1 and Step 2, the following equation holds. With the same initial global occupancy state,

$$\lim_{N\to\infty} \mathbf{q}_1^N(\cdot) = \lim_{N\to\infty} \mathbf{q}_1'^N(\cdot) = \lim_{N\to\infty} \mathbf{q}_2'^N(\cdot) = \lim_{N\to\infty} \mathbf{q}_2^N(\cdot),$$

where the first and last equalities are because of Step 1 and the second equality is because of Step 2.

Step 4. Use Theorem 2 to note that when the sequence $\{G^N\}_N$ satisfies the assumption that for each $m \in \mathcal{M}$, $X_j^N(0)$, $j \in V_m^N$, are i.i.d., the scaled global occupancy process $\mathbf{q}^N$ converge weakly to $\mathbf{q}$ described by the system of ODEs in (3.7).

Step 5. By Steps 3 and 4, Theorem 3 holds.

In the proof scheme, observe that all that remains is to show Step 1, which is given here.

Proof of Theorem 3. For Step 1 described in the proof scheme, by Proposition 3, it is sufficient to show that for any $\varepsilon^* > 0$ and $\delta^* > 0$, there exists an $N_0 \geq 1$ such that

$$\mathbb{P}\left(\sup_{t\in[0,T]} \Delta^N(t)/N \geq \varepsilon^*\right) \leq \delta^*, \quad \forall N \geq N_0. \tag{4.30}$$

Fix an $\varepsilon > 0$, which will be chosen later. Let $\mathscr{G}_N^\varepsilon(t)$ and $\mathscr{B}_N^\varepsilon(t)$ be the numbers of ε-good and ε-bad dispatchers in the system G^N at time t, respectively. We couple the evolution of the system G^N with that of the system G'^N by the optimal coupling method. In system G^N, let $(x_{k,m,l}^N(t), m \in \mathcal{M}, l \in \mathbb{N}_0)$ be the global weighted queue-length distribution of cluster $k \in \mathcal{K}$ and $(\hat{x}_{i,m,l}^N(t), m \in \mathcal{M}, l \in \mathbb{N}_0)$ be the local queue-length distribution of the dispatcher $i \in W_k^N$, $k \in \mathcal{K}$. Also, let $(x_{k,m,l}'^N(t), m \in \mathcal{M}, l \in \mathbb{N}_0)$ be the global weighted queue-length distribution of cluster $k \in \mathcal{K}$ in system G'^N. Denote $\rho_k^N(t) = \sum_{m\in\mathcal{M}}\sum_{l\in\mathbb{N}_0} |x_{k,m,l}^N(t) - x_{k,m,l}'^N(t)|$. At an arrival epoch $t \geq 0$, if a task arrives at an ε-good dispatcher $i \in W_k^N$, then

$$\begin{aligned}
&\sum_{m\in\mathcal{M}}\sum_{l\in\mathbb{N}_0} |\hat{x}_{i,m,l}^N(t-) - x_{k,m,l}'^N(t-)| \\
&\leq \sum_{m\in\mathcal{M}}\sum_{l\in\mathbb{N}_0} |\hat{x}_{i,m,l}^N(t-) - x_{k,m,l}^N(t-)| + \sum_{m\in\mathcal{M}}\sum_{l\in\mathbb{N}_0} |x_{k,m,l}^N(t-) - x_{k,m,l}'^N(t-)| = \varepsilon + \rho_k^N(t).
\end{aligned} \tag{4.31}$$

Recall the uniform random variable U and $\overline{p}_{m,l}^N$ defined in the description of the optimal coupling method. The probability that the systems have a mismatch at such arrival epoch is bounded by

$$\begin{aligned}
\mathbb{P}\left(U \notin \left[0, \sum_{m\in\mathcal{M}}\sum_{l\in\mathbb{N}_0} \overline{p}_{m,l}^N\right]\right) &= 1 - \sum_{m\in\mathcal{M}}\sum_{l\in\mathbb{N}_0} \overline{p}_{m,l}^N = \sum_{m\in\mathcal{M}}\sum_{l\in\mathbb{N}_0} p_{m,l}'^N - \sum_{m\in\mathcal{M}}\sum_{l\in\mathbb{N}_0} \overline{p}_{m,l}^N \\
&\leq \sum_{m\in\mathcal{M}}\sum_{l\in\mathbb{N}_0} |p_{m,l}'^N - p_{m,l}^N| \leq L_1 \sum_{m\in\mathcal{M}}\sum_{l\in\mathbb{N}_0} |\hat{x}_{i,m,l}^N(t-) - x_{k,m,l}'^N(t-)| \leq L_1(\rho_k^N(t) + \varepsilon),
\end{aligned} \tag{4.32}$$

where the second inequality is from Lemma 9. At an arrival epoch $t \geq 0$, if a task arrives at an ε-bad dispatcher $i \in W_k^N$, then with probability at most one, the systems have a mismatch. Because of the Poisson thinning property, we can construct an independent unit-rate Poisson process $(Z(t))_{t\geq 0}$ so that $\Delta^N(t)$ can be upper bounded by a random time change of Z as the following; for all $t \in [0,T]$,

$$\begin{aligned}\Delta^N(t) &\leq Z\left(\sum_{k\in\mathcal{K}}\sum_{i\in W_k^N}\lambda\int_0^t [\mathbb{1}_{(i\in\mathscr{G}_N^\varepsilon(s-))}L_1(\rho_k^N(s-)+\varepsilon)+\mathbb{1}_{(i\in\mathscr{B}_N^\varepsilon(s-))}\cdot 1]ds\right)\\ &\leq Z\left(\sum_{k\in\mathcal{K}}\sum_{i\in W_k^N}\lambda\int_0^t [\mathbb{1}_{(i\in\mathscr{G}_N^\varepsilon(s-))}L_1(L\Delta^N(s-)/N+\varepsilon)+\mathbb{1}_{(i\in\mathscr{B}_N^\varepsilon(s-))}\cdot 1]ds\right)\\ &= Z\left(\lambda\int_0^t [\mathscr{G}_N^\varepsilon(s-)L_1(L\Delta^N(s-)/N+\varepsilon)+\mathscr{B}_N^\varepsilon(s-)\cdot 1]ds\right),\end{aligned} \tag{4.33}$$

where the second inequality is because of Lemma 8. By Proposition 2, we have that for any $\varepsilon' > 0$, there exists an $N(\varepsilon')$ such that for all $N \geq N(\varepsilon')$,

$$\mathbb{P}\left(\sup_{t\in[0,T]}\mathscr{B}_N^\varepsilon(t) \geq \varepsilon' |W^N|\right) \leq \frac{\varepsilon'}{2}. \tag{4.34}$$

Hence, by (4.33), (4.34), and Tonelli's theorem, we have that for all $N \geq N(\varepsilon')$ and $t \in [0,T]$,

$$\mathbb{E}\left(\frac{\Delta^N(t)}{N}\right) \leq \lambda\int_0^t\left[L_1\left(L\frac{W(N)}{N}\frac{\mathbb{E}(\Delta^N(s-))}{N}+\varepsilon\right)+\frac{W(N)}{N}\frac{3\varepsilon'}{2}\right]ds. \tag{4.35}$$

Also, by the assumption that $\lim_{N\to\infty}\frac{W(N)}{N} = \zeta$, there exists N_0 such that $\frac{W(N)}{N} \leq 2\zeta$. Hence, we have that for all $N \geq \max(N(\varepsilon'), N_0)$ and $t \in [0,T]$,

$$\mathbb{E}\left(\frac{\Delta^N(t)}{N}\right) \leq \lambda\int_0^t\left[L_1\left(2L\zeta\frac{\mathbb{E}(\Delta^N(s-))}{N}+\varepsilon\right)+3\zeta\varepsilon'\right]ds. \tag{4.36}$$

By applying Grönwall's inequality to (4.36), we have

$$\mathbb{E}\left(\frac{\Delta^N(t)}{N}\right) \leq \lambda(L_1\varepsilon+3\zeta\varepsilon')t\,\exp(2LL_1\zeta\lambda t). \tag{4.37}$$

Because $\Delta^N(t)$ is nonnegative, by Markov's inequality and (4.37), we have

$$\mathbb{P}\left(\sup_{t\in[0,T]}\Delta^N(t)/N \geq \varepsilon^*\right) \leq \frac{1}{\varepsilon^*}\lambda(L_1\varepsilon+3\zeta\varepsilon')t\,\exp(2LL_1\zeta\lambda t), \tag{4.38}$$

and we can choose small-enough ε and ε' such that (4.30) holds. □

5. Proof of Interchange of Limits

5.1. Properties of the Limiting System of ODEs

First, we define the fixed point of the ODE (3.7). Recall $\delta_k = \sum_{m\in\mathcal{M}}p_{k,m}v_m$ and $\tilde{q}_{k,l}(t) = \sum_{m\in\mathcal{M}}\frac{v_m p_{k,m}}{\delta_k}q_{m,l}(t)$. Let $\mathbf{q}^* = (q^*_{m,l} \in \mathbb{R}_+, m \in \mathcal{M}, l \in \mathbb{N}_0)$ be a fixed point of the ODE (3.7) if for all $m \in \mathcal{M}, l \in \mathbb{N}$,

$$u_m(q^*_{m,l}-q^*_{m,l+1}) = \lambda\zeta(q^*_{m,l-1}-q^*_{m,l})\sum_{k\in\mathcal{K}}\frac{p_{k,m}w_k}{\delta_k}\frac{(\tilde{q}^*_{k,l-1})^d-(\tilde{q}^*_{k,l})^d}{\tilde{q}^*_{k,l-1}-\tilde{q}^*_{k,l}}, \tag{5.1}$$

with $q^*_{m,0} = 1$, $m \in \mathcal{M}$. The next proposition shows some important properties of the fixed point $\mathbf{q}$ of the ODE (3.7).

Proposition 4. *If there exists a fixed point* $\mathbf{q}^*$ *of the ODE* (3.7) *such that for each* $m \in \mathcal{M}$, $q_{m,0} = 1$ *and* $q_{m,l} \overset{l\to\infty}{\to} 0$, *then for each* $m \in \mathcal{M}$, *the sequence* $\{q_{m,l}, l \in \mathbb{N}_0\}$ *decreases doubly exponentially*.

The proof of Proposition 4 is provided in Appendix F. The key observation used in the proof is that by (5.1), $q^*_{m,l}$ can be expressed in terms of $q^*_{m,l-1}$ and $q^*_{m,l-2}$. Thus, we can recursively characterize the values of $q^*_{m,l}$, $l \geq 2$, if we know $q^*_{m,0}$ and $q^*_{m,1}$, $m \in \mathcal{M}$.

By Proposition 4, we know that if $\mathbf{q}^*$ is a fixed point of the ODE (3.7) and for all $m \in \mathcal{M}$, $q^*_{m,l} \overset{l\to\infty}{\to} 0$, then such $\mathbf{q}^*$ must be in $\mathcal{S}$, so we only need to show that such $\mathbf{q}^*$ exists. For the proof of the existence of such $\mathbf{q}^*$, we need a technical lemma, which will be used in (5.4).

Lemma 10. *Consider a sequence* $\{G^N\}_N$ *satisfying Condition 1. If* $\{G^N\}_N$ *is proportionally sparse and in the subcritical regime, then for any* $(\alpha_1, \ldots, \alpha_M) \in [0,1]^M$ *with* $\sum_{m\in\mathcal{M}} \alpha_m > 0$, *the following holds*:

$$\left(\sum_{m\in\mathcal{M}} \alpha_m v_m u_m\right)^{-1} \lambda\zeta \sum_{k\in\mathcal{K}} w_k \left(\frac{\sum_{m\in\mathcal{M}} \alpha_m p_{k,m} v_m}{\delta_k}\right)^d \leq \rho < 1. \tag{5.2}$$

The proof of Lemma 10 is provided in Appendix G.

Proof of Theorem 4. We prove the *existence* of the fixed point first. From (5.1), we know that if $(q^*_{m,1}, m \in \mathcal{M})$ are fixed, then all $(q^*_{m,l}, m \in \mathcal{M}, l \geq 2)$ are determined as well. Hence, $\mathbf{q}^*$ can be the viewed as the function of $(q^*_{m,1}, m \in \mathcal{M})$. Moreover, in the steady state, $\sum_{m\in\mathcal{M}} q^*_{m,1} = \lambda\zeta$, which implies that $q^*_{M,1}$ can be decided by the values of $q^*_{m,1}$, $m \in \mathcal{M} \setminus \{M\}$. Hence, we construct the sequence $\mathbf{q}(\overline{\alpha}) = (q_{m,l}(\overline{\alpha}), m \in \mathcal{M}, l \in \mathbb{N}_0)$ as functions of the vector $\overline{\alpha} = (\alpha_1, \ldots, \alpha_{M-1}) \in (0,1)^{M-1}$ as follows:

$$\begin{aligned}
&q_{m,0}(\overline{\alpha}) = 1, \ \forall m \in \mathcal{M},\\
&q_{m,1}(\overline{\alpha}) = \alpha_m, \quad m \in \mathcal{M} \setminus \{M\}, \quad \text{and} \quad q_{M,1} = \frac{\lambda\zeta - \sum_{m\in\mathcal{M}\setminus\{M\}} \alpha_m v_m u_m}{v_M u_M},\\
&u_m(q_{m,l}(\overline{\alpha}) - q_{m,l+1}(\overline{\alpha})) = \lambda\zeta(q_{m,l-1}(\overline{\alpha}) - q_{m,l}(\overline{\alpha})) \sum_{k\in\mathcal{K}} \frac{p_{k,m} w_k}{\delta_k} \frac{(\tilde{q}_{k,l-1}(\overline{\alpha}))^d - (\tilde{q}_{k,l}(\overline{\alpha}))^d}{\tilde{q}_{k,l-1}(\overline{\alpha}) - \tilde{q}_{k,l}(\overline{\alpha})}, l \geq 1.
\end{aligned} \tag{5.3}$$

Because for all $m \in \mathcal{M}$, $q_{m,1}(\overline{\alpha})$ should be in (0, 1), then $\overline{\alpha} = (\alpha_1, \ldots, \alpha_{M-1})$ must lie in the polyhedron $\mathbf{P}_1$ defined as follows:

$$\begin{aligned}
\mathbf{P}_1 := \Bigg\{ \alpha_m \in \left(\max\left(0, \frac{\lambda\zeta - \sum_{m'\in\mathcal{M}\setminus\{m\}} v_{m'} u_{m'}}{v_m u_m}\right), \min\left(\frac{\lambda\zeta}{v_m u_m}, 1\right)\right), \ \forall m \leq M-1,\\
\text{and } \lambda\zeta - v_M u_M < \sum_{m\leq M-1} \alpha_m v_m u_m < \lambda\zeta \Bigg\}.
\end{aligned}$$

For all $\overline{\alpha} \in \mathbf{P}_1$, we have $1 = q_{m,0}(\overline{\alpha}) > q_{m,1}(\overline{\alpha}) > 0$, $\forall m \in \mathcal{M}$. Consider $l = 2$. By (5.3), we have that when $\alpha_m = 0$, $m \in \mathcal{M} \setminus \{M\}$,

$$u_m(0 - q_{m,2}(\overline{\alpha})) = \lambda\zeta(1-0) \sum_{k\in\mathcal{K}} \frac{p_{k,m} w_k}{\delta_k} \frac{1 - (\tilde{q}_{k,1}(\overline{\alpha}))^d}{1 - \tilde{q}_{k,1}(\overline{\alpha})},$$

implying that $q_{m,2}(\overline{\alpha}) < 0$; when $\alpha_m = 1$, $m \in \mathcal{M} \setminus \{M\}$,

$$u_m(1 - q_{m,2}(\overline{\alpha})) = 0,$$

implying that $q_{m,2}(\overline{\alpha}) = 1 > 0$. When $\alpha_m = \frac{\lambda\zeta}{v_m u_m}$, $m \in \mathcal{M} \setminus \{M\}$,

$$u_m\left(\frac{\lambda\zeta}{v_m u_m} - q_{m,2}(\overline{\alpha})\right) = \lambda\zeta\left(1 - \frac{\lambda\zeta}{v_m u_m}\right) \sum_{k\in\mathcal{K}} \frac{p_{k,m} w_k}{\delta_k} \frac{1 - (\tilde{q}_{k,1}(\overline{\alpha}))^d}{1 - \tilde{q}_{k,1}(\overline{\alpha})},$$

implying that

$$\begin{aligned} q_{m,2}(\overline{\alpha}) &= \frac{\lambda\zeta}{v_m u_m} - \frac{\lambda\zeta}{u_m}\left(1 - \frac{\lambda\zeta}{v_m u_m}\right)\sum_{k\in\mathcal{K}} \frac{p_{k,m} w_k}{\delta_k} \frac{1-(\tilde{q}_{k,1}(\overline{\alpha}))^d}{1-\tilde{q}_{k,1}(\overline{\alpha})} \\ &> \frac{\lambda\zeta}{v_m u_m} - \frac{\lambda\zeta}{u_m}\left(1 - \frac{\lambda\zeta}{v_m u_m}\right)\sum_{k\in\mathcal{K}} \frac{p_{k,m} w_k}{\delta_k} \\ &> \frac{\lambda\zeta}{v_m u_m} - \frac{\lambda\zeta}{u_m}\left(1 - \frac{\lambda\zeta}{v_m u_m}\right)\sum_{k\in\mathcal{K}} \frac{p_{k,m} w_k}{p_{k,m} v_m} \\ &= \frac{\lambda\zeta}{v_m u_m} - \frac{\lambda\zeta}{v_m u_m}\left(1 - \frac{\lambda\zeta}{v_m u_m}\right) > 0. \end{aligned}$$

Let $r_{m,1}$, $m \le M-1$ be the maximum number, which satisfies the following:

1. $r_{m,1} < \min\left(\frac{\lambda\zeta}{v_m u_m}, 1\right)$,
2. $\exists \overline{\alpha} \in \mathbf{P}_1$ with $\alpha_m = r_{m,1}$ such that $q_{m,2}(\overline{\alpha}) = 0$.

