The Annals of Applied Probability
2024, Vol. 34, No. 6, 5228–5257
https://doi.org/10.1214/24-AAP2090

MEAN-FIELD ANALYSIS FOR LOAD BALANCING ON SPATIAL GRAPHS

BY DAAN RUTTEN[a] AND DEBANKUR MUKHERJEE[b]

Industrial and Systems Engineering, Georgia Institute of Technology, [a]*drutten@gatech.edu,*
[b]*debankur.mukherjee@isye.gatech.edu*

The analysis of large-scale, parallel-server load balancing systems has relied heavily on mean-field analysis. A pivotal assumption for this framework is that servers are exchangeable. However, modern data-centers have *data locality constraints*, such that tasks of a particular type can only be routed to a small subset of servers. An emerging line of research, therefore, considers load balancing algorithms on bipartite graphs where vertices represent task types and servers, respectively. Due to the lack of exchangeability in this model, mean-field techniques fundamentally break down. Recent progress has been made on graphs with strong edge-expansion properties, that is, where *any* two large subsets of vertices are well-connected. However, data locality often leads to spatial graphs that do not have strong expansion properties.

In this paper, we develop a novel coupling-based approach to establish mean-field approximation for a large class of graphs which includes spatial graphs. The method extends the scope of mean-field analysis far beyond the classical full-flexibility setup. En route, we prove that, starting from suitable states, the occupancy process becomes close to its steady state in a time that is independent of system size, which might be of independent interest. Numerical experiments are conducted, which positively support the theoretical results.

1. Introduction.

1.1. *Background and motivation*. The study of load balancing algorithms for large-scale systems started with the seminal works of Mitzenmacher [27] and Vvedenskaya et al. [39]. Since then, there has been a significant development in our understanding of the performance of various load balancing policies and their tradeoffs between quantities like user-perceived delay, communication overhead, implementation complexity and energy consumption; see, for example, [3, 4, 11, 12, 15, 17, 22, 32, 37, 38, 40] for a few recent, representative works from various research domains. A pivotal methodological tool behind this success has been *mean-field analysis*. The history of mean-field analysis, in its current form, goes back to the foundational works of Kurtz [18–20], Norman [30, 31] and Barbour [5]. The high-level idea is to represent the system state by aggregate Markovian quantities and characterize their rate of change as the system size grows large. In the context of load balancing, this representation is the occupancy process $\boldsymbol{q}^N(t) = (q_i^N(t))_{i\geq 1}$, where $q_i^N(t)$ denotes the fraction of servers with queue length at least i in a system with N servers at time t. As $N \to \infty$, $\boldsymbol{q}^N(t)$ tends to behave like a deterministic, continuous system described by an ordinary differential equation (ODE) that is analytically tractable. A pivotal assumption for the above scheme to work is that the aggregate quantity $\boldsymbol{q}^N(t)$ is Markovian such that its rate of change can be expressed as a function of its current state. If $\boldsymbol{q}^N(t)$ is not Markovian, not only does this technique break down, the mean-field approximation may even turn out to be highly inaccurate.

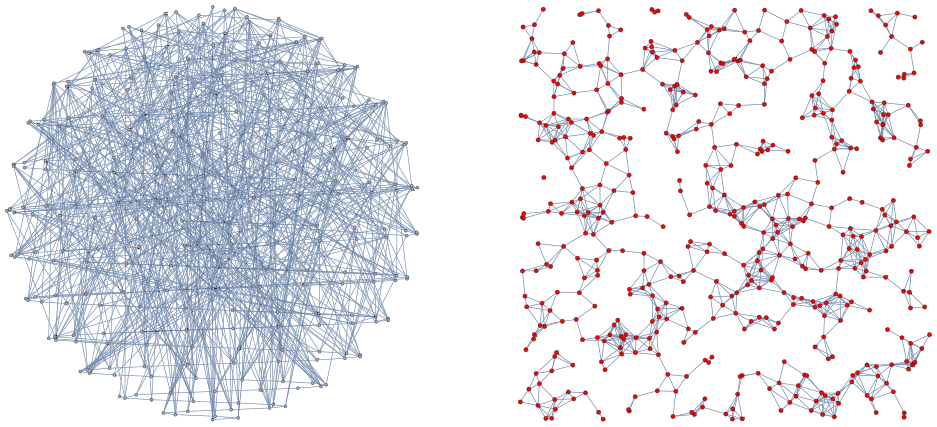
Received January 2023; revised June 2024.
MSC2020 subject classifications. Primary 60J27; secondary 60G55.
Key words and phrases. Meanfield approximation, power-of-d, stochastic coupling, load balancing on network, data locality, many-server asymptotics, queueing theory.