Define $\mathbf{P}_1' \subseteq \mathbf{P}_1$ as the following:

$$\mathbf{P}_1' := \left\{ \alpha_m \in \left(\max\left(r_{m,1}, \frac{\lambda\zeta - \sum_{m\le M-1} v_{m'} u_{m'}}{v_m u_m} \right), \min\left(\frac{\lambda\zeta}{v_m u_m}, 1 \right) \right),\ \forall m \le M-1, \right.$$
$$\left. \text{and } \lambda\zeta - v_M u_M < \sum_{m\le M-1} \alpha_m v_m u_m < \lambda\zeta \right\}.$$

Again, by using (5.3), we get that when $\sum_{m\le M-1}\alpha_m u_m = \lambda\zeta - v_M u_M$ (i.e., $q_{M,1}(\overline{\alpha}) = 1$),

$$u_M(1 - q_{M,2}(\overline{\alpha})) = 0,$$

implying that $q_{M,2}(\overline{\alpha}) = 1 > 0$; when $\sum_{m\le M-1}\alpha_m u_m = \lambda\zeta$ (i.e., $q_{M,1}(\overline{\alpha}) = 0$),

$$u_M(0 - q_{M,2}(\overline{\alpha})) = \lambda\zeta(1-0)\sum_{k\in\mathcal{K}} \frac{p_{k,M} w_k}{\delta_k} \frac{1-(\tilde{q}_{k,1}(\overline{\alpha}))^d}{1-\tilde{q}_{k,1}(\overline{\alpha})},$$

implying that $q_{M,2}(\overline{\alpha}) < 0$.

Let r_1 be the minimum number that satisfies the following:

1. $r_1 < \lambda\zeta$.
2. There exists $\overline{\alpha} \in \mathbf{P}_1'$ such that $\sum_{m\le M-1}\alpha_m v_m u_m = r_1$ and $q_{M,2}(\overline{\alpha}) = 0$.

Define $\mathbf{P}_2 \subseteq \mathbf{P}_1' \subseteq \mathbf{P}_1$ as the following:

$$\mathbf{P}_2 := \left\{ \alpha_m \in \left(\max\left(r_{m,1}, \frac{\lambda\zeta - \sum_{m'\le M-1} v_{m'} u_{m'}}{v_m u_m} \right), \min\left(\frac{r_1}{v_m u_m}, 1 \right) \right),\ \forall m \le M-1 \right.$$
$$\left. \text{and } \lambda\zeta - v_M u_M \le \sum_{m\le M-1} \alpha_m v_m u_m \le r_1 \right\}.$$

Hence, for all $\overline{\alpha} \in \mathbf{P}_2$, we have $1 = q_{m,0}(\overline{\alpha}) > q_{m,1}(\overline{\alpha}) > q_{m,2}(\overline{\alpha}) > 0,\ \forall m \in \mathcal{M}$. Continuing this process, we can define a sequence $\{\mathbf{P}_1 \supseteq \mathbf{P}_2 \supseteq \cdots\}$ of polyhedra such that for all $\overline{\alpha} \in \mathbf{P}_n$, we have $1 = q_{m,0}(\overline{\alpha}) > q_{m,1}(\overline{\alpha}) > \cdots > q_{m,n}(\overline{\alpha}) > 0,\ \forall m \in \mathcal{M}$. Thus, we can get decreasing sequences $\{q_{m,l}(\overline{\alpha})\}_{l\in\mathbb{N}_0}$, $m \in \mathcal{M}$ for some $\overline{\alpha}$. Because $q_{m,l} \ge 0,\ \forall m \in \mathcal{M}$, $l \in \mathbb{N}_0$, then $\forall m \in \mathcal{M}$, $\exists x_m^*$ such that $\lim_{l\to\infty} q_{m,l}(\overline{\alpha}) = x_m^*$. Next, we need to show that $x_m^* = 0,\ \forall m \in \mathcal{M}$. By (F.2), we have

$$\sum_{m\in\mathcal{M}} v_m u_m x_m^* = \lambda\zeta \sum_{k\in\mathcal{K}} w_k \left(\sum_{m\in\mathcal{M}} \frac{p_{k,m} v_m}{\delta_k} x_m^* \right)^d. \tag{5.4}$$

Clearly, $x^*_m = 0,\ \forall m \in \mathcal{M}$ is a solution of (5.4). It must be the unique solution because by Lemma 10, for all $(x^*_m, m \in \mathcal{M}) \in [0,1]^M$ with $\sum_{m\in\mathcal{M}} x^*_m > 0$,

$$\left(\sum_{m\in\mathcal{M}} v_m u_m x^*_m\right)^{-1} \lambda\zeta \sum_{k\in\mathcal{K}} w_k \left(\sum_{m\in\mathcal{M}} \frac{p_{k,m} v_m}{\delta_k} x^*_m\right)^d < 1,$$

implying that (5.4) does not hold. Now, let $q^*_{m,l} = q_{m,l}(\overline{\alpha}),\ \forall m \in \mathcal{M}, l \in \mathbb{N}_0$.

Now, we are going to show the *uniqueness*. The proof of the uniqueness is based on a monotonicity property of the system, which is stated in the following claim.

Claim 1. If $\mathbf{q} \le \hat{\mathbf{q}}$ for $\mathbf{q}, \hat{\mathbf{q}} \in \mathcal{S}$, then $\overline{\mathbf{q}}(t, \mathbf{q}) \le \overline{\mathbf{q}}(t, \hat{\mathbf{q}})$ for all t.

Proof. Consider any $\mathbf{q} \le \hat{\mathbf{q}} \in \mathcal{S}$. It is easy to construct two copies of the Nth systems with initial states $\{X^N_j(0), j \in V^N\}$ and $\{\hat{X}^N_j(0), j \in V^N\}$ satisfying the following.

1. For all $j \in V^N$, $X^N_j(0) \le \hat{X}^N_j(0)$.
2. $\{X^N_j(0), j \in V^N\}$ has the corresponding global occupancy $\mathbf{q}^N(0) = \mathbf{q} \in \mathcal{S}$; similarly, $\{\hat{X}^N_j(0), j \in V^N\}$ has $\hat{\mathbf{q}}^N(0) = \hat{\mathbf{q}} \in \mathcal{S}$.

By a natural coupling, we have that for all $j \in V^N$ and $t \ge 0$, $X^N_j(t) \le \hat{X}^N_j(t)$, implying that $\mathbf{q}^N(t) \le \hat{\mathbf{q}}^N(t)$. Because systems are stable, then $\mathbf{q}^N(t), \hat{\mathbf{q}}^N(t) \in \mathcal{S}$ for all $t \ge 0$. Moreover, by Theorem 2, the claim follows. □

We continue the proof of the uniqueness. Now, it is sufficient to show that $\lim_{t\to\infty} \overline{\mathbf{q}}(t, \mathbf{q}_0) = \mathbf{q}^*$, in which either $\mathbf{q}_0 \le \mathbf{q}^*$ or $\mathbf{q}_0 \ge \mathbf{q}^*$ component wise, because Claim 1 implies that

$$\overline{\mathbf{q}}(t, \min(\mathbf{q}_0, \mathbf{q}^*)) \le \overline{\mathbf{q}}(t, \mathbf{q}_0) \le \overline{\mathbf{q}}(t, \max(\mathbf{q}_0, \mathbf{q}^*)), \quad \forall \mathbf{q}_0 \in \overline{\mathbf{S}}, t \ge 0.$$

We will prove the case that if $\mathbf{q}_0 \le \mathbf{q}^*$, then

$$\lim_{t\to\infty} \overline{\mathbf{q}}(t, \mathbf{q}_0) = \mathbf{q}^*.$$

The case that $\mathbf{q}_0 \ge \mathbf{q}^*$ is similar. Also, note that $q_{m,l}(\infty),\ \forall m \in \mathcal{M}, l \ge 2$ can be solved recursively by (5.1) when $q_{m,1}(\infty),\ \forall m \in \mathcal{M}$ are determined, so it is sufficient to show that $q_{m,1}(\infty) = q^*_{m,1},\ \forall m \in \mathcal{M}$. By ODE (3.7), we have

$$\frac{d}{dt} \sum_{m\in\mathcal{M}} v_m q_{m,1}(t) = -\sum_{m\in\mathcal{M}} v_m u_m q_{m,1}(t) + \lambda\zeta.$$

Because $\mathbf{q}_0 \le \mathbf{q}^*$, then $\overline{\mathbf{q}}(t, \mathbf{q}_0) \le \overline{\mathbf{q}}(t, \mathbf{q}^*) = \mathbf{q}^*$. Observe that $\sum_{m\in\mathcal{M}} v_m u_m q^*_{m,1} = \lambda\zeta$. Hence, if for some $m \in \mathcal{M}$, $q_{m,1}(t) < q^*_{m,1}$, then $\frac{d\sum_{m\in\mathcal{M}} v_m q_{m,1}(t)}{dt} > 0$, which implies that

$$\lim_{t\to\infty} \sum_{m\in\mathcal{M}} v_m q_{m,1}(t) = \sum_{m\in\mathcal{M}} v_m q_{m,1}(\infty) = \lambda\zeta.$$

Because for all $m \in \mathcal{M}$ and $t \ge 0$, $q_{m,1}(t) \le q^*_{m,1}$, then $\lim_{t\to\infty} q_{m,1}(t) = q^*_{m,1}$ must hold for all $m \in \mathcal{M}$. □

Proof of Theorem 7. The result holds immediately from Proposition 4 and Theorem 4. □

5.2. Proof of Tightness and Interchange of Limits

Next, we are going to prove the tightness of the steady-state occupancy processes $\{\mathbf{q}^N(\infty)\}_N$. Let $\overline{q}^N_l(\infty) = \sum_{m\in\mathcal{M}} q^N_{m,l}(\infty)$ and $\overline{\mathbf{q}}^N(\infty) = (\overline{q}^N_l(\infty), l \in \mathbb{N}_0)$. In order to show the tightness of $\{\mathbf{q}^N(\infty)\}_N$, it is sufficient to show that the sequence is $\{\overline{\mathbf{q}}^N(\infty)\}_N$, which is stated in the next proposition. For showing the tightness, we will bound the tail of the expected global occupancy of the stationary state first.

Lemma 11. *Let $\{G^N\}_N$ be a sequence of proportionally sparse graphs satisfying Condition 1. There exists an N_0 such that for all $N \ge N_0$ and $\ell \ge 1$,*

$$\sum_{l=\ell}^{\infty} \mathbb{E}(\overline{q}^N_l(\infty)) \le \frac{(1+\rho)/2}{1-(1+\rho)/2} \mathbb{E}(\overline{q}^N_{\ell-1}(\infty)). \tag{5.5}$$

Furthermore,

$$\mathbb{E}(\overline{q}^N_\ell(\infty)) \le \left(\frac{1+\rho}{2}\right)^\ell, \quad \forall \ell \in \mathbb{N}_0. \tag{5.6}$$

The proof of Lemma 11 is similar to Rutten and Mukherjee (2022, lemma 3). We define a sequence $\{L^N_{m,\ell}\}_{m\in\mathcal{M},\ell\in\mathbb{N}_0}$ of Lyapunov functions and bound the drift of $L^N_{m,\ell}$, which enables us to bound the tail sum of $\overline{q}^N_l(\infty)$ starting from ℓ. Given the Nth system state, $X^N = (X^N_j, j \in V^N)$. Let $Q^N_{m,l}(X)$ be the set of servers of type $m \in \mathcal{M}$ with queue length at least $l \in \mathbb{N}_0$. For each $m \in \mathcal{M}$, we define a sequence of Lyapunov functions $L^N_{m,\ell}(X) = \sum_{i=\ell}^{\infty}\sum_{l=i}|Q^N_{m,l}(X)|$, $\ell \in \mathbb{N}_0$. The complete proof is provided in Appendix H.

The next lemma from Mukherjee et al. (2018a) gives us the criterion for ℓ_1-tightness.

Lemma 12 (Mukherjee et al. 2018b, lemma 2). *Let $\{\mathbf{X}^N\}$ be a sequence of random variables in $\mathcal{S}'$, where $\mathcal{S}' = \{\mathbf{x} \in [0,1]^{\mathbb{N}_0} : x_i \le x_{i-1},\ \forall i \in \mathbb{N}_0,$ and $\sum_i x_i < \infty\}$. Then, the following are equivalent.*

i. *$\{\mathbf{X}^N\}$ is tight with respect to the product topology, and for all $\varepsilon > 0$,*

$$\lim_{k\to\infty}\overline{\lim_{N\to\infty}}\ \mathbb{P}\left(\sum_{i\ge k} x^N_i > \varepsilon\right) = 0. \tag{5.7}$$

ii. *$\{\mathbf{X}^N\}$ is tight with respect to the ℓ_1-topology.*

Proof of Theorem 5. Because for all $l \in \mathbb{N}_0$, $\overline{q}^N_l \in [0,1]$, then it is easy to check that $\{\overline{\mathbf{q}}^N(\infty)\}$ is tight with respect to the product topology. Hence, it is sufficient to show that for any $\varepsilon > 0$,

$$\lim_{\ell\to\infty}\overline{\lim_{N\to\infty}}\ \mathbb{P}\left(\sum_{l\ge\ell}\overline{q}^N_l(\infty) > \varepsilon\right) = 0. \tag{5.8}$$

By Markov's inequality and Lemma 11, we have that for all $N \ge N_0$,

$$\mathbb{P}\left(\sum_{l\ge\ell}\overline{q}^N_l(\infty) > \varepsilon\right) \le \frac{1}{\varepsilon}\mathbb{E}\left(\sum_{l\ge\ell}\overline{q}^N_l(\infty)\right) \le \frac{1}{\varepsilon}\frac{(1+\rho)/2}{1-(1+\rho)/2}\mathbb{E}(\overline{q}^N_{\ell-1}(\infty)) \le \frac{1}{\varepsilon}\frac{((1+\rho)/2)^{\ell}}{1-(1+\rho)/2}, \tag{5.9}$$

which implies that (5.8) holds. By Lemma 12, the desired result holds. □

Proof of Theorem 6. By Theorem 5, $\{\mathbf{q}^N(\infty)\}_N$ is tight with respect to the ℓ_1-topology. Then, any subsequence has a convergent further subsequence. Let $\{\mathbf{q}^{N_n}(\infty)\}_n$ be such convergent subsequence, and assume $\mathbf{q}^N(\infty)\xrightarrow{d}\mathbf{q}^*$. Clearly, $\mathbf{q}^*$ must be in the space $\mathcal{S}$. Now, initiate the Nth system at its stationarity. Then, the system is in steady state at any fixed finite time $t \ge 0$. That is, we have $\mathbf{q}^{N_n}(t) \sim \mathbf{q}^{N_n}(\infty)$ for all $t \in [0,T]$. Also, by Theorem 3, $\mathbf{q}^{N_n}(t)\xrightarrow{d}\mathbf{q}(t)$. Thus, for all $t \in [0,T]$, $\mathbf{q}(t) \sim \mathbf{q}^*$, which implies that $\mathbf{q}^*$ is a stationary point of the limiting system. By Theorem 4, we know that $\mathbf{q}^*$ is unique. Therefore, the desired result holds. □

6. Numerical Results

In this section, we will present the simulation to validate the theoretical results. Using the insights from the theoretical results, we will also show that systems with carefully designed compatibility structure perform much better than the classical, fully flexible systems. Throughout this section, we set the system parameters as follows: $K = 2$: two clusters of dispatchers; $M = 3$: three types of servers; $d = 2$: the system follows the JSQ(2) policy; $\boldsymbol{\mu} = (1,5,10)$, where each μ_m, $m = 1,2,3$, is the service rate of type m servers; $\lambda = 3$, which is the arrival rate at each dispatcher is λ; $Q = \begin{bmatrix} 0.2 & 0.5 & 0.3 \\ 0.5 & 0 & 0.5 \\ 0.9 & 0.1 & 0 \end{bmatrix}$, where each $q_{m,l}$ is the probability that type m, $m = 1,2,3$ server's initial queue length is l, $l = 1,2,3$; fraction of types of dispatchers: $[w_1 \quad w_2] = [0.2 \quad 0.8]$; fraction of types of servers: $[v_1 \quad v_2 \quad v_3] = [0.5 \quad 0.3 \quad 0.2]$; and $\zeta = 1$: the relationship between the number of dispatchers and that of servers in the system.

In the setting, the capacity sufficiency is satisfied, $\lambda\zeta = 3 < \sum_{m\in\mathcal{M}} v_m u_m = 4$. The first experiment is to compare the performance of the classical, fully flexible system with that of the system with carefully designed compatibility structure.

6.1. Complete Bipartite vs. Designed Compatibility Structure

The complete bipartite is the case that the compatibility matrix $\mathbf{p}^0 = (p^0_{m,k}, m \in \mathcal{M}, k \in \mathcal{K})$ is a matrix with all elements equal to one. From Lemma 1, we have that an Nth system under JSQ(d) is stable if and only if it satisfies the

following:

$$\rho^N = \max_{\substack{U\subseteq V^N\\ U\neq\emptyset}} \left\{ \left(\sum_{j\in U}\sum_{m\in\mathcal{M}} \mathbb{1}_{(j\in V_m^N)} u_m \right)^{-1} \sum_{i\in W^N} \sum_{S\subseteq (U\cap \mathcal{N}_w^N(i)):\, |S|=d} \frac{\lambda}{\binom{\delta_i^N}{d}} \right\} < 1.$$

By Lemma 10, for the complete bipartite case, we have that

$$\lim_{N\to\infty} \rho^N \geq \max_{\mathcal{M}'\subseteq\mathcal{M}} \left(\sum_{m\in\mathcal{M}'} v_m u_m \right)^{-1} \lambda\zeta \sum_{k\in\mathcal{K}} w_k \left(\sum_{m\in\mathcal{M}'} v_m \right)^d \geq (0.5\times 1)^{-1} \times 3 \times (0.5)^2 > 1,$$

which implies that for large-enough N, the system under JSQ(2) is unstable. The bottleneck here is that the type 1 servers with poor performance receive heavy workload. By Proposition 1, if the capacity sufficiency is satisfied, then there always exists a compatibility matrix $\mathbf{p}^1 \in [0,1]^{K\times M}$ making all large-enough systems stable under JSQ(2). Checking the feasible region defined in Lemma 2, we get one of the appropriate matrices $\mathbf{p}^1$ defined as $\mathbf{p}^1 = \begin{bmatrix} 0.05 & 0.6 & 1 \\ 0.1 & 0.7 & 1 \end{bmatrix}$. The intuition for designing the compatibility matrix, like $\mathbf{p}^1$, is to lower the traffic intensity for type 1 servers by decreasing the fraction of the type 1 servers in the neighborhood of each dispatcher. For the experiment, we set the number of servers $n = 1{,}000$ and consider two systems $S1$ and $S2$. $S1$ is a system with complete bipartite graph structure; $S2$, generated by $\text{IRG}(\mathbf{p}^1)$ (Definition 3), is a system with compatibility matrix $\mathbf{p}^1$. We simulate the evolution of each system 100 times and plot the mean sample path in Figure 1.

Figure 1 shows that the average queue length of type 1 servers in $S1$ almost monotonically increases as t increases, which implies that the average queue length of type 1 servers in $S1$ is unbounded. However, in the system $S2$, the average queue length of each type of servers is bounded. From this numerical result, we observe that with an appropriately designed graph structure, the performance of the system can be improved. Although we tried to plot the 95% confidence interval (CI) for each point $t = 0.5, 1.0, 1.5, 2.0, 2.5$, the CI is narrow, and its size is smaller than that of markers in the plot. One reasonable explanation for such a narrow CI is that for large-enough N, the scaled occupancy process $\mathbf{q}^N$ is close to the fluid limit $\mathbf{q}$. In other words, the error of the mean-field approximation is quite small, which can be of independent interest. With a similar heterogeneous setting, Allmeier and Gast (2022) show that the error of the mean-field approximation is $O(1/N)$.

6.2. Convergence of Global Occupancy States

In this experiment, we generate systems by $\text{IRG}(\mathbf{p}^1)$ and simulate the evolution of systems with size $n = 100$, 500, 1,000. For each system, we also simulate 100 times and plot the mean trajectories of $q_{m,1}^N$ and $q_{m,2}^N$, $m \in \{1,2,3\}$ in Figure 2. Also, we plot the evolution of $q_{m,1}$ and $q_{m,2}$, $m = 1, 2, 3$, of the limit system. The simulation results show that the evolution of the global occupancy of the Nth system converges to that of the limit system as N goes to infinity. From the simulation result, we find that $q_{1,1}^N$ and especially, $q_{1,2}^N$ decrease very fast when their initial values are

Figure 1. Complete Bipartite vs. Appropriate Designed Structure

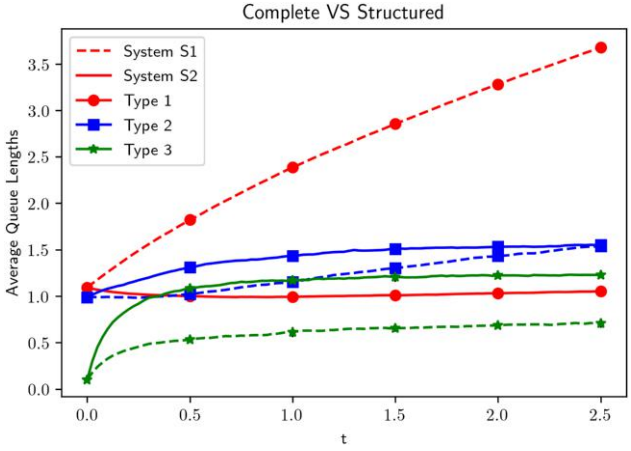
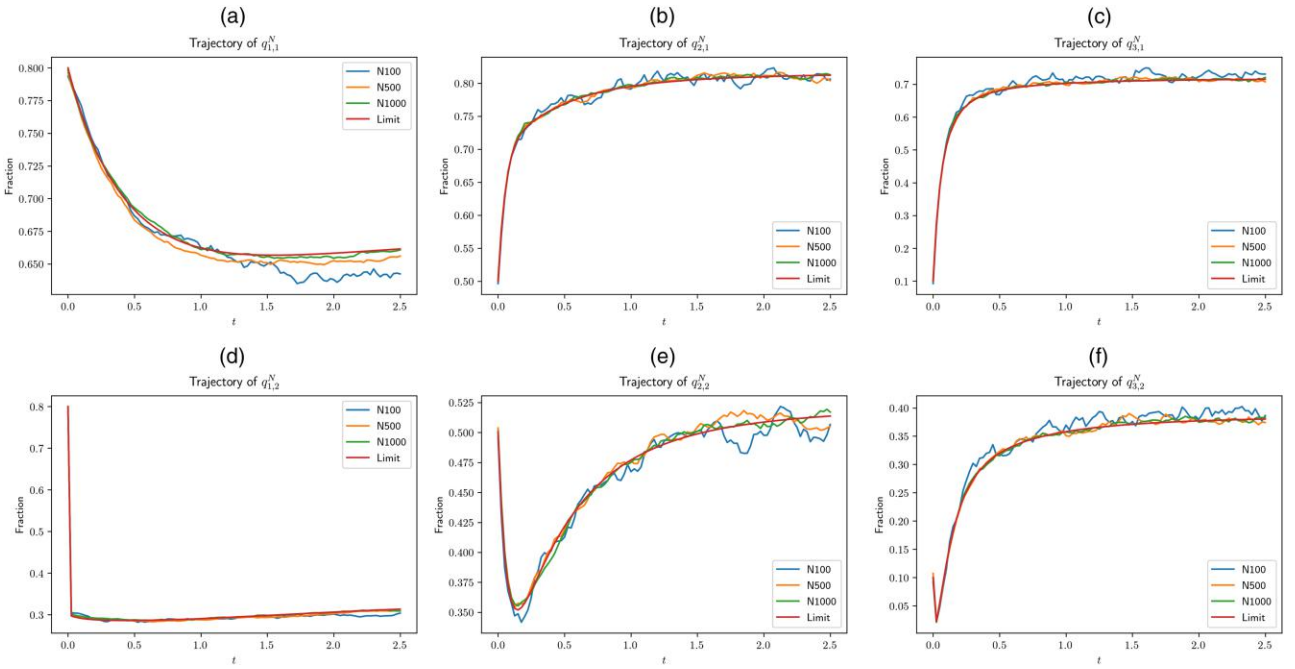


Figure 2. The Simulated Trajectories of $q^N_{m,1}$ and $q^N_{m,2}$, $m = 1, 2, 3$ Converging to the Solution of the System of ODEs as N Increases



large. In other words, when the average queue length of type 1 servers is large, it will decrease very fast. The reason is because of our designed compatibility matrix such that compared with other type servers, type 1 servers are sampled much less often.

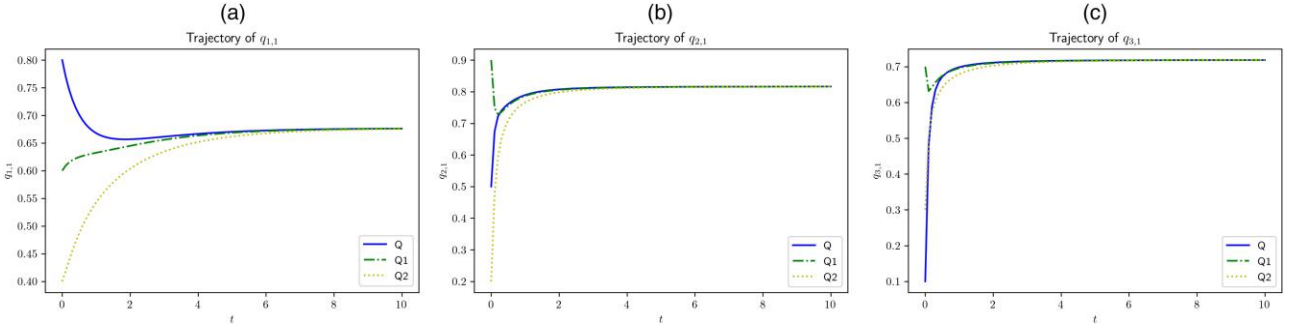
6.3. Uniqueness of the Fixed Point of the Limit System

From Theorem 4, we have that for all $\mathbf{q} \in \mathcal{S}$, $\lim_{\to\infty} \overline{\mathbf{q}}(t, \mathbf{q}_0) = \mathbf{q}^*$. In order to verify this, we use a simulation of the evolution of $\overline{\mathbf{q}}(t, \mathbf{q}_0)$ with different $\mathbf{q}_0 \in \mathcal{S}$ (i.e., consider the different Q mentioned). We also simulate the system with $Q_1 = \begin{bmatrix} 0.4 & 0.3 & 0.3 \\ 0.1 & 0.8 & 0.1 \\ 0.3 & 0.6 & 0.1 \end{bmatrix}$ and $Q_2 = \begin{bmatrix} 0.6 & 0.3 & 0.1 \\ 0.8 & 0.1 & 0.1 \\ 0.7 & 0.2 & 0.1 \end{bmatrix}$. Figure 3 shows that with different $\mathbf{q} \in \mathcal{S}$, $\lim_{t\to\infty} q_{m,1}(t)$, $m = 1, 2, 3$, are the same. If $q_{m,1}$, $m = 1, 2, 3$ are fixed, then the values of all $q_{m,l}$, $l \geq 2$, $m = 1, 2, 3$ are fixed as well by using (5.1). Hence, Figure 3 verifies the uniqueness of the fixed point.

7. Conclusion

In this paper, we model a heterogeneous system as a bipartite graph and investigate how we can impose the data locality to significantly improve the system performance even if the individual task assignment remains oblivious to the service rates. We figure out that if the sequence of systems satisfies the capacity sufficiency, we can always

Figure 3. Multiple Trajectories of $q_{m,1}$, $m = 1, 2, 3$ in the Limit System Converging to the Fixed Point



design an appropriate graph structure between dispatchers and servers such that the vanilla JSQ(d) policy achieves maximal throughput and the tail of queue-length distribution decreases doubly exponentially. However, it is worthwhile to mention that although we consider the data locality, which restricts the compatible servers for each dispatcher, our work is not really to investigate the sparsest system. It is an interesting future research direction to see what is the sparsest in which such a compatibility graph can achieve similar favorable properties, like the double-exponential tail decay.

Appendix A. Proofs for Stability Results

The goal of this appendix is to prove Proposition 1. We start by proving Lemma 2, for which we need the next technical lemma. This lemma will help us to upper bound the probability that a new task will be assigned to a specific subset of servers (in particular, (A.5)).

Lemma A.1. *Consider the following optimization problem*:

$$\max \sum_{i=1}^{N} \binom{x_i}{d} \quad s.t. \ \sum_{i=1}^{N} x_i = C \ and \ x_i \in [0, D],$$

where C and D are positive integers. Let $k^* = \lfloor C/D \rfloor$. Then, the optimal value is $k^* \binom{D}{d} + \binom{C - Dk^*}{d}$, if $N > k^*$; otherwise, the optimal value is $N \binom{D}{d}$.

Proof. We will prove by contradictions. Suppose the maximizer $\{x_i^* : i = 1, \dots, N\}$ contains some $x_j^*, x_k^* \in \{1, \dots, D-1\}$ for some $j \neq k$. Note that

$$\binom{x_j^*}{d} + \binom{x_k^*}{d} < \binom{\tilde{x}_j}{d} + \binom{\tilde{x}_k}{d},$$

where $\tilde{x}_j = \min\{x_j^* + x_k^*, D\}$ and $\tilde{x}_k = x_j^* + x_k^* - \tilde{x}_j$; that is, the pair (x_j^*, x_k^*) gives a smaller value than the extremer pair $(\tilde{x}_j, \tilde{x}_k)$. This contradicts the assumption that $\{x_i^* : i = 1, \dots, N\}$ is the maximizer. Therefore the maximizer $\{x_i^* : i = 1, \dots, N\}$ must contain at most one $x_j^* \in \{1, \dots, D-1\}$, with all the other x_i^* being either zero or D. This completes the proof. □

Proof of Lemma 2. Suppose that (3.1) holds. Because $\mathcal{M}$ is finite, then there exists a $\rho \in (0,1)$ such that $\frac{\lambda\zeta}{u_m} \sum_{k \in \mathcal{K}} \frac{w_k p_{k,m}}{\delta_k} < \rho$ for all $m \in \mathcal{M}$. Fix any $\varepsilon \in \left(0, \frac{1-\rho}{1+3\rho}\right)$. Recall $\delta_i^N = |\mathcal{N}_w^N(i)|$. By our model assumption and Condition 1, there exists $N_\varepsilon \in \mathbb{N}_0$ such that for all $m \in \mathcal{M}$ and $j \in V_m^N$,

$$p_{k,m} w_k W(N)(1-\varepsilon) \le \deg_v^N(k,j) \le p_{k,m} w_k W(N)(1+\varepsilon), \quad \forall k \in \mathcal{K}, \tag{A.1}$$

and for all $K \in \mathcal{K}$ and $i \in W_k^N$,

$$N\delta_k(1-\varepsilon) \le \delta_i^N \le N\delta_k(1+\varepsilon). \tag{A.2}$$

Consider the Nth system. Consider any nonempty subset $U \subseteq V$ of servers. If $|U| \le C(\lambda, \rho) := \frac{\lambda}{\rho \min_{m \in \mathcal{M}} u_m}$, then there exists an $N_1 \in \mathbb{N}_0$ such that for all $N \ge (N_\varepsilon \vee N_1)$,

$$\left(\sum_{j \in U} \sum_{m \in \mathcal{M}} \mathbb{1}_{(j \in V_m^N)} u_m \right)^{-1} \sum_{i \in W^N} \sum_{\substack{S \subseteq (U \cap \mathcal{N}_w^N(i)) : \\ |S| = d}} \frac{\lambda}{\binom{\delta_i^N}{d}} \le \frac{1}{|U| \min_{m \in \mathcal{M}} u_m} \sum_{k \in \mathcal{K}} \sum_{i \in W_k^N} \frac{\lambda \binom{|C(\lambda,\rho)|}{d}}{\binom{\delta_i^N}{d}} \le \rho,$$

and for all $i \in W^N$, δ_i^N goes to infinity as $N \to \infty$ uniformly by (A.2). Next, consider the case $|U| > C(\lambda, \rho)$. Denote $\alpha_m = |U \cap V_m^N| / |V_m^N|$ for each $m \in \mathcal{M}$. Then,

$$\left(\sum_{j \in U} \sum_{m \in \mathcal{M}} \mathbb{1}_{(j \in V_m^N)} u_m \right)^{-1} \sum_{i \in W^N} \sum_{\substack{S \subseteq (U \cap \mathcal{N}_w^N(i)) : \\ |S| = d}} \frac{\lambda}{\binom{\delta_w^N(i)}{d}} \le \left(\sum_{m \in \mathcal{M}} \lfloor |V_m^N| \alpha_m \rfloor u_m \right)^{-1} \sum_{k \in \mathcal{K}} \sum_{i \in W_k^N} \frac{\lambda \binom{|U \cap \mathcal{N}_w^N(i)|}{d}}{\binom{\delta_w^N(i)}{d}}. \tag{A.3}$$

By (A.1), we have that for each $k \in \mathcal{K}$,

$$\sum_{i \in W_k^N} |U \cap \mathcal{N}_w^N(i)| = \sum_{j \in U} \deg_v^N(k,j) \le \sum_{m \in \mathcal{M}} |V_m^N| \alpha_m p_{k,m} w_k W(N)(1+\varepsilon). \tag{A.4}$$

By Lemma A.1, (A.2), and (A.4),

$$\text{(A.3)} \le \left(\sum_{m\in\mathcal{M}} \lfloor |V_m^N|\alpha_m \rfloor u_m\right)^{-1} \lambda \sum_{k\in\mathcal{K}} \left(\left\lfloor \frac{\sum_{m\in\mathcal{M}} |V_m^N|\alpha_m p_{k,m} w_k W(N)(1+\varepsilon)}{\delta_k N(1-\varepsilon)} \right\rfloor + 1\right) \frac{\binom{\delta_k N(1+\varepsilon)}{d}}{\binom{\delta_k N(1-\varepsilon)}{d}}$$

$$\le C_1(N)\left(\frac{1+\varepsilon}{1-\varepsilon}\right)^d \left(\sum_{m\in\mathcal{M}} v_m\alpha_m u_m\right)^{-1} \lambda\zeta \sum_{k\in\mathcal{K}} \left(w_k \left\lfloor \frac{\sum_{m\in\mathcal{M}} v_m\alpha_m p_{k,m}(1+\varepsilon)}{\delta_k(1-\varepsilon)} \right\rfloor + \frac{1}{N}\right), \tag{A.5}$$

where $C_1(N)$ only depends on N and goes to one as $N \to \infty$. Let

$$\mathcal{K}' := \left\{k \in \mathcal{K} : \left\lfloor \frac{\sum_{m\in\mathcal{M}} v_m\alpha_m p_{k,m}(1+\varepsilon)}{\delta_k(1-\varepsilon)} \right\rfloor \ge 1\right\}.$$

If $\mathcal{K}' = \emptyset$, then

$$\text{(A.5)} \le C_1(N)\left(\frac{1+\varepsilon}{1-\varepsilon}\right)^d \frac{\lambda\zeta}{N\sum_{m\in\mathcal{M}} v_m\alpha_m u_m}$$

$$\le C_1(N)\left(\frac{1+\varepsilon}{1-\varepsilon}\right)^d \frac{\lambda\zeta}{C(\lambda,\rho)\min_{m\in\mathcal{M}} u_m} \le \rho\left(\frac{1+\varepsilon}{1-\varepsilon}\right)^d. \tag{A.6}$$

Consider the case $\mathcal{K}' \neq \emptyset$. Then, we get

$$\text{(A.5)} \le C_1(N)\left(\frac{1+\varepsilon}{1-\varepsilon}\right)^d \left(\sum_{m\in\mathcal{M}} v_m\alpha_m u_m\right)^{-1} \lambda\zeta \left(\sum_{k\in\mathcal{K}} w_k \frac{\sum_{m\in\mathcal{M}} v_m\alpha_m p_{k,m}(1+\varepsilon)}{\delta_k(1-\varepsilon)} + \frac{1}{N}\right)$$

$$\le C_1(N)\left(\frac{1+\varepsilon}{1-\varepsilon}\right)^d \left(\sum_{m\in\mathcal{M}} v_m\alpha_m u_m\right)^{-1} \lambda\zeta \left(\sum_{k\in\mathcal{K}} w_k \frac{\sum_{m\in\mathcal{M}} v_m\alpha_m p_{k,m}(1+\varepsilon)}{\delta_k(1-\varepsilon)}\right)$$

$$+ C_1(N)\left(\frac{1+\varepsilon}{1-\varepsilon}\right)^d \frac{\lambda\zeta}{N\sum_{m\in\mathcal{M}} v_m\alpha_m u_m}. \tag{A.7}$$