In load balancing systems, if servers are exchangeable, then $\boldsymbol{q}^N(t)$ is indeed Markovian. However, the growing heterogeneity in the types of tasks processed by modern data centers has recently motivated the research community to consider systems beyond the exchangeability assumption. The main reason stems from *data locality*, that is, the fact that servers need to store resources to process tasks of a particular type locally and have only limited storage space. Examples of these resources may include databases or machine learning models specific to particular tasks. This limits the flexibility of the assignment of a task to a queue, which now needs to ensure that the corresponding server is able to process the assigned task. In fact, the lack of flexibility also arises in much broader contexts such as due to a spatially constrained network architecture (e.g., in bike-sharing), see [14, 25, 32], or in the context of geographically distributed data centers [21, 24]. An emerging line of work thus considers a bipartite graph between task types and servers; see, for example, [9, 10, 29, 35, 36, 41]. In this *compatibility* graph, an edge between a server and a task type represents the server's ability to process these tasks. In this model, if the graph is complete bipartite, then the problem reduces to the classical case of a fully flexible system. In reality, the storage capacity or geographical constraints forces a server to process only a small subset of all task types, leading to sparser network topologies. This motivates the study of load balancing in systems with suitably *sparse* bipartite compatibility graphs.

1.2. *Fundamental barriers*. The analysis of sparse systems poses significant challenges, mainly due to the fact that the vector $\boldsymbol{q}^N(t)$ is no longer Markovian. In fact, for general graphs, there does not even exist a Markovian state descriptor that is an aggregate quantity such as $\boldsymbol{q}^N(t)$, and one needs to keep track of the evolution of the entire system in order to know the instantaneous transition rates. These barriers are the reason, as noted as early as by Mitzenmacher in his thesis [27], that a network topology is a "*very interesting question*... (but) *seems to require different techniques*". One key question to understand here is: *Under what conditions on the (sparse) compatibility graph does the system behavior retain the performance benefits (in terms of the queue length behavior) of the fully flexible system?* From a more foundational standpoint, this is equivalent to understanding how much the validity of the mean-field approximation can be extended to nontrivial graphs.

A few recent works have made successful attempts in analyzing compatibility graphs that possess the proper edge-expansion properties [29, 35, 41], of which [35] is most relevant to the current work. Here, the JSQ(d) policy was considered, where each arriving task joins the shortest of d randomly selected compatible queues. The authors showed that if the graph is "well-connected", the limiting occupancy process is indistinguishable from the fully flexible system both in the transient limit and in steady state. Even though the well-connectedness condition allows the graph to be sparse, it requires the graph to have strong edge-expansion properties in the following sense: *Pick any subset of servers of size δN for $\delta > 0$ however small. Then, asymptotically, almost all task types should be connected to this set and have a δ fraction of their compatible servers in that set.* This condition allows the authors in [35] to ensure that, for any occupancy measure, each task type observes approximately the same queue length distribution within their set of compatible servers. As a result, the evolution of the queue length distribution in any neighborhood happens in the same way and this ensures that, asymptotically, the process evolves in the same way as the fully flexible system.

The well-connectedness property is not satisfied by spatial graphs such as random geometric graphs [33]. The edges in a spatial graph are "local", and hence dispatchers in one location cannot assign tasks to servers in spatially distant locations (see Figure 1). However, as already pointed out via numerical simulations in [35], in steady state, sparse graphs still retain the performance benefits of a fully flexible system, even though the neighborhood coupling based method in [35] fails for these graphs.



(a) Erdős–Rényi graph (b) Random geometric graph

FIG. 1. *Examples of graph topologies generated by an Erdős–Rényi graph and a random geometric graph with same average degree* $\ln N$, *where* N *is the number of vertices. The picture illustrates the fundamental difference between the nature of global vs. local connections in the two graphs.*

Aside from the technical difficulties, there is a fundamental barrier that prevents the mean-field approximation from being applied to spatial graphs. This can be understood by a simple counterexample: If all the high queues in the system are located in a small spatial region, then the behavior of the system will be qualitatively different from when they are spread out across the system, as it will take more time for the congestion to disperse throughout the rest of the graph. In general, in these situations, the behavior of the system in a local neighborhood of the graph may be very different from the global behavior. Therefore, *one cannot expect the transient behavior of a system with spatial compatibility constraints to coincide with the fully flexible system.* However, in steady state, it may happen that the situation described in the above counterexample does not occur with high probability, making the steady state still behave like the fully flexible system. Thus, one needs to characterize the limit of the steady state distribution *without* proceeding via the process-level mean-field limit, as the transient limit will be provably different. Alternative techniques such as the moment generating function (MGF) method and Lyapunov approaches may allow moment bounds on the steady-state via Stein's method, but cannot commonly be used for the exact characterization of the limit of stationary distributions. Although Stein's method has been successfully used for analyzing the join-shortest-queue (JSQ) policy [41], these results critically rely on the *state space collapse* or the degeneracy of the steady state, observed as a consequence of JSQ (i.e., all queues are of length zero or one, asymptotically). When the limit of the stationary distributions is nondegenerate, as is the case in the current paper, we enter uncharted waters in the mean-field approximation literature, and formalizing a new method to take care of the above difficulties is one of the main contributions of this paper.