By (3.1), we have that for all $m \in \mathcal{M}$ and $\alpha_m \in (0,1)$,

$$(\alpha_m v_m u_m)^{-1} \lambda\zeta \left(\sum_{k\in\mathcal{K}} w_k \frac{\alpha_m p_{k,m} v_m}{\delta_k}\right) \frac{1+\varepsilon}{1-\varepsilon} \le \frac{\rho(1+\varepsilon)}{1-\varepsilon} < \frac{1+\rho}{2},$$

which implies that

$$\lambda\zeta \left(\sum_{k\in\mathcal{K}} w_k \frac{\sum_{m\in\mathcal{M}} v_m\alpha_m p_{k,m}(1+\varepsilon)}{\delta_k(1-\varepsilon)}\right) < \frac{1+\rho}{2}\left(\sum_{m\in\mathcal{M}} v_m\alpha_m u_m\right). \tag{A.8}$$

Because $\mathcal{K}'$ is nonempty, then we assume $k' \in \mathcal{K}$ is in $\mathcal{K}'$ (i.e., $\sum_{m\in\mathcal{M}} v_m\alpha_m p_{k',m}(1+\varepsilon) \ge \delta_{k'}(1-\varepsilon)$). Hence,

$$\frac{\lambda\zeta}{N\sum_{m\in\mathcal{M}} v_m\alpha_m u_m} \le \frac{\lambda\zeta(1+\varepsilon)}{N\delta_{k'}(1-\varepsilon)\min_{m\in\mathcal{M}} u_m} \le \frac{\lambda\zeta\rho}{N\delta_{k'}\min_{m\in\mathcal{M}} u_m}, \tag{A.9}$$

which implies that there exists $N_2 \in \mathbb{N}_0$ such that for all $N \ge N_2$,

$$\frac{\lambda\zeta}{N\sum_{m\in\mathcal{M}} v_m\alpha_m u_m} \le \frac{\lambda\zeta(1+\varepsilon)}{N\delta_{k'}(1-\varepsilon)\min_{m\in\mathcal{M}} u_m} < C_2(N) \xrightarrow{N\to\infty} 0. \tag{A.10}$$

We choose ε such that $\left(\frac{1+\varepsilon}{1-\varepsilon}\right)^d \frac{1+\rho}{2} < 1$. By (A.8) and (A.10), we have that there exists a positive integer $N_3 \ge (N_\varepsilon \vee N_1 \vee N_2)$ such that for all $N_3 \in \mathbb{N}_0$,

$$\left(\sum_{j\in U}\sum_{m\in\mathcal{M}'} \mathbb{1}_{(j\in V_m^N)} u_m\right)^{-1} \sum_{i\in W^N} \sum_{\substack{S\subseteq(U\cap\mathcal{N}_w^N(i)):\\ |S|=d}} \frac{\lambda}{\binom{\delta_w^N(i)}{d}} < \left(\frac{1+\varepsilon}{1-\varepsilon}\right)^d \left(C_1(N)\frac{1+\rho}{2} + C_2(N)\right) < 1. \tag{A.11}$$

We choose ε such that $\left(\frac{1+\varepsilon}{1-\varepsilon}\right)^d \frac{1+\rho}{2} < 1$. Now, because the subset $U \subseteq V^N$ is arbitrary, then for all $N \ge N_3$, the Nth system is stable under JSQ(d) policy. □

Proof of Proposition 1. By Lemma 2, it is sufficient to show that there exists some $\mathbf{p}$ such that for each $m \in \mathcal{M}$,

$$\lambda\zeta\sum_{k\in\mathcal{K}} w_k \frac{v_m p_{k,m}}{\delta_k} < v_m u_m. \tag{A.12}$$

Let $x_{m,k} = \frac{v_m p_{k,m}}{\delta_k} \in [0,1]$, $k \in \mathcal{K}$, $m \in \mathcal{M}$ with $\sum_{m\in\mathcal{M}} x_{m,k} = 1$. Now, we can formulate a linear optimization problem as the following; the objective is min ρ, and the constraints are

$$\begin{aligned} \lambda\zeta\sum_{k\in\mathcal{K}} w_k x_{k,m} &\le \rho v_m u_m, \quad \forall m \in \mathcal{M}, \\ \sum_{m\in\mathcal{M}} x_{k,m} &= 1, \quad \forall k \in \mathcal{K}, \\ x_{k,m} &\in [0,1], \quad \forall k \in \mathcal{K}, m \in \mathcal{M}. \end{aligned} \tag{A.13}$$

Next, we construct a specific solution $\mathbf{x}' = (x'_{k,m}, k \in \mathcal{K}, m \in \mathcal{M})$ satisfying Constraints (A.13) with $\rho_0 = \lambda\zeta/\sum_{m\in\mathcal{M}} v_m u_m$. Note that $\rho_0 < 1$ by (2.1). For convenience, we denote $x'_{k,0} = 0$ for all $k \in \mathcal{K}$. First, consider $k = 1$. Let $x'_{1,1} = \frac{\min(\rho_0 v_1 u_1, \lambda\zeta w_1)}{\lambda\zeta w_1}$, and for $m \ge 2$,

$$x'_{1,m} = \frac{\min(\rho_0 v_m u_m, \lambda\zeta w_1(1 - \sum_{m'<m} x'_{1,m'}))}{\lambda\zeta w_1}.$$

Because $\lambda\zeta w_1 \le \lambda\zeta = \rho_0 \sum_{m\in\mathcal{M}} v_m u_m$, then $\sum_{m\in\mathcal{M}} x'_{1,m} = 1$ and $m_1 := \min\{m \in \mathcal{M} : \rho_0 v_m u_m - x'_{1,m}\lambda\zeta w_1 > 0\} \in \mathcal{M}$. Then, consider $k = 2$. For all $m < m_1$, let $x'_{2,m} = 0$. Let

$$x'_{2,m_1} = \frac{\min(\rho_0 v_{m_1} u_{m_1} - x'_{1,m_1}\lambda\zeta w_1, \lambda\zeta w_2)}{\lambda\zeta w_2},$$

and let

$$x'_{2,m} = \frac{\min(\rho_0 v_m u_m, \lambda\zeta w_2(1 - \sum_{m'<m} x'_{2,m'}))}{\lambda\zeta w_2}, \quad m > m_1.$$

Again, because $\lambda\zeta(w_1 + w_2) \le \lambda\zeta \le \rho_0 \sum_{m\in\mathcal{M}} v_m u_m$, then $\sum_{m\in\mathcal{M}} x'_{2,m} = 1$ and $m_2 := \min\{m \ge m_1 : \rho_0 v_m u_m - x'_{2,m}\lambda\zeta w_2 > 0\} \in \mathcal{M}$. We can construct $x'_{k,m}$, $m \in \mathcal{M}$, $k \ge 3$ by following the steps of the construction of $x'_{2,m}$, $m \in \mathcal{M}$. Hence, we get a specific solution $\mathbf{x}'$ satisfying (A.13) with $\rho_0 < 1$. Therefore, min ρ is strictly less than one, and our desired result holds. □

Appendix B. Approximation of the Graph Structure for Large *N* Systems

Proof of Lemma 5. Consider any fixed $m \in \mathcal{M}$ and fixed $j \in V_m$. Also, fix any $k \in \mathcal{K}$ and $(M_2, \dots, M_d) \in \mathcal{M}^{d-1}$:

$$\left| \sum_{i\in W_k^N} \zeta_{i,j}^N \sum_{\substack{(j_2,\dots,j_d)\in \mathtt{set}^N(j) \\ \text{s.t. } j_2\in V_{M_2}^N,\dots,j_d\in V_{M_d}^N}} \frac{\xi_{i,j_2}^N \times \cdots \times \xi_{i,j_d}^N}{\binom{\delta_i^N}{d}(d-1)!} - d\zeta \frac{p_{k,m} w_k}{\delta_k} \prod_{h=2}^{d} \frac{v_{M_h} p_{k,M_h}}{\delta_k} \right|$$

$$\le \left| \sum_{i\in W_k^N} \xi_{i,j}^N \sum_{\substack{(j_2,\dots,j_d)\in \mathtt{set}^N(j) \\ \text{s.t. } j_2\in V_{M_2}^N,\dots,j_d\in V_{M_d}^N}} \frac{\xi_{i,j_2}^N \times \cdots \times \xi_{i,j_d}^N}{\binom{\delta_i^N}{d}(d-1)!} - \sum_{i\in W_k^N} \xi_{i,j}^N \frac{\binom{\delta_i^N}{d-1}}{\binom{\delta_i^N}{d}} \prod_{h=2}^{d} \frac{v_{M_h} p_{k,M_h}}{\delta_k} \right| \tag{B.1}$$

$$+ \left| \sum_{i\in W_k^N} \xi_{i,j}^N \frac{\binom{\delta_i^N}{d-1}}{\binom{\delta_i^N}{d}} \prod_{h=2}^{d} \frac{v_{M_h} p_{k,M_h}}{\delta_k} - d\zeta \frac{p_{k,m} w_k}{\delta_k} \prod_{h=2}^{d} \frac{v_{M_h} p_{k,M_h}}{\delta_k} \right|. \tag{B.2}$$

First,

$$\max_{i\in W_k^N}\left|\sum_{\substack{(j_2,\dots,j_d)\in \mathtt{set}^N(j)\\ \text{s.t.}\quad j_2\in V^N_{M_2},\dots,j_d\in V^N_{M_d}}}\frac{\xi^N_{i,j_2}\times\cdots\times\xi^N_{i,j_d}}{\binom{\delta_i^N}{d-1}(d-1)!}-\prod_{h=2}^{d}\frac{v_{M_h}p_{k,M_h}}{\delta_k}\right|$$

$$\le\max_{i\in W_k^N}\left|\sum_{\substack{(j_2,\dots,j_d)\in \mathtt{set}^N(j)\\ \text{s.t.}\quad j_2\in V^N_{M_2},\dots,j_d\in V^N_{M_d}}}\frac{\xi^N_{i,j_2}\times\cdots\times\xi^N_{i,j_d}}{\binom{\delta_i^N}{d-1}(d-1)!}-\frac{\deg_w^N(i,M_2)\times\cdots\times\deg_w^N(i,M_d)}{\binom{\delta_i^N}{d-1}(d-1)!}\right|$$

$$+\max_{i\in W_k^N}\left|\frac{\deg_w^N(i,M_2)\times\cdots\times\deg_w^N(i,M_d)}{\binom{\delta_i^N}{d-1}(d-1)!}-\prod_{h=2}^{d}\frac{v_{M_h}p_{k,M_h}}{\delta_k}\right|.$$

For large-enough N,

$$\max_{i\in W_k^N}\left|\sum_{\substack{(j_2,\dots,j_d)\in \mathtt{set}^N(j)\\ \text{s.t.}\quad j_2\in V^N_{M_2},\dots,j_d\in V^N_{M_d}}}\frac{\xi^N_{i,j_2}\times\cdots\times\xi^N_{i,j_d}}{\binom{\delta_i^N}{d-1}(d-1)!}-\frac{\deg_w^N(i,M_2)\times\cdots\times\deg_w^N(i,M_d)}{\binom{\delta_i^N}{d-1}(d-1)!}\right|$$

$$\le\max_{i\in W_k^N}\frac{d(d-1)}{\binom{\delta_i^N}{d-1}(d-1)!}\max_{m\in\mathcal{M}}\left(\deg_w^N(i,m)\right)^{d-2}$$

$$\le\frac{d(d-1)}{\min_{i\in W_k^N}\binom{\delta_i^N}{d-1}(d-1)!}\max_{i\in W_k^N}\max_{m\in\mathcal{M}}\left(\deg_w^N(i,m)\right)^{d-2}$$

$$\le c^N(m,k)d(d-1)\frac{\left(N\max_{m\in\mathcal{M}}v_mp_{k,m}\right)^{d-2}}{(N\delta_k)^{d-1}}\xrightarrow{N\to\infty}0, \tag{B.3}$$

where $c^N(m,k)$ goes to one as N goes to infinity and only depends on k and m for each N. The last inequality comes from Condition 1, Lemma 3, and $\frac{\delta_i^N\times\cdots\times(\delta_i^N-d+2)}{(\delta_i^N)^{d-1}}\xrightarrow{N\to\infty}1$. Similarly, we have

$$\max_{i\in W_k^N}\left|\frac{\deg_w^N(i,M_2)\times\cdots\times\deg_w^N(i,M_d)}{\binom{\delta_i^N}{d-1}(d-1)!}-\prod_{h=2}^{d}\frac{v_{M_h}p_{k,M_h}}{\delta_k}\right|$$

$$\le\max\left(\prod_{h=2}^{d}\left(\frac{\max_{i\in W_k^N}\deg_w^N(i,M_h)}{\min_{i\in W_k^N}(\delta_i^N-d)}-\frac{v_{M_h}p_{k,M_h}}{\delta_k}\right),\prod_{h=2}^{d}\left(\frac{\min_{i\in W_k^N}\deg_w^N(i,M_h)}{\max_{i\in W_k^N}\delta_i^N}-\frac{v_{M_h}p_{k,M_h}}{\delta_k}\right)\right)$$

$$\le c^N(m,k,M_2,\dots,M_d)\xrightarrow{N\to\infty}0, \tag{B.4}$$

where $c^N(m,k,M_2,\dots,M_d)$ depends on $m,k,M_2,\dots,M_d$. By (B.3) and (B.4), we have

$$\max_{i\in W_k^N}\left|\sum_{\substack{(j_2,\dots,j_d)\in \mathtt{set}^N(j)\\ \text{s.t.}\quad j_2\in V^N_{M_2},\dots,j_d\in V^N_{M_d}}}\frac{\xi^N_{i,j_2}\times\cdots\times\xi^N_{i,j_d}}{\binom{\delta_i^N}{d-1}(d-1)!}-\prod_{h=2}^{d}\frac{v_{M_h}p_{k,M_h}}{\delta_k}\right|\le c_1^N(m,k,M_2,\dots,M_d)\xrightarrow{N\to\infty}0, \tag{B.5}$$

where $c_1^N(m,k,M_2,\dots,M_d)$ depends on $m,k,M_2,\dots,M_d$. By Lemma 3, we have

$$\lim_{N\to\infty}\max_{i\in W_k^N}\frac{N\binom{\delta_i^N}{d-1}}{\binom{\delta_i^N}{d}}=\lim_{N\to\infty}\min_{i\in W_k^N}\frac{N\binom{\delta_i^N}{d-1}}{\binom{\delta_i^N}{d}}=\frac{d}{\delta_k},$$

$$\lim_{N\to\infty}\max_{j\in V_m^N}\frac{\deg_v^N(k,j)}{N}=\lim_{N\to\infty}\min_{j\in V_m^N}\frac{\deg_v^N(k,j)}{N}=\zeta p_{k,m}w_k.$$

Then,

$$\left|\sum_{i\in W_k^N}\xi_{i,j}^N\frac{\binom{\delta_i^N}{d-1}}{\binom{\delta_i^N}{d}}-d\zeta\frac{p_{k,m}w_k}{\delta_k}\right|\le\left|\sum_{i\in W_k^N}\xi_{i,j}^N\frac{\binom{\delta_i^N}{d-1}}{\binom{\delta_i^N}{d}}-\deg_v^N(k,j)\frac{d}{N\delta_k}\right|+\left|\deg_v^N(k,j)\frac{d}{N\delta_k}-d\zeta\frac{p_{k,m}w_k}{\delta_k}\right|$$

$$\le c_1^N(m,k)\xrightarrow{N\to\infty}0, \tag{B.6}$$

where $c_1^N(m,k)$ only depends on m and k.

Consider (B.1).

$$\left|\sum_{i\in W_k^N}\xi_{i,j}^N\sum_{\substack{(j_2,\dots,j_d)\in \mathrm{set}^N(j)\\ \text{s.t. } j_2\in V_{M_2}^N,\dots,j_d\in V_{M_d}^N}}\frac{\xi_{i,j_2}^N\times\cdots\times\xi_{i,j_d}^N}{\binom{\delta_i^N}{d}(d-1)!}-\sum_{i\in W_k^N}\xi_{i,j}^N\frac{\binom{\delta_i^N}{d-1}}{\binom{\delta_i^N}{d}}\prod_{h=2}^d\frac{v_{M_h}p_{k,M_h}}{\delta_k}\right|$$

$$=\left|\sum_{i\in W_k^N}\xi_{i,j}^N\frac{\binom{\delta_i^N}{d-1}}{\binom{\delta_i^N}{d}}\sum_{\substack{(j_2,\dots,j_d)\in \mathrm{set}^N(j)\\ \text{s.t. } j_2\in V_{M_2}^N,\dots,j_d\in V_{M_d}^N}}\frac{\xi_{i,j_2}^N\times\cdots\times\xi_{i,j_d}^N}{\binom{\delta_i^N}{d-1}(d-1)!}-\sum_{i\in W_k^N}\xi_{i,j}^N\frac{\binom{\delta_i^N}{d-1}}{\binom{\delta_i^N}{d}}\prod_{h=2}^d\frac{v_{M_h}p_{k,M_h}}{\delta_k}\right|$$

$$\le\sum_{i\in W_k^N}\xi_{i,j}^N\frac{\binom{\delta_i^N}{d-1}}{\binom{\delta_i^N}{d}}\left|\sum_{\substack{(j_2,\dots,j_d)\in \mathrm{set}^N(j)\\ \text{s.t. } j_2\in V_{M_2}^N,\dots,j_d\in V_{M_d}^N}}\frac{\xi_{i,j_2}^N\times\cdots\times\xi_{i,j_d}^N}{\binom{\delta_i^N}{d-1}(d-1)!}-\prod_{h=2}^d\frac{v_{M_h}p_{k,M_h}}{\delta_k}\right|$$

$$\overset{(a)}{\le}\sum_{i\in W_k^N}\xi_{i,j}^N\frac{\binom{\delta_i^N}{d-1}}{\binom{\delta_i^N}{d}}c_2^N(m,k,M_2,\dots,M_d)$$

$$\overset{(b)}{\le}c_2^N(m,k,M_2,\dots,M_d)c_2^N(m,k)d\zeta\frac{p_{k,m}w_k}{\delta_k}\xrightarrow{N\to\infty}0, \tag{B.7}$$

where $c_2^N(m,k,M_2,\dots,M_d)\xrightarrow{N\to\infty}0$ and $c_2^N(m,k)\xrightarrow{N\to\infty}1$. (a) is from (B.5), and (b) is from (B.6). Hence, (B.1) goes to zero as $N\to\infty$. Then,

$$\left|\sum_{i\in W_k^N}\xi_{i,j}^N\sum_{\substack{(j_2,\dots,j_d)\in \mathrm{set}^N(j)\\ \text{s.t. } j_2\in V_{M_2}^N,\dots,j_d\in V_{M_d}^N}}\frac{\xi_{i,j_2}^N\times\cdots\times\xi_{i,j_d}^N}{\binom{\delta_i^N}{d}(d-1)!}-d\zeta\frac{p_{k,m}w_k}{\delta_k}\prod_{h=2}^d\frac{v_{M_h}p_{k,M_h}}{\delta_k}\right|$$

$$\le c_3^N(m,k,M_2,\dots,M_d)\xrightarrow{N\to\infty}0, \tag{B.8}$$

where $c_3^N(m,k,M_2,\dots,M_d)$ only depends on $m,k,M_2,\dots,M_d$. Because $k\in\mathcal{K}$ and $(M_2,\dots,M_d)\in\mathcal{M}^{d-1}$ are arbitrary and because $\mathcal{K}$ and $\mathcal{M}^{d-1}$ are finite sets, we have

$$\max_{k\in\mathcal{K}}\max_{(M_2,\dots,M_d)\in\mathcal{M}^{d-1}}\left|\sum_{i\in W_k^N}\xi_{i,j}^N\sum_{\substack{(j_2,\dots,j_d)\in \mathtt{sett}^N(j)\\ \text{s.t.}\quad j_2\in V_{M_2}^N,\dots,j_d\in V_{M_d}^N}}\frac{\xi_{i,j_2}^N\times\cdots\times\xi_{i,j_d}^N}{\binom{\delta_i^N}{d}(d-1)!}-d\zeta\frac{p_{k,m}w_k}{\delta_k}\prod_{h=2}^d\frac{v_{M_h}p_{k,M_h}}{\delta_k}\right|$$
$$\le c^N(m)\xrightarrow{N\to\infty}0, \tag{B.9}$$

where $c^N(m)$ only depends on m. Because $c^N(m)$ does not depend on $j\in V_m^N$, (4.9) holds. □

Proof of Lemma 6. Fix any $m\in\mathcal{M}$ and $j\in V_m$. Consider (4.10). When $\xi_{i,j}^N=1$, by the definition (4.2) of $\mathtt{sett}^N(\cdot)$,

$$\sum_{\mathtt{sett}^N(j)}\frac{\xi_{i,j}^N\times\xi_{i,j_2}^N\times\cdots\times\xi_{i,j_d}^N}{\binom{\delta_i^N}{d}(d-1)!}\frac{\xi_{i,j}^N\times\xi_{i,j_2'}^N\times\cdots\times\xi_{i,j_d'}^N}{\binom{\delta_i^N}{d}(d-1)!}=\frac{\left[(d-1)!\binom{\delta_i^N-1}{d-1}\right]^2-(2d-2)!\binom{\delta_i^N-1}{2d-2}}{\binom{\delta_i^N}{d}^2((d-1)!)^2}.$$

Also, by Lemma 3, we have that for all $k\in\mathcal{K}$ and $i\in W_k$,

$$\frac{\left[(d-1)!\binom{\delta_i^N-1}{d}\right]^2-(2d-2)!\binom{\delta_i^N-1}{2d-2}}{\binom{\delta_i^N}{d}^2((d-1)!)^2}\le\frac{\left[(d-1)!\max_{i\in W_k^N}\binom{\delta_i^N-1}{d}\right]^2-(2d-2)!\min_{i\in W_k^N}\binom{\delta_i^N-1}{2d-2}}{\min_{i\in W_k^N}\binom{\delta_i^N}{d}^2((d-1)!)^2}$$
$$\le c_1(N)\frac{\left[(d-1)!\binom{N\delta_k}{d-1}\right]^2-(2d-2)!\binom{N\delta_k}{2d-2}}{\binom{N\delta_k}{d}^2((d-1)!)^2},$$

where $c_1(N)$ only depends on N and goes to one as $N\to\infty$. By Lemma 3, we have that for all $k\in\mathcal{K}$, $\max_{j\in W_m^N}\deg_v^N(k,j)\le c_2(N,m)|W_k^N|p_{k,m}$, where $c_2(N,m)$ only depends on N and m and goes to one as $N\to\infty$. Hence,

$$\sum_{i\in W^N}\sum_{\mathtt{sett}^N(j)}\frac{\xi_{i,j}^N\times\xi_{i,j_2}^N\times\cdots\times\xi_{i,j_d}^N}{\binom{\delta_i^N}{d}(d-1)!}\frac{\xi_{i,j}^N\times\xi_{i,j_2'}^N\times\cdots\times\xi_{i,j_d'}^N}{\binom{\delta_i^N}{d}(d-1)!}$$
$$=\sum_{k\in\mathcal{K}}\sum_{i\in W_k^N}\frac{\left[(d-1)!\binom{\delta_i^N-1}{d}\right]^2-(2d-2)!\binom{\delta_i^N-1}{2d-2}}{\binom{\delta_i^N}{d}^2((d-1)!)^2}$$
$$\le c_1(N)\sum_{k\in\mathcal{K}}\deg_v^N(k,j)\frac{\left[(d-1)!\binom{N\delta_k}{d-1}\right]^2-(2d-2)!\binom{N\delta_k}{2d-2}}{\binom{N\delta_k}{d}^2((d-1)!)^2}$$
$$\le c_1(N)c_2(N,m)\sum_{k\in\mathcal{K}}|W_k^N|p_{k,m}\frac{\left[(d-1)!\binom{N\delta_k}{d-1}\right]^2-(2d-2)!\binom{N\delta_k}{2d-2}}{\binom{N\delta_k}{d}^2((d-1)!)^2}.$$

Let $c_3(N) = \max_{m\in\mathcal{M}} c_1(N)c_2(N,m)$ with $c_3(N) \overset{N\to\infty}{\to} 1$. Then, we have that for large-enough N,

$$\sum_{i\in W^N}\sum_{\mathtt{sett}^N(j)} \frac{\xi^N_{i,j}\times\xi^N_{i,j_2}\times\cdots\times\xi^N_{i,j_d}}{\binom{\delta^N_i}{d}(d-1)!}\frac{\xi^N_{i,j}\times\xi^N_{i,j'_2}\times\cdots\times\xi^N_{i,j'_d}}{\binom{\delta^N_i}{d}(d-1)!}$$

$$\le c_3(N)\sum_{k\in\mathcal{K}}|W^N_k|p_{k,m}\frac{\left[(d-1)!\binom{N\delta_k}{d-1}\right]^2-(2d-2)!\binom{N\delta_k}{2d-2}}{\binom{N\delta_k}{d}^2((d-1)!)^2}$$

$$\le 2\sum_{k\in\mathcal{K}}|W^N_k|p_{k,m}\frac{\left[(d-1)!\binom{N\delta_k}{d-1}\right]^2-(2d-2)!\binom{N\delta_k}{2d-2}}{\binom{N\delta_k}{d}^2((d-1)!)^2}. \tag{B.10}$$

Because $\lim_{N\to\infty}|W^N_k|/N = \zeta w_k$ and $\left[(d-1)!\binom{x}{d-1}\right]^2-(2d-2)!\binom{x}{2d-2} \le C_3x^{2d-3}$ for some constant C_3, then by choosing C_1 appropriately, (4.10) holds for all large-enough N. We can get (4.11) in a similar way. □

Appendix C. Unique Solution of ODE (3.7)

Proof of Lemma 7. Recall that $\overline{\mathbf{q}}(t,\mathbf{q}_0)$ is a solution of (3.7) given the initial point $\mathbf{q}^N(0)=\mathbf{q}_0$. For convenience, we denote $\overline{\mathbf{q}}(t,\mathbf{q}_0)$ as $\overline{\mathbf{q}}(t)$ and write the ODE (3.7) as the following:

$$\overline{\mathbf{q}}(0)=\mathbf{q}_0,\quad \dot{\overline{\mathbf{q}}}(t)=\overline{\mathbf{h}}(\overline{\mathbf{q}}(t)), \tag{C.1}$$

where for all $m\in\mathcal{M}$,

$$\overline{h}_{m,0}(\mathbf{q})=0,$$

$$\overline{h}_{m,l}(\mathbf{q})=-u_m(q_{m,l}-q_{m,l+1})+\lambda\zeta(q_{m,l-1}-q_{m,l})\sum_{k\in\mathcal{K}}\frac{p_{k,m}w_k}{\delta_k}\frac{(\tilde{q}_{k,l-1})^d-(\tilde{q}_{k,l})^d}{\tilde{q}_{k,l-1}-\tilde{q}_{k,l}},\quad l\ge 1. \tag{C.2}$$

Observe that under (C.2), if $q_{m,l}(t)=q_{m,l+1}(t)$ for some $m\in\mathcal{M}, l\in\mathbb{N}_0, t\ge 0$, then $\overline{h}_{m,l}(\mathbf{q}(t))\ge 0$ and $\overline{h}_{m,l+1}(\mathbf{q}(t))\le 0$; if $q_{m,l}(t)=0$ for some $m\in\mathcal{M}, l\in\mathbb{N}_0, t\ge 0$, then $\overline{h}_{m,l}(\mathbf{q}(t))\ge 0$. Hence, if $\mathbf{q}\in\overline{\mathcal{S}}$, then any solution of (C.1) and (C.2) remains within $\overline{\mathcal{S}}$. In order to show the existence and the uniqueness, we use the Picard successive approximation method (Martin and Suhov 1999, theorem 1(i)). In the rest of the proof, we use the norm

$$\|\mathbf{q}\|=\sup_{m\in\mathcal{M}}\sup_{l\in\mathbb{N}_0}\frac{|q_{m,l}|}{l+1}.$$

For any $\mathbf{q},\mathbf{q}'\in\overline{\mathcal{S}}$,

$$\|\overline{\mathbf{h}}(\mathbf{q})\|\le K_1,\quad \|\overline{\mathbf{h}}(\mathbf{q})-\overline{\mathbf{h}}(\mathbf{q}')\|\le K_2\|\mathbf{q}-\mathbf{q}'\|, \tag{C.3}$$

where $K_1:=\max_{m\in\mathcal{M}}u_m+\lambda\zeta$ and $K_2:=2\max_{m\in\mathcal{M}}u_m+2d\lambda\zeta$. For $t\ge 0$, let $\mathbf{q}^{(0)}(t)=\mathbf{q}_0$, and by the Picard successive approximation method, let

$$\mathbf{q}^{(n)}(t)=\mathbf{q}_0+\int_0^t\overline{\mathbf{h}}(\mathbf{q}^{(n-1)}(s))ds,\quad n\in\mathbb{N}.$$

By induction, we have that $\mathbf{q}^{(n)}(t)$ is continuous w.r.t. t on $[0,\infty)$ for all n and that

$$\|\mathbf{q}^{(n+1)}(t)-\mathbf{q}^{(n)}(t)\|\le\frac{K_1K_2^nt^{n+1}}{(n+1)!},\quad \forall n\in\mathbb{N}, t\ge 0.$$

Hence, for all $t\ge 0$, $\mathbf{q}^{(\infty)}=\lim_{n\to\infty}\mathbf{q}^{(n)}$ exists uniformly for $s\in[0,t]$. Also, by (C.3) and the dominated convergence theorem, the following holds:

$$\mathbf{q}^{(\infty)}(t)=\mathbf{q}_0+\int_0^t\overline{\mathbf{h}}(\mathbf{q}^{(\infty)}(s))ds. \tag{C.4}$$

Next, we show the uniqueness by contradiction. Assume that $\tilde{\mathbf{q}}^{(\infty)}$ also satisfies

$$\tilde{\mathbf{q}}^{(\infty)}(t) = \mathbf{q}_0 + \int_0^t \overline{\mathbf{h}}(\tilde{\mathbf{q}}^{(\infty)}(s))ds.$$

Then, we have

$$\tilde{\mathbf{q}}^{(\infty)}(t) - \mathbf{q}^{(n)}(t) = \int_0^t [\overline{\mathbf{h}}(\tilde{\mathbf{q}}^{(\infty)}(s)) - \overline{\mathbf{h}}(\mathbf{q}^{(n-1)}(s))]ds.$$

Similarly, we get

$$\|\tilde{\mathbf{q}}^{(\infty)}(t) - \mathbf{q}^{(n)}(t)\| \le \frac{K_1 K_2^n t^{n+1}}{(n+1)!},$$

which implies that $\tilde{\mathbf{q}}^{(\infty)}(t) = \lim_{n\to\infty} \mathbf{q}^{(n)}(t) = \mathbf{q}^{(\infty)}$. □

Appendix D. Proof of Proposition 2

Lemma D.1. *If $\mathbf{q}^N(0)$ weakly converges to $\mathbf{q}(0) = \mathbf{q}^\infty \in \mathcal{S}$, then for any $\varepsilon > 0$, $\delta > 0$, and $T > 0$, there exist $\ell \in \mathbb{N}_0$ and $N_\ell \in \mathbb{N}_0$, depending on $\mathbf{q}^\infty$, ε, δ, and T, such that for all $N \ge N_1$,*

$$\mathbb{P}\left(\sup_{t\in[0,T]} \sup_{m\in\mathcal{M}} q^N_{m,\ell}(t) \ge \varepsilon\right) < \delta. \tag{D.1}$$

Proof. Fix any $\varepsilon > 0$ and $\delta > 0$. Because $\mathbf{q}^\infty \in \mathcal{S}$, then there exists $\ell_1 \in \mathbb{N}_0$ such that $\sup_{m\in\mathcal{M}} q^\infty_{m,\ell_1} \le \varepsilon/4$. By the weak convergence $\mathbf{q}^N(0) \Rightarrow \mathbf{q}^\infty$, there exists $N_1 \in \mathbb{N}_0$ such that for all $N \ge N_1$,

$$\mathbb{P}(q^N_{m,\ell_1}(0) \le \varepsilon/2) \le \mathbb{P}(\|\mathbf{q}^N(0) - \mathbf{q}^\infty\|_1 \ge \varepsilon/4) < \frac{\delta}{2}. \tag{D.2}$$

Let $\ell = \ell_1 + \sup_{m\in\mathcal{M}}\lceil \frac{4\zeta\lambda T}{v_m \varepsilon}\rceil$. Hence,

$$\begin{aligned}&\mathbb{P}\left(\sup_{t\in[0,T]} \sup_{m\in\mathcal{M}} q^N_{m,\ell}(t) \ge \varepsilon\right)\\ &\le \mathbb{P}\left(\sup_{t\in[0,T]} \sup_{m\in\mathcal{M}} q^N_{m,\ell}(t) \ge \varepsilon \,\middle|\, \sup_{m\in\mathcal{M}} q^N_{m,\ell_1}(0) < \varepsilon/2\right) + \mathbb{P}(q^N_{m,\ell_1}(0) \le \varepsilon/2).\end{aligned} \tag{D.3}$$

Because given $\sup_{m\in\mathcal{M}} q^N_{m,\ell_1}(0) < \varepsilon/2$ (i.e., for all $m \in \mathcal{M}$, $q^N_{m,\ell_1}(0)|V^N_m| < \varepsilon/2|V^N_m|$), then if for some $t \in [0,T]$ and $m \in \mathcal{M}$, $q^N_{m,\ell}(t) \ge \varepsilon$ (i.e., $q^N_{m,\ell}(t)|V^N_m| \ge \varepsilon|V^N_m|$), there must be at least $\inf_{m\in\mathcal{M}}|V^N_m|\varepsilon(\ell-\ell_1)/2$ tasks arriving in the system. By using the standard concentration inequality for Poisson random variables (Habib et al. 1998, theorem 2.3(b)), we have

$$\begin{aligned}&\mathbb{P}\left(\sup_{t\in[0,T]} \sup_{m\in\mathcal{M}} q^N_{m,\ell}(t) \ge \varepsilon \,\middle|\, \sup_{m\in\mathcal{M}} q^N_{m,\ell_1}(0) < \varepsilon/2\right) \le \mathbb{P}(\mathrm{Po}(W(N)\lambda) \ge \inf_{m\in\mathcal{M}}|V^N_m|\varepsilon(\ell-\ell_1)/2)\\ &\le \mathbb{P}(\mathrm{Po}(N\zeta\lambda T) \ge 2C(N)N\zeta\lambda T) \le \exp\left(-\frac{((2C(N)-1)N\zeta\lambda T)^2}{2(N\zeta\lambda T + ((2C(N)-1)N\zeta\lambda T)/3)}\right) \xrightarrow{N\to\infty} 0,\end{aligned} \tag{D.4}$$

where $\mathrm{Po}(\cdot)$ is a unit-rate Poisson random variable and $C(N)$ is a positive constant only dependent on N that goes to one as N goes to infinity. The second inequality comes from the assumption that $W(N)/N \to \zeta$ and $|V^N_m|/N \to v_m$, $\forall m \in \mathcal{M}$. By (D.4), there exists $N_2 \in \mathbb{N}_0$ such that for all $N \ge N_2$,

$$\mathbb{P}\left(\sup_{t\in[0,T]} \sup_{m\in\mathcal{M}} q^N_{m,\ell}(t) \ge \varepsilon \,\middle|\, \sup_{m\in\mathcal{M}} q^N_{m,\ell_1}(0) < \varepsilon/2\right) < \frac{\delta}{2}. \tag{D.5}$$

Let $N_0 = \max(N_1, N_2)$. By (D.2), (D.3), and (D.5),

$$\mathbb{P}\left(\sup_{t\in[0,T]} \sup_{m\in\mathcal{M}} q^N_{m,\ell}(t) \ge \varepsilon\right) < \delta. \quad \square$$

Lemma D.2. *For each $m \in \mathcal{M}$ and $k \in \mathcal{K}$,*

$$\sup_{U\subseteq V^N_m}\left|\frac{|E^N_k(U)|}{|E^N_k(V^N)|} - \frac{v_m p_{k,m}}{\delta_k}\frac{|U|}{|V^N_m|}\right| \to 0 \text{ as } N \to \infty. \tag{D.6}$$

Proof. Fix any $\varepsilon > 0$. By Condition 1 and Lemma 3, there exists $N(\varepsilon) \in \mathbb{N}_0$ such that for all $N \geq N(\varepsilon)$,

$$(1-\varepsilon)p_{k,m}|W_k^N||U| \leq |E_k^N(U)| \leq (1+\varepsilon)p_{k,m}|W_k^N||U|, \quad \forall U \subseteq V_m^N, \tag{D.7}$$

and

$$(1-\varepsilon)\sum_{m\in\mathcal{M}} p_{k,m}|W_k^N||V_m^N| \leq |E_k^N(V^N)| \leq (1+\varepsilon)\sum_{m\in\mathcal{M}} p_{k,m}|W_k^N||V_m^N|. \tag{D.8}$$

Hence, for all $N \geq N(\varepsilon)$,

$$\sup_{U\subseteq V_m^N}\left|\frac{|E_k^N(U)||V_m^N|}{|E_k^N(V^N)||U|} - \frac{v_m p_{k,m}}{\delta_k}\right| \leq \max\{\varepsilon_1(\varepsilon,N), \varepsilon_2(\varepsilon,N)\}, \tag{D.9}$$

where $\varepsilon_1(\varepsilon,N) = \left|\frac{(1-\varepsilon)p_{k,m}|W_k^N||V_m^N|}{(1+\varepsilon)\sum_{m\in\mathcal{M}}p_{k,m}|W_k^N||V_m^N|} - \frac{v_m p_{k,m}}{\delta_k}\right|$ and $\varepsilon_2(\varepsilon,N) = \left|\frac{(1+\varepsilon)p_{k,m}|W_k^N||V_m^N|}{(1-\varepsilon)\sum_{m\in\mathcal{M}}p_{k,m}|W_k^N||V_m^N|} - \frac{v_m p_{k,m}}{\delta_k}\right|$. Again, by Condition 1 and Lemma 3,

$$\begin{aligned}
&\lim_{N\to\infty}\sup_{U\subseteq V_m^N}\left|\frac{|E_k^N(U)||V_m^N|}{|E_k^N(V^N)||U|} - \frac{v_m p_{k,m}}{\delta_k}\right| \\
&\leq \lim_{N\to\infty}\max\{\varepsilon_1(\varepsilon,N), \varepsilon_2(\varepsilon,N)\} \\
&= \max\left\{\left|\frac{(1-\varepsilon)v_m p_{k,m}}{(1+\varepsilon)\delta_k} - \frac{v_m p_{k,m}}{\delta_k}\right|, \left|\frac{(1+\varepsilon)v_m p_{k,m}}{(1-\varepsilon)\delta_k} - \frac{v_m p_{k,m}}{\delta_k}\right|\right\}.
\end{aligned} \tag{D.10}$$

Because (D.10) holds for any $\varepsilon > 0$, we have

$$\begin{aligned}
&\lim_{N\to\infty}\sup_{U\subseteq V_m^N}\left|\frac{|E_k^N(U)||V_m^N|}{|E_k^N(V^N)||U|} - \frac{v_m p_{k,m}}{\delta_k}\right| \\
&\leq \lim_{\varepsilon\downarrow 0}\max\left\{\left|\frac{(1-\varepsilon)v_m p_{k,m}}{(1+\varepsilon)\delta_k} - \frac{v_m p_{k,m}}{\delta_k}\right|, \left|\frac{(1+\varepsilon)v_m p_{k,m}}{(1-\varepsilon)\delta_k} - \frac{v_m p_{k,m}}{\delta_k}|\right\} = 0. \quad \square
\end{aligned} \tag{D.11}$$

Proof of Proposition 2. Consider any fixed $k \in \mathcal{K}$. Also, fix $\varepsilon_1 > 0$ and $\varepsilon_2 > 0$. By the triangle inequality, we have

$$\begin{aligned}
&\mathbb{P}\left(\sup_{t\in[0,T]}\left|\left\{i \in W_k^N : \sum_{m\in\mathcal{M}}\sum_{l\in\mathbb{N}_0}|\hat{x}_{i,m,l}^N(t) - x_{k,m,l}^N(t)| > \varepsilon_1\right\}\right| \geq \varepsilon_2 M(N)/K\right) \\
&\leq \mathbb{P}\left(\sup_{t\in[0,T]}\left|\left\{i \in W_k^N : \sum_{m\in\mathcal{M}}\sum_{0\leq l\leq \ell-1}|\hat{x}_{i,m,l}^N(t) - x_{k,m,l}^N(t)| > \varepsilon_1/4\right\}\right| \geq \varepsilon_2 M(N)/(4K)\right) \\
&+ \mathbb{P}\left(\sup_{t\in[0,T]}\left|\left\{i \in W_k^N : \sum_{m\in\mathcal{M}}\sum_{l\geq\ell}\hat{x}_{i,m,l}^N(t) > \varepsilon_1/2\right\}\right| \geq \varepsilon_2 M(N)/(2K)\right) \\
&+ \mathbb{P}\left(\sup_{t\in[0,T]}\left|\left\{i \in W_k^N : \sum_{m\in\mathcal{M}}\sum_{l\geq\ell}x_{k,m,l}^N(t) > \varepsilon_1/4\right\}\right| \geq \varepsilon_2 M(N)/(4K)\right) \\
&\leq \sum_{0\leq l\leq\ell-1}\mathbb{P}\left(\sup_{t\in[0,T]}\left|\left\{i \in W_k^N : \sum_{m\in\mathcal{M}}|\hat{x}_{i,m,l}^N(t) - x_{k,m,l}^N(t)| > \varepsilon_1/(4\ell)\right\}\right| \geq \varepsilon_2 M(N)/(4\ell K)\right) \\
&+ \mathbb{P}\left(\sup_{t\in[0,T]}\left|\left\{i \in W_k^N : \sum_{m\in\mathcal{M}}\left|\sum_{l\geq\ell}(\hat{x}_{i,m,l}^N(t) - x_{k,m,l}^N(t))\right| > \varepsilon_1/4\right\}\right| \geq \varepsilon_2 M(N)/(4K)\right) \\
&+ 2\mathbb{P}\left(\sup_{t\in[0,T]}\left|\left\{i \in W_k^N : \sum_{m\in\mathcal{M}}\sum_{l\geq\ell}x_{k,m,l}^N(t) > \varepsilon_1/4\right\}\right| \geq \varepsilon_2 M(N)/(4K)\right).
\end{aligned} \tag{D.12}$$

By the triangle inequality and Markov's inequality,

$$\begin{aligned}
&\sum_{0\le l\le \ell-1} \mathbb{P}\left(\sup_{t\in[0,T]} \left| \left\{ i\in W_k^N : \sum_{m\in\mathcal{M}} |\hat{x}_{i,m,l}^N(t) - x_{k,m,l}^N(t)| > \varepsilon_1/(4\ell) \right\} \right| \ge \varepsilon_2 M(N)/(4\ell K) \right) \\
&\le \sum_{0\le l\le \ell-1} \left(\mathbb{P}\left(\sup_{t\in[0,T]} \left| \left\{ i\in W_k^N : \sum_{m\in\mathcal{M}} \left| \hat{x}_{i,m,l}^N(t) - \frac{|E_k^N(U_{m,l}^N(t))|}{|E_k^N(V^N)|} \right| > \varepsilon_1/(8\ell) \right\} \right| \ge \varepsilon_2 M(N)/(4\ell K) \right) \right. \\
&\quad \left. + \mathbb{P}\left(\sup_{t\in[0,T]} \sum_{m\in\mathcal{M}} \left| \frac{|E_k^N(U_{m,l}^N(t))|}{|E_k^N(V^N)|} - x_{k,m,l}^N \right| > \varepsilon_1/(8\ell) \right) \right) \\
&\le \frac{4\ell K}{\varepsilon_2 M(N)} \sum_{0\le l\le \ell-1} \mathbb{E}\left(\sup_{t\in[0,T]} \left| \left\{ i\in W_k^N : \sum_{m\in\mathcal{M}} \left| \hat{x}_{i,m,l}^N(t) - \frac{\left|E_k^N(U_{m,l}^N(t))\right|}{|E_k^N(V^N)|} \right| > \varepsilon_1/(8\ell) \right\} \right| \right) \\
&\quad + \sum_{0\le l\le \ell-1} \mathbb{P}\left(\sum_{m\in\mathcal{M}} \sup_{U\in V_m^N} \left| \frac{|E_k^N(U)|}{|E_k^N(V^N)|} - \frac{v_m}{\delta_k} \frac{|U|}{|V_m^N|} \right| > \varepsilon_1/(8\ell) \right) \\
&\le \frac{4\ell K}{\varepsilon_2 M(N)} \sum_{0\le l\le \ell-1} \sum_{m\in\mathcal{M}} \mathbb{E}\left(\sup_{t\in[0,T]} \left| \left\{ i\in W_k^N : \left| \hat{x}_{i,m,l}^N(t) - \frac{|E_k^N(U_{m,l}^N(t))|}{|E_k^N(V^N)|} \right| > \varepsilon_1/(8M\ell) \right\} \right| \right) \\
&\quad + \sum_{0\le l\le \ell-1} \sum_{m\in\mathcal{M}} \mathbb{P}\left(\sup_{U\in V_m^N} \left| \frac{|E_k^N(U)|}{|E_k^N(V^N)|} - \frac{v_m p_{k,m}}{\delta_k} \frac{|U|}{|V_m^N|} \right| > \varepsilon_1/(8M\ell) \right) \\
&\le \frac{4\ell K}{\varepsilon_2 M(N)} \sum_{0\le l\le \ell-1} \sum_{m\in\mathcal{M}} \sup_{U\in V_m^N} \left| \left\{ i\in W_k^N : \left| \frac{|\mathcal{N}_w^N(i)\cap U|}{|\mathcal{N}_w^N(i)|} - \frac{|E_k^N(U)|}{|W_k^N(V^N)|} \right| > \varepsilon_1/(8M\ell) \right\} \right| \\
&\quad + \sum_{0\le l\le \ell-1} \sum_{m\in\mathcal{M}} \mathbb{P}\left(\sup_{U\in V_m^N} \left| \frac{|E_k^N(U)|}{|E_k^N(V^N)|} - \frac{v_m p_{k,m}}{\delta_k} \frac{|U|}{|V_m^N|} \right| > \varepsilon_1/(8M\ell) \right) \\
&\le \frac{4\ell^2 KM}{\varepsilon_2 M(N)} \sup_{U\in V^N} \left| \left\{ i\in W_k^N : \left| \frac{|\mathcal{N}_w^N(i)\cap U|}{|\mathcal{N}_w^N(i)|} - \frac{|E_k^N(U)|}{|W_k^N(V^N)|} \right| > \varepsilon_1/(8M\ell) \right\} \right| \\
&\quad + \sum_{0\le l\le \ell-1} \sum_{m\in\mathcal{M}} \mathbb{P}\left(\sup_{U\in V_m^N} \left| \frac{|E_k^N(U)|}{|E_k^N(V^N)|} - \frac{v_m p_{k,m}}{\delta_k} \frac{|U|}{|V_m^N|} \right| > \varepsilon_1/(8M\ell) \right). \qquad \text{(D.13)}
\end{aligned}$$

By Lemma D.2, there exists $N_1 \in \mathbb{N}_0$ such that for all $N \ge N_1$,

$$\sup_{m\in\mathcal{M}} \sup_{U\subseteq V_m^N} \left| \frac{|E_k^N(U)|}{|E_k^N(V^N)|} - \frac{v_m p_{k,m}}{\delta_k} \frac{|U|}{|V_m^N|} \right| \le \varepsilon_1/(8M\ell), \qquad \text{(D.14)}$$

implying

$$\begin{aligned}
&\sum_{0\le l\le \ell-1} \mathbb{P}\left(\sup_{t\in[0,T]} \left| \left\{ i\in W_k^N : \sum_{m\in\mathcal{M}} |\hat{x}_{i,m,l}^N(t) - x_{k,m,l}^N(t)| > \varepsilon_1/(4\ell) \right\} \right| \ge \varepsilon_2 M(N)/(4\ell K) \right) \\
&\le \frac{4\ell^2 KM}{\varepsilon_2 M(N)} \sup_{U\in V^N} \left| \left\{ i\in W_k^N : \left| \frac{|\mathcal{N}_w^N(i)\cap U|}{|\mathcal{N}_w^N(i)|} - \frac{|E_k^N(U)|}{|W_k^N(V^N)|} \right| > \varepsilon_1/(8M\ell) \right\} \right|. \qquad \text{(D.15)}
\end{aligned}$$

Similarly, we have that there exists $N_2 \in \mathbb{N}_0$ such that $N \ge N_2$,

$$\begin{aligned}
&\mathbb{P}\left(\sup_{t\in[0,T]} \left| \left\{ i\in W_k^N : \sum_{m\in\mathcal{M}} \left| \sum_{l\ge\ell} (\hat{x}_{i,m,l}^N(t) - x_{k,m,l}^N(t)) \right| > \varepsilon_1/4 \right\} \right| \ge \varepsilon_2 M(N)/(4K) \right) \\
&\le \frac{4K}{\varepsilon_2 M(N)} \sup_{U\in V^N} \left| \left\{ i\in W_k^N : \left| \frac{|\mathcal{N}_w^N(i)\cap U|}{|\mathcal{N}_w^N(i)|} - \frac{|E_k^N(U)|}{|W_k^N(V^N)|} \right| > \varepsilon_1/4 \right\} \right|. \qquad \text{(D.16)}
\end{aligned}$$

By (D.12), (D.15), and (D.16), there exists $N_3 = \max(N_1, N_2)$ such that for all $N \geq N_3$,

$$\mathbb{P}\left(\sup_{t\in[0,T]}\left|\left\{i \in W_k^N : \sum_{m\in\mathcal{M}}\sum_{l\in\mathbb{N}_0}\left|\hat{x}_{i,m,l}^N(t) - x_{k,m,l}^N(t)\right| > \varepsilon_1\right\}\right| \geq \varepsilon_2 M(N)/K\right)$$

$$\leq \frac{8\ell^2 KM}{\varepsilon_2 M(N)}\sup_{U\in V^N}\left|\left\{i \in W_k^N : \left|\frac{|\mathcal{N}_w^N(i)\cap U|}{|\mathcal{N}_w^N(i)|} - \frac{|E_k^N(U)|}{|W_k^N(V^N)|}\right| > \varepsilon_1/4\right\}\right|$$

$$+ 2\mathbb{P}\left(\sup_{t\in[0,T]}\left|\left\{i \in W_k^N : \sum_{m\in\mathcal{M}}\sum_{l\geq\ell} x_{k,m,l}^N(t) > \varepsilon_1/4\right\}\right| \geq \varepsilon_2 M(N)/(4K)\right). \tag{D.17}$$

Fix any $\varepsilon_3 > 0$. By Definition 2, there exists $N_4 \in \mathbb{N}_0$ such that for all $N \geq N_4$,

$$\sup_{U\in V^N}\left|\left\{i \in W_k^N : \left|\frac{|\mathcal{N}_w^N(i)\cap U|}{|\mathcal{N}_w^N(i)|} - \frac{|E_k^N(U)|}{|W_k^N(V^N)|}\right| > \varepsilon_1/4\right\}\right| \leq \frac{\varepsilon_2 M(N)\varepsilon_3}{16\ell^2 KM}. \tag{D.18}$$

By Lemma D.1, there exists $N_5 \in \mathbb{N}_0$ such that for all $N \geq N_5$,

$$\mathbb{P}\left(\sup_{t\in[0,T]}\left|\left\{i \in W_k^N : \sum_{m\in\mathcal{M}}\sum_{l\geq\ell} x_{k,m,l}^N(t) > \varepsilon_1/4\right\}\right| \geq \varepsilon_2 M(N)/(4K)\right) \leq \frac{\varepsilon_3}{4}. \tag{D.19}$$

Hence,

$$\mathbb{P}\left(\sup_{t\in[0,T]}\left|\left\{i \in W_k^N : \sum_{m\in\mathcal{M}}\sum_{l\in\mathbb{N}_0}\left|\hat{x}_{i,m,l}^N(t) - x_{k,m,l}^N(t)\right| > \varepsilon_1\right\}\right| \geq \varepsilon_2 M(N)/K\right) \leq \varepsilon_3. \tag{D.20}$$

Because $\varepsilon_3 > 0$ are arbitrary, then the desired result holds. □

Appendix E. Bound the Mismatch

Proof of Lemma 9. Define a function $F_{m,l}^N(\cdot) : \mathcal{S} \to [0,1]$ as for $\mathbf{x} = (x_{m,l}, m \in \mathcal{M}, l \in \mathbb{N}_0) \in \mathcal{S}$,

$$F_{m,l}^N(\mathbf{x}) = \frac{\sum_{r=1}^d \sum_{r_1=1}^r \frac{r_1}{r}\binom{Nx_{m,l}}{r_1}\binom{N\sum_{\mathcal{M}\setminus\{m\}} x_{m,l}}{r-r_1}\binom{N\sum_{\mathcal{M}}\sum_{l'\geq l+1} x_{m,l}}{d-r}}{\binom{N}{d}}. \tag{E.1}$$

Also, define a function $f_{m,l}(\cdot)$ as for $\mathbf{x} \in \mathcal{S}$,

$$f_{m,l}(\mathbf{x}) = \sum_{r=1}^d \sum_{r_1=1}^r \frac{r_1}{r}\frac{d!}{r_1!(r-r_1)!(d-r)!}(x_{m,l})^{r_1}\left(\sum_{\mathcal{M}\setminus\{m\}} x_{m,l}\right)^{r-r_1}\left(\sum_{\mathcal{M}}\sum_{l'\geq l+1} x_{m,l'}\right)^{d-r}. \tag{E.2}$$

Note that for any $0 \le y \le x \le 1$ and $1 \le k \le d$, $x^k - (x-y)^k \le kxy \le ky$. Then, we have

$$\begin{aligned}
&\sum_{m\in\mathcal{M}}\sum_{l\in\mathbb{N}_0}|F^N_{m,l}(\mathbf{x})-f_{m,l}(\mathbf{x})|\\
&\le \sum_{m\in\mathcal{M}}\sum_{l\in\mathbb{N}_0}\sum_{r=1}^{d}\sum_{r_1=1}^{r}\frac{r_1}{r}\frac{d!}{r_1!(r-r_1)!(d-r)!}\left((x_{m,l})^{r_1}\left(\sum_{\mathcal{M}\setminus\{m\}}x_{m,l}\right)^{r-r_1}\left(\sum_{\mathcal{M}}\sum_{l'\ge l+1}x_{m,l'}\right)^{d-r}\right.\\
&\quad\left.-(x_{m,l}-\frac{r_1}{N})^{r_1}\left(\sum_{\mathcal{M}\setminus\{m\}}x_{m,l}-\frac{r-r_1}{N}\right)^{r-r_1}\left(\sum_{\mathcal{M}}\sum_{l'\ge l+1}x_{m,l'}-\frac{d-r}{n}\right)^{d-r}\right)\\
&\le \sum_{m\in\mathcal{M}}\sum_{l\in\mathbb{N}_0}\sum_{r=1}^{d}\sum_{r_1=1}^{r}\frac{r_1}{r}\frac{d!r_1(r-r_1)(d-r)}{r_1!(r-r_1)!(d-r)!}x_{m,l}\left(\sum_{\mathcal{M}\setminus\{m\}}x_{m,l}\right)\left(\sum_{\mathcal{M}}\sum_{l'\ge l+1}x_{m,l'}\right)\left(\frac{d}{N}\right)^{d}\\
&\le \sum_{m\in\mathcal{M}}\sum_{l\in\mathbb{N}_0}\sum_{r=1}^{d}\sum_{r_1=1}^{r}\frac{r_1}{r}\frac{d!r_1(r-r_1)(d-r)}{r_1!(r-r_1)!(d-r)!}x_{m,l}\left(\frac{d}{N}\right)^{d}\\
&= \sum_{r=1}^{d}\sum_{r_1=1}^{r}\frac{r_1}{r}\frac{d!r_1(r-r_1)(d-r)}{r_1!(r-r_1)!(d-r)!}\left(\frac{d}{N}\right)^{d}\to 0 \text{ as } 0.
\end{aligned}\tag{E.3}$$

Let $\hat{\mathbf{x}}^N_i=(\hat{x}^N_{i,m,l}, m\in\mathcal{M}, l\in\mathbb{N}_0)$ and $\mathbf{x}'^N_k=(x'^N_{k,m,l}, m\in\mathcal{M}, l\in\mathbb{N}_0)$. By (4.22) and (4.23), $p^N_{m,l}(i)=F^N_{m,l}(\hat{\mathbf{x}}^N_i)$ and $p'^N_{m,l}(k)=F^N_{m,l}(\mathbf{x}'^N_k)$. By the optimal coupling, we have

$$\begin{aligned}
\mathbb{P}(\mathit{Mismatch}) &\le \sum_{m\in\mathcal{M}}\sum_{l\in\mathbb{N}_0}|F^N_{m,l}(\hat{\mathbf{x}}^N_i)-F^N_{m,l}(\mathbf{x}'^N_k)|\\
&\le \sum_{m\in\mathcal{M}}\sum_{l\in\mathbb{N}_0}|F^N_{m,l}(\hat{\mathbf{x}}^N_i)-f_{m,l}(\hat{\mathbf{x}}^N_i)|+\sum_{m\in\mathcal{M}}\sum_{l\in\mathbb{N}_0}|F^N_{m,l}(\mathbf{x}'^N_k)-f_{m,l}(\mathbf{x}'^N_k)|\\
&\quad+\sum_{m\in\mathcal{M}}\sum_{l\in\mathbb{N}_0}|f_{m,l}(\hat{\mathbf{x}}^N_i)-f_{m,l}(\mathbf{x}'^N_k)|.
\end{aligned}\tag{E.4}$$

Next, we are going to show that $f(\cdot)$ is Lipschitz continuous for $\mathbf{x}\in\mathcal{S}$:

$$\begin{aligned}
&\sum_{m\in\mathcal{M}}\sum_{l\in\mathbb{N}_0}|f_{m,l}(\hat{\mathbf{x}}^N_i)-f_{m,l}(\mathbf{x}'^N_k)|\\
&\le \sum_{m\in\mathcal{M}}\sum_{l\in\mathbb{N}_0}\sum_{r=1}^{d}\sum_{r_1=1}^{r}\frac{r_1}{r}\frac{d!}{r_1!(r-r_1)!(d-r)!}\left|(\hat{x}^N_{i,m,l})^{r_1}\left(\sum_{\mathcal{M}\setminus\{m\}}\hat{x}^N_{i,m,l}\right)^{r-r_1}\left(\sum_{\mathcal{M}}\sum_{l'\ge l+1}\hat{x}^N_{i,m,l'}\right)^{d-r}\right.\\
&\quad\left.-(x'^N_{k,m,l})^{r_1}\left(\sum_{\mathcal{M}\setminus\{m\}}x'^N_{k,m,l}\right)^{r-r_1}\left(\sum_{\mathcal{M}}\sum_{l'\ge l+1}x'^N_{k,m,l'}\right)^{d-r}\right|\\
&\le \sum_{m\in\mathcal{M}}\sum_{l\in\mathbb{N}_0}\sum_{r=1}^{d}\sum_{r_1=1}^{r}\frac{r_1}{r}\frac{d!r_1(r-r_1)(d-r)}{r_1!(r-r_1)!(d-r)!}\left|(\hat{x}^N_{i,m,l}-x'^N_{k,m,l})\right.\\
&\quad\left.\left(\sum_{\mathcal{M}\setminus\{m\}}\hat{x}^N_{i,m,l}-\sum_{\mathcal{M}\setminus\{m\}}x'^N_{k,m,l}\right)\left(\sum_{\mathcal{M}}\sum_{l'\ge l+1}\hat{x}^N_{i,m,l'}-\sum_{\mathcal{M}}\sum_{l'\ge l+1}x'^N_{k,m,l'}\right)\right|\\
&\le \sum_{m\in\mathcal{M}}\sum_{l\in\mathbb{N}_0}\sum_{r=1}^{d}\sum_{r_1=1}^{r}\frac{r_1}{r}\frac{d!r_1(r-r_1)(d-r)}{r_1!(r-r_1)!(d-r)!}|(\hat{x}^N_{i,m,l}-x'^N_{k,m,l})|\\
&= \sum_{r=1}^{d}\sum_{r_1=1}^{r}\frac{r_1}{r}\frac{d!r_1(r-r_1)(d-r)}{r_1!(r-r_1)!(d-r)!}\|\hat{\mathbf{x}}^N_i-\mathbf{x}^N_k\|_1.
\end{aligned}\tag{E.5}$$

Let $L=2\sum_{r=1}^{d}\sum_{r_1=1}^{r}\frac{r_1}{r}\frac{d!r_1(r-r_1)(d-r)}{r_1!(r-r_1)!(d-r)!}$. By (E.3), (E.4), and (E.5), we have that for large-enough N,

$$\mathbb{P}(\mathit{Mismatch})\le L\|\hat{\mathbf{x}}^N_i-\mathbf{x}^N_k\|_1. \quad\square\tag{E.6}$$

Appendix F. Doubly Exponential Decay

Proof of Proposition 4. Because $\mathbf{q}$ is a fixed point of (3.7), then we have

$$u_m(q_{m,l}-q_{m,l+1})=\lambda\zeta(q_{m,l-1}-q_{m,l})\sum_{k\in\mathcal{K}}\frac{p_{k,m}w_k}{\delta_k}\frac{(\tilde{q}_{k,l-1})^d-(\tilde{q}_{k,l})^d}{\tilde{q}_{k,l-1}-\tilde{q}_{k,l}}.$$

Multiplying both sides by v_m and summing over $m\in\mathcal{M}$ gives

$$\sum_{m\in\mathcal{M}}v_mu_m(q_{m,l}-q_{m,l+1})=\lambda\zeta\sum_{k\in\mathcal{K}}w_k((\tilde{q}_{k,l-1})^d-(\tilde{q}_{k,l})^d). \tag{F.1}$$

Also, because $q_{m,l}\overset{l\to\infty}{\to}0,\ \forall m\in\mathcal{M}$, then for $\ell\geq1$, by adding $l\geq\ell$, we have

$$\sum_{m\in\mathcal{M}}v_mu_mq_{m,\ell}=\lambda\zeta\sum_{k\in\mathcal{K}}w_k(\tilde{q}_{k,\ell-1})^d. \tag{F.2}$$

From (F.2) and $\sum_{k\in\mathcal{K}}w_k=1$, we have

$$\sum_{m\in\mathcal{M}}v_mu_mq_{m,\ell}\leq\lambda\zeta(\tilde{q}^*_{\ell-1})^d,$$

where $\tilde{q}^*_{\ell-1}=\max_{k\in\mathcal{K}}\tilde{q}_{k,\ell-1}$. Hence, for all $m\in\mathcal{M}$,

$$q_{m,\ell}\leq\frac{\lambda\zeta}{v_mu_m}(\tilde{q}^*_{\ell-1})^d\leq c^*(m,\ell-1)\tilde{q}^*_{\ell-1},$$

where $c^*(m,\ell-1)=(\tilde{q}^*_{\ell-1})^{d-1}\max_{m\in\mathcal{M}}\lambda\zeta/(v_mu_m)$. Because we assume that $q_{m,\ell}\overset{\ell\to\infty}{\to}0$ for all $m\in\mathcal{M}$, then we can choose a large-enough ℓ such that $c^*(m,\ell-1)<1$. By definition, for each $k\in\mathcal{K}$,

$$\tilde{q}_{k,\ell}=\sum_{m\in\mathcal{M}}\frac{v_mp_{k,m}}{\delta_k}q_{m,\ell}\leq c^*(m,\ell-1)(\tilde{q}^*_{\ell-1})^{d-1},$$

which implies that $\tilde{q}^*_\ell\leq c^*(m,\ell-1)\tilde{q}^*_{\ell-1}$ and

$$q_{m,\ell+1}\leq\frac{\lambda\zeta}{v_mu_m}(\tilde{q}^*_\ell)^d\leq(c^*(m,\ell-1)\tilde{q}^*_{\ell-1})^d\max_{m\in\mathcal{M}}\lambda\zeta/(v_mu_m)=(c^*(m,\ell-1))^{d+1}\tilde{q}^*_{\ell-1}.$$

By induction, we obtain that for $n\in\mathbb{N}_0$,

$$q_{m,\ell+n}\leq(c^*(m,\ell-1))^{e(n)}\tilde{q}^*_{\ell-1}\leq(c^*(m,\ell-1))^{d^n}\tilde{q}^*_{\ell-1}, \tag{F.3}$$

where $e(n)=\sum_{i=0}^nd^i$. (F.3) implies that $\{q_{m,l},l\in\mathbb{N}_0\}$ decreases doubly exponentially. □

Remark F.1. Recall $\tilde{q}_{k,l}=\sum_{m\in\mathcal{M}}\frac{v_mp_{k,m}}{\delta_k}q_{m,l}$. From Proposition 4, we know that $\{\tilde{q}_{k,l},l\in\mathbb{N}_0\}$ decreases doubly exponentially. In fact, they do not decay further faster. To see this, let $c_0=\min_{k\in\mathcal{K}}\min_{m\in\mathcal{M}}\frac{p_{k,m}}{\delta_k}\in(0,1]$. Then, $\tilde{q}_{k,l}=\sum_{m=1}^M\frac{v_mp_{k,m}}{\delta_k}q_{m,l}\geq c_0\sum_{m\in\mathcal{M}}v_mq_{m,l}$. It then follows from (F.2) that

$$\min_{k\in\mathcal{K}}\tilde{q}_{k,\ell}\geq c_0\sum_{m\in\mathcal{M}}v_mq_{m,\ell}=\lambda c_0\sum_{k\in\mathcal{K}}w_k(\tilde{q}_{k,\ell-1})^d\geq\lambda c_0\left(\min_{k\in\mathcal{K}}\tilde{q}_{k,\ell-1}\right)^d.$$

So,

$$(\lambda c_0)^{\frac{1}{d-1}}\min_{k\in\mathcal{K}}\tilde{q}_{k,\ell}\geq\left((\lambda c_0)^{\frac{1}{d-1}}\min_{k\in\mathcal{K}}\tilde{q}_{k,\ell-1}\right)^d\geq\cdots\geq\left((\lambda c_0)^{\frac{1}{d-1}}\min_{k\in\mathcal{K}}\tilde{q}_{k,0}\right)^{d^\ell},$$

and hence, $\min_{k\in\mathcal{K}}\tilde{q}_{k,\ell}\geq(\lambda c_0)^{\frac{d^\ell-1}{d-1}}$.

Appendix G. Proof of Lemma 10

Proof of Lemma 10. Fix any $(\alpha_1,\dots,\alpha_M)\in(0,1)^M$ with $\sum_{m\in\mathcal{M}}\alpha_m>0$. Consider any sequence $\{U^N\}_N$ of subsets with $U^N\subseteq V^N$ and $\lim_{N\to\infty}\frac{|U^N\cap V^N_m|}{|V^N_m|}=\alpha_m$ for all $m\in\mathcal{M}$. By Condition 1, we have that for all $k\in\mathcal{K}$ and $m\in\mathcal{M}$,

$$\lim_{N\to\infty}\frac{|E^N_k(U^N\cap V^N_m)|}{|E^N_k(V^N_m)|}=\alpha_mv_m. \tag{G.1}$$

Fix any $\varepsilon>0$, which will be chosen later. Let $\mathscr{G}^N_{k,\varepsilon}=\left\{i\in W^N_k:\left|\frac{|\mathcal{N}^N_w(i)\cap v|}{|\mathcal{N}^N_w(i)|}-\frac{|E^N_k(v)|}{|E^N_k(V^N)|}\right|\geq\varepsilon\right\}$ and $\mathscr{B}^N_{k,\varepsilon}=W^N_k\setminus\mathscr{G}^N_{k,\varepsilon}$. By (G.1), for all large-enough N and $i\in\mathscr{G}^N_{k,\varepsilon}$,

$$N(1-2\varepsilon)\sum_{m\in\mathcal{M}}\alpha_mv_mp_{k,m}\leq|\mathcal{N}^N_w(i)\cap U^N|\leq N(1+2\varepsilon)\sum_{m\in\mathcal{M}}\alpha_mv_mp_{k,m}. \tag{G.2}$$

Also, by Condition 1, for all large-enough N,

$$N\delta_k(1-\varepsilon) \le \delta_i^N \le N\delta_k(1+\varepsilon). \tag{G.3}$$

Because the sequence $\{G^N\}_N$ is in the subcritical, then for large-enough N,

$$\begin{aligned}
\rho \ge \rho^N &\ge \left(\sum_{j\in U^N}\sum_{m\in\mathcal{M}'} \mathbb{1}_{(j\in V_m^N)} u_m\right)^{-1} \sum_{i\in W^N} \sum_{\substack{S\subseteq (U^N\cap \mathcal{N}_w^N(i)):\\ |S|=d}} \frac{\lambda}{\binom{|\mathcal{N}_w^N(i)|}{d}} \\
&\ge c'(N)\left(\sum_{m\in\mathcal{M}'} N v_m\alpha_m u_m\right)^{-1} \sum_{k\in\mathcal{K}}\sum_{i\in W_k^N} \frac{\lambda\binom{|U^N\cap\mathcal{N}_w^N(i)|}{d}}{\binom{|\mathcal{N}_w^N(i)|}{d}} \\
&\ge \left(\sum_{m\in\mathcal{M}'} N v_m\alpha_m u_m\right)^{-1} \sum_{k\in\mathcal{K}} \frac{\lambda|\mathcal{G}_{k,\varepsilon}^N|\binom{N(1-2\varepsilon)\sum_{m\in\mathcal{M}}\alpha_m v_m p_{k,m}}{d}}{\binom{N\delta_k(1+\varepsilon)}{d}},
\end{aligned} \tag{G.4}$$

where $c'(N)$ is a constant only depending on N with $c'(N)\overset{N\to\infty}{\to} 1$. Because the sequence $\{G^N\}$ is proportionally sparse, then $\lim_{N\to\infty}\frac{|\mathcal{G}_{k,\varepsilon}^N|}{|W_k^N|}=1$. Then, we have

$$\rho \ge \left(\sum_{m\in\mathcal{M}} v_m\alpha_m p_{k,m}\right)^{-1} \lambda\zeta \sum_{m\in\mathcal{M}} w_k\left(\frac{(1-2\varepsilon)\sum_{m\in\mathcal{M}}\alpha_m v_m p_{k,m}}{\delta_k(1+\varepsilon)}\right)^d. \tag{G.5}$$

Because (G.5) holds for all $\varepsilon>0$, then

$$\rho \ge \left(\sum_{m\in\mathcal{M}} v_m\alpha_m p_{k,m}\right)^{-1} \lambda\zeta \sum_{m\in\mathcal{M}} w_k\left(\frac{\sum_{m\in\mathcal{M}}\alpha_m v_m p_{k,m}}{\delta_k}\right)^d. \quad \square \tag{G.6}$$

Appendix H. Proof of Lemma 11

Proof of Lemma 11. Given the system state X^N, when a task arrives at the system, by the Poisson thinning property, the probability that the task will be assigned to a server in the set $Q_{m,l}^N(X^N)$ is

$$\mathbb{P}(\mathcal{E}(Q_{m,l}^N)) = \frac{1}{W(N)}\sum_{i\in W^N}\sum_{\substack{U\subseteq (Q_{m,l}^N\cap\mathcal{N}_w^N(i))\\ |U|=d}} \frac{1}{\binom{\mathcal{N}_w^N(i)}{d}}, \tag{H.1}$$

where $\mathcal{E}(Q_{m,l}^N)$:= the event that the new task will be assigned to $Q_{m,l}^N(X^N)$. Fix any $\varepsilon>0$. Because the sequence $\{G^N\}$ is subcritical, then for large-enough N, we have that

$$\mathbb{P}(\mathcal{E}(Q_{m,l}^N)) \le \frac{N}{W(N)}\frac{\rho^N}{\lambda}\frac{|Q_{m,l}^N(X^N)|u_m}{N} \le \frac{\rho}{\lambda\zeta} q_{m,l}^N u_m(1+\varepsilon). \tag{H.2}$$

We consider the system state at event times $t_0=0<t_1<t_2<\cdots<t_i<\cdots$; for all i, t_i can be an arrival or a potential departure epoch. Define the drift $\Delta L_{m,\ell}^N(X^N)$ as

$$\Delta L_{m,\ell}^N(X^N) = \mathbb{E}(L_{m,\ell}^N(X^N(t_1)) - L_{m,\ell}^N(X^N)|X^N(t_0)=X^N). \tag{H.3}$$

Again, by the Poisson thinning property, we have that for all large N,

$$\begin{aligned}
\Delta L_{m,\ell}^N(X^N) &= \sum_{i=\ell}^{\infty}\left(\frac{\lambda W(N)}{\lambda W(N)+\sum_{m\in\mathcal{M}}|V_m^N|u_m}\mathbb{P}(\mathcal{E}(Q_{m,i-1}^N)) - \frac{\sum_{m\in\mathcal{M}}|V_m^N|u_m}{\lambda W(N)+\sum_{m\in\mathcal{M}}|V_m^N|u_m}\frac{|Q_{m,i}^N|u_m}{\sum_{m\in\mathcal{M}}|V_m^N|u_m}\right) \\
&\le \sum_{i=\ell}^{\infty}\left(\frac{\rho q_{m,i-1}^N u_m(1+\varepsilon)}{\lambda\zeta+\sum_{m\in\mathcal{M}}v_m u_m} - \frac{q_{m,i}^N u_m}{\lambda\zeta+\sum_{m\in\mathcal{M}}v_m u_m}\right) \\
&= \frac{\rho q_{m,\ell-1}^N u_m(1+\varepsilon)}{\lambda\zeta+\sum_{m\in\mathcal{M}}v_m u_m} - \frac{1-(1+\varepsilon)\rho}{\lambda\zeta+\sum_{m\in\mathcal{M}}v_m u_m}\sum_{i=\ell}^{\infty} q_{m,i}^N u_m.
\end{aligned} \tag{H.4}$$

By the definition of the steady state, $\mathbb{E}(\Delta L^N_{m,\ell} X^N(\infty)) = 0$. Choosing ε such that $(1+\varepsilon)\rho \le (1+\rho)/2 < 1$, we have

$$\sum_{i=\ell}^{\infty} \mathbb{E}(q^N_{m,i}(\infty)) \le \frac{(1+\rho)/2}{1-(1+\rho)/2} \mathbb{E}(q^N_{m,\ell-1}). \tag{H.5}$$

Finally, summing over $m \in \mathcal{M}$, we get the desired result. □

Appendix I. Proof for the Sequence of Random Graphs

Proof of Theorem 8. First, to show that the sequence $\{G^N\}_N$ satisfies Condition 1, consider any fixed $k \in \mathcal{K}$ and $m \in \mathcal{M}$. Let $e_{i,j}$ be a Bernoulli random variable with probability $p_{k,m}$ for each $i \in W^N_k$ and $j \in V^N_m$. Then, $E^N(k,m) = \sum_{(i,j)\in W^N_k \times V^N_m} e_{i,j}$, and by the L.L.N., we have that

$$\lim_{N\to\infty} \frac{E^N(k,m)}{|W^N_k| \times |V^N_m|} = p_{k,m},$$

which implies that Condition 1(a) holds. Next, we prove that Condition 1(b) holds. Based on the definition $\deg^N_w(i)$, we have $\deg^N_w(i) = \sum_{j\in V^N_m} e_{i,j}$, which is a binomial random variable Binomial($|V^N_m|, p_{k,m}$). By the Chernoff bound (Cheng and Yang 2005, theorem 2.4), it follows that for $i \in W^N_k$,

$$\mathbb{P}(|\deg^N_w(i) - \mathbb{E}(\deg^N_w(i))| \ge x) \le 2\exp\left(-\frac{x^2}{2\mathbb{E}(\deg^N_w(i)) + 2x/3}\right).$$

Let $X(N) = p_{k,m} N^{3/4} (\ln(N))^{1/4}$. Then, for some $c_1 \in (0,\infty)$,

$$\mathbb{P}(|\deg^N_w(i) - |V^N_m| p_{k,m}| \ge X(N)) \le c_1 \exp(-c_1 p_{k,m} N^{1/2} (\ln(N))^{1/2}/v_m) \tag{I.1}$$

for sufficiently large N. Also, by $\lim_{N\to\infty} \frac{W^N_k}{W(N)} = w_k$, $\lim_{N\to\infty} \frac{W(N)}{N} = \zeta$, and the union bound, we have that there exists $c_2 \in (0,\infty)$ such that for large-enough N,

$$\mathbb{P}(\cup_{i\in W^N_k} |\deg^N_w(i) - |V^N_m| p_{k,m}| \ge X(N)) \le c_2 w_k \zeta N \exp(-c_1 p_{k,m} N^{1/2} (\ln(N))^{1/2}/v_m). \tag{I.2}$$

Then, the RHS of (I.2) is summable over N. From the Borel–Cantelli lemma, we get that a.s., for all large-enough N,

$$|\deg^N_w(i) - |V^N_m| p_{k,m}| \le X(N), \quad i \in W^N_k,$$

which implies that the following equation holds:

$$1 \le \lim_{N\to\infty} \frac{\max_{i\in W^N_k} \deg^N_w(i)}{\min_{i\in W^N_k} \deg^N_w(i)} \le \lim_{N\to\infty} \frac{|V^N_m| p_{k,m} + X(N)}{|V^N_m| p_{k,m} - X(N)} = 1, \quad \text{a.s.}.$$

Thus, Condition 1(b) holds.

Now, we show that the sequence $\{G^N\}_N$ is clustered proportionally sparse. Fix any $k \in \mathcal{K}$, $i \in W^N_k$, $\varepsilon > 0$, and $U \subseteq V^N$. Let $B_i(U)$ be the event that the dispatcher i is bad w.r.t. the set U: that is,

$$B_i(U) := \left\{ \left| \frac{\mathcal{N}^N_w(i) \cap U}{\mathcal{N}^N_w(i)} - \frac{E^N_k(U)}{E^N_k(V^N)} \right| \ge \varepsilon \right\}. \tag{I.3}$$

Define $\alpha_m := \frac{|U \cap V^N_m|}{|V^N|}$ for each $m \in \mathcal{M}$. By the union bound, we have that

$$\begin{aligned}
\mathbb{P}(B_i(U)) \le \mathbb{P}\Bigg(& B_i(U), \left| |\mathcal{N}^N_w(i) \cap U| - \sum_{m\in\mathcal{M}} |V^N_m \cap U| p_{k,m} \right| < \varepsilon_1 \sum_{m\in\mathcal{M}} |V^N_m| p_{k,m}, \\
& \left| \frac{E^N_k(U)}{E^N_k(V^N)} - \frac{\sum_{m\in\mathcal{M}} \alpha_m p_{k,m}}{\sum_{v_m p_{k,m}}} \right| < \varepsilon_2, \text{ and } \left| \mathcal{N}^N_w(i) - \sum_{m\in\mathcal{M}} |V^N_m| p_{k,m} \right| < \varepsilon_3 \sum_{m\in\mathcal{M}} |V^N_m| p_{k,m} \Bigg) \\
& + \mathbb{P}\left(\left| |\mathcal{N}^N_w(i) \cap U| - \sum_{m\in\mathcal{M}} |V^N_m \cap U| p_{k,m} \right| \ge \varepsilon_1 \sum_{m\in\mathcal{M}} |V^N_m| p_{k,m} \right) \\
& + \mathbb{P}\left(\left| \mathcal{N}^N_w(i) - \sum_{m\in\mathcal{M}} |V^N_m| p_{k,m} \right| \ge \varepsilon_2 \sum_{m\in\mathcal{M}} |V^N_m| p_{k,m} \right).
\end{aligned} \tag{I.4}$$

We will bound each term of the RHS of (I.4). By choosing ε_1, ε_2 and ε_3 satisfying

$$\frac{\varepsilon_3 \sum_{m\in\mathcal{M}} \alpha_m p_{k,m} + \varepsilon_1 \sum_{m\in\mathcal{M}} v_m p_{k,m}}{(1-\varepsilon_3)\sum_{m\in\mathcal{M}} v_m p_{k,m}} + \varepsilon_2 < \varepsilon, \tag{I.5}$$

we have that

$$\frac{\mathcal{N}_w^N(i)\cap U}{\mathcal{N}_w^N(i)}-\frac{E_k^N(U)}{E_k^N(V^N)}=\frac{\mathcal{N}_w^N(i)\cap U}{\mathcal{N}_w^N(i)}-\frac{\sum_{m\in\mathcal{M}}\alpha_m p_{k,m}}{\sum_{v_m p_{k,m}}}+\frac{\sum_{m\in\mathcal{M}}\alpha_m p_{k,m}}{\sum_{v_m p_{k,m}}}-\frac{E_k^N(U)}{E_k^N(V^N)}$$
$$<\frac{\varepsilon_3\sum_{m\in\mathcal{M}}\alpha_m p_{k,m}+\varepsilon_1\sum_{m\in\mathcal{M}}v_m p_{k,m}}{(1-\varepsilon_3)\sum_{m\in\mathcal{M}}v_m p_{k,m}}+\varepsilon_2<\varepsilon \tag{I.6}$$

and

$$\frac{\mathcal{N}_w^N(i)\cap U}{\mathcal{N}_w^N(i)}-\frac{E_k^N(U)}{E_k^N(V^N)}=\frac{\mathcal{N}_w^N(i)\cap U}{\mathcal{N}_w^N(i)}-\frac{\sum_{m\in\mathcal{M}}\alpha_m p_{k,m}}{\sum_{v_m p_{k,m}}}+\frac{\sum_{m\in\mathcal{M}}\alpha_m p_{k,m}}{\sum_{v_m p_{k,m}}}-\frac{E_k^N(U)}{E_k^N(V^N)}$$
$$>-\frac{\varepsilon_3\sum_{m\in\mathcal{M}}\alpha_m p_{k,m}+\varepsilon_1\sum_{m\in\mathcal{M}}v_m p_{k,m}}{(1+\varepsilon_3)\sum_{m\in\mathcal{M}}v_m p_{k,m}}-\varepsilon_2>-\varepsilon, \tag{I.7}$$

which implies that the first term is equal to zero with ε_1, ε_2, and ε_3. Using the Chernoff bound again, we can bound the second term and the third term as follows; for some $c_3\in(0,\infty)$ and large-enough N,

$$\mathbb{P}\left(\left|\,|\mathcal{N}_w^N(i)\cap U|-\sum_{m\in\mathcal{M}}|V_m^N\cap U|p_{k,m}\right|\geq\varepsilon_1\sum_{m\in\mathcal{M}}|V_m^N|p_{k,m}\right)\leq c_3\exp\left(-c_3N\sum_{m\in\mathcal{M}}v_m p_{k,m}\right) \tag{I.8}$$

and

$$\mathbb{P}\left(\left|\mathcal{N}_w^N(i)-\sum_{m\in\mathcal{M}}|V_m^N|p_{k,m}\right|\geq\varepsilon_2\sum_{m\in\mathcal{M}}|V_m^N|p_{k,m}\right)\leq c_3\exp\left(-c_3N\sum_{m\in\mathcal{M}}v_m p_{k,m}\right). \tag{I.9}$$

Therefore, for large-enough N, we have

$$\mathbb{P}(B_i(U))\leq 2c_3\exp\left(-c_3N\sum_{m\in\mathcal{M}}v_m p_{k,m}\right) \tag{I.10}$$

and

$$\mathbb{P}(\cup_{i\in W_k^N}B_i(U))\leq 2c_3|W_k^N|\exp\left(-c_3N\sum_{m\in\mathcal{M}}v_m p_{k,m}\right). \tag{I.11}$$

Moreover, for some $c_4\in(0,\infty)$ and large-enough N,

$$\mathbb{P}\left(\sup_{U\subseteq V^N}\cup_{i\in W_k^N}B_i(U)\right)\leq\exp(-c_4N). \tag{I.12}$$

The RHS of (I.12) is summable over N, and the set $\mathcal{K}$ is finite; so, by the Borel Cantelli lemma, the sequence is clustered proportionally sparse.

If $\mathbf{p}$ satisfies (3.1), by Lemma 2, there exists an $N_0\in\mathbb{N}_0$ such that for all $N\geq N_0$, the queue-length process $(X_j^N(t))_{j\in V^N}$ under the local JSQ(d) policy is ergodic, which implies that all assumptions of Theorem 6 hold. □

References

Adler M, Chakrabarti S, Mitzenmacher M, Rasmussen L (1995) Parallel randomized load balancing. *Proc. 27th Ann. ACM Sympos. Theory Comput. (STOC '95)* (Association for Computing Machinery, New York), 238–247.

Allmeier S, Gast N (2022) Mean field and refined mean field approximations for heterogeneous systems: It works! *Proc. ACM Measurement Anal. Comput. Systems* 6(1):13.

Bhambay S, Mukhopadhyay A (2022) Asymptotic optimality of speed-aware JSQ for heterogeneous service systems. *Performance Evaluation* 157–158(2022):10232.

Bramson M (2011) Stability of join the shortest queue networks. *Ann. Appl. Probab.* 21(4):1568–1625.

Budhiraja A, Mukherjee D, Wu R (2019) Supermarket model on graphs. *Ann. Appl. Probab.* 29(3):1740–1777.

Cheng QM, Yang H (2005) Inequalities for eigenvalues of a clamped plate problem. *Trans. Amer. Math. Soc.* 358(6):2625–2635.

Cruise J, Jonckheere M, Shneer S (2020) Stability of JSQ in queues with general server-job class compatibilities. *Queueing Systems* 95(3–4):271–279.

Ethier SN, Kurtz TG (2009) *Markov Processes: Characterization and Convergence* (John Wiley & Sons, Hoboken, NJ).

Foss SG, Chernova NI (1998) On the stability of a partially accessible multi-station queue with state-dependent routing. *Queueing Systems* 29(1):55–73.

Gardner K, Abdul Jaleel J, Wickeham A, Doroudi S (2021) Scalable load balancing in the presence of heterogeneous servers. *Performance Evaluation* 145(2021):102151.

Gast N (2015) The power of two choices on graphs: The pair-approximation is accurate. *Performance Evaluation Rev.* 43(2):69–71.

Habib M, McDiarmid C, Ramirez-Alfonsin J, Reed B (1998) *Probabilistic Methods for Algorithmic Discrete Mathematics*, vol. 16 (Springer, Berlin).

Hurtado-Lange D, Maguluri ST (2021) Throughput and delay optimality of power-of-d choices in inhomogeneous load balancing systems. *Oper. Res. Lett.* 49(4):616–622.

Martin JB, Suhov YM (1999) Fast Jackson networks. *Ann. Appl. Probab.* 9(3):854–870.

Méléard S (1996) Asymptotic behaviour of some interacting particle systems; Mckean–Vlasov and Boltzmann models. Talay D, Tubaro L, eds. *Probabilistic Models for Nonlinear Partial Differential Equations*, Lecture Notes in Mathematics, vol. 1627 (Springer, Berlin), 42–95.

Mitzenmacher M (1996a) Load balancing and density dependent jump Markov processes. *Proc. 37th Conf. Foundations Comput. Sci.* (IEEE, Piscataway, NJ), 213–222.

Mitzenmacher M (1996b) The power of two choices in randomized load balancing. PhD thesis, University of California, Berkeley.

Mukherjee D, Borst SC, Van Leeuwaarden JSH (2018a) Asymptotically optimal load balancing topologies. *Proc. ACM Measurement Anal. Comput. Systems* 2(1):14.

Mukherjee D, Borst SC, van Leeuwaarden JSH, Whiting PA (2018b) Universality of power-of-d load balancing in many-server systems. *Stochastic Systems* 8(4):265–292.

Mukhopadhyay A, Mazumdar RR (2016) Analysis of randomized Join-the-Shortest-Queue (JSQ) schemes in large heterogeneous processor-sharing systems. *IEEE Trans. Control Network Systems* 3(2):116–126.

Mukhopadhyay A, Karthik A, Mazumdar RR (2016) Randomized assignment of jobs to servers in heterogeneous clusters of shared servers for low delay. *Stochastic Systems* 6(1):90–131.

Rutten D, Mukherjee D (2022) Load balancing under strict compatibility constraints. *Math. Oper. Res.* 48(1):227–256.

Stolyar AL (2005) Optimal routing in output-queued flexible server systems. *Probab. Engrg. Inform. Sci.* 19(2):141–189.

Stolyar AL (2015) Pull-based load distribution in large-scale heterogeneous service systems. *Queueing Systems* 80(4):341–361.

Stolyar AL (2017) Pull-based load distribution among heterogeneous parallel servers: The case of multiple routers. *Queueing Systems* 85(1):31–65.

Sznitman AS (1991) *Topics in Propagation of Chaos* (Springer, Berlin).

Tirmazi M, Barker A, Deng N, Haque ME, Qin ZG, Hand S, Harchol-Balter M, Wilkes J (2020) Borg: The next generation. *Proc. 15th Eur. Conf. Comput. Systems (EuroSys '20)* (Association for Computing Machinery, New York), 1–14.

Tsitsiklis JN, Xu K (2013) Queueing system topologies with limited flexibility. *Proc. ACM SIGMETRICS/Internat. Conf. Measurement and Modeling Comput. Systems (SIGMETRICS '13)* (Association for Computing Machinery, New York), 167–178.

Tsitsiklis JN, Xu K (2017) Flexible queueing architectures. *Oper. Res.* 65(5):1398–1413.

Turner SR (1998) The effect of increasing routing choice on resource pooling. *Probab. Engrg. Inform. Sci.* 12(1):109–124.

van der Boor M, Borst S, van Leeuwaarden J, Mukherjee D (2022) Scalable load balancing in networked systems: A survey of recent advances. *SIAM Rev.* 64(3):554–622.

Vvedenskaya ND, Dobrushin RL, Karpelevich FI (1996) Queueing system with selection of the shortest of two queues: An asymptotic approach. *Problemy Peredachi Informatsii* 32(1):20–34.

Weng W, Zhou X, Srikant R (2020) Optimal load balancing with locality constraints. *Proc. ACM Measurement Anal. Comput. Systems* 4(3):45.