2. Main contributions. Let $G_N = (V_N, W_N, E_N)$ be a bipartite graph, where V_N denotes the set of servers, W_N denotes the set of task types and $E_N \subseteq V_N \times W_N$ denotes the compatibility constraints. Throughout, we will use the words task-types and dispatchers interchangeably. Here, $N := |V_N|$ equals the number of servers and $M(N) := |W_N|$ equals the number of task types. Let $\mathcal{N}_v := \{w \in W_N : (v, w) \in E_N\}$ be the compatible task types for a server $v \in V_N$ and $\mathcal{N}_w := \{v \in V_N : (v, w) \in E_N\}$ be the compatible servers for a task type $w \in W_N$. Denote $d_v^N = |\mathcal{N}_v|$ and $d_w^N = |\mathcal{N}_w|$. Tasks of each type arrive as independent Poisson processes of rate $\lambda N/M(N)$ and each task requires an independent and exponentially distributed service time with mean one. Thus, the total arrival rate is λN and we assume

$\lambda < 1$ to ensure stability of the system. If a task arrives at a dispatcher $w \in W_N$, then $d \geq 2$ servers are sampled uniformly at random from $\mathcal{N}_w$ with replacement, and the task is assigned to the shortest queue among the selected servers, breaking ties at random. The tasks in the queue are handled one at a time in first come, first served order.

The criteria for ergodicity of the queue length process for such a system are known and have been developed, for example, by Bramson [6] and Cardinaels et al. [10]. However, in this paper, we work with a slightly stronger, but simplified condition on the graph as follows. Let

$$\rho(G_N) := \max_{v \in V_N} \frac{\lambda N}{M(N)} \sum_{w \in \mathcal{N}_v} \frac{1}{d_w^N}. \tag{2.1}$$

Using Lyapunov arguments, it is not hard to show that $\rho(G_N) < 1$ implies that the queue length process of the system is ergodic for any $d \geq 2$ (Proposition 4.1). Conceptually, $\rho(G_N)$ is the maximum load on a server if each dispatcher uses random routing ($d = 1$) and hence it should seem natural that this condition implies stability also for $d \geq 2$. To avoid heavy-traffic behavior as $N \to \infty$, we will assume that $\rho(G_N) \leq \rho_0$ for all $N \geq 1$ for a constant $\rho_0 < 1$ throughout.

REMARK 1. In comparison, the stability condition in [10] reduces to: the queue length process is ergodic if for all $w \in W_N$ and $U \subseteq V_N$, there exists a probability distribution $p_{w,U}(\cdot)$ on U such that

$$\max_{v \in V_N} \frac{\lambda N}{M(N)} \sum_{w \in W_N} \binom{|\mathcal{N}_w|}{d}^{-1} \sum_{\substack{U \subseteq \mathcal{N}_w \\ |U| = \min(d, |\mathcal{N}_w|)}} p_{w,U}(v) < 1. \tag{2.2}$$

However, to prove the mean-field approximation, we require later that $\frac{N}{M(N)} \sum_{w \in \mathcal{N}_v} \frac{1}{d_w^N} \approx 1$ for all $v \in V_N$ (see the definition of $\gamma(G_N)$ in (2.3) and Corollary 3.2) and hence $\rho(G_N) \approx \lambda < 1$ follows immediately. We will therefore work with the simplified stability condition in (2.1).

We make contributions on four fronts: (a) We establish bounds on a *large-scale mixing time* of the underlying Markov process; (b) we quantify how much the transient behavior deviates from the mean-field ODE, starting from i.i.d queue lengths, in terms of certain graph parameters; (c) we combine (a) and (b) to formulate a criterion of when the global quantity $\boldsymbol{q}^N(t)$ is asymptotically indistinguishable from the fully flexible system in steady state; and finally (d) we show how standard generative models for sparse spatial graphs and a large class of sparse regular graphs satisfy this criterion for convergence.

(*a*) *Large-scale mixing time bounds*. Mixing time bounds for large-scales systems are known to be hard to obtain. Even without any compatibility constraints, bounding the mixing time for the JSQ(d) policy for large N requires significant work [23]. First, as discussed in [23], a major challenge is posed by the effect of the starting state. As the state space is infinite, if the system starts from a bad corner of the state space, it may take a very long time to come back to the "regular states", which may even render a mixing time bound useless for our purposes. Second, in the presence of a compatibility graph structure, regenerative arguments, such as bounding the time the Markov process takes to hit a fixed state [13], cannot be used either since these regeneration lengths are typically exponential in N. In fact, for large-scale analysis we do not require the conventional notion of mixing time. Instead, we introduce a notion of *large-scale mixing time* as follows: starting from a set of suitable states, if we compare the distribution of $\boldsymbol{q}^N(t)$ and its steady-state distribution, when can we say that they are

"close" in a suitable sense? Here, it is worth pointing out that, since $\boldsymbol{q}^N(t)$ is not a Markov process, by its steady-state distribution we mean the functional $\boldsymbol{q}^N(t)$ evaluated on the system in steady state. We show that this mixing time does not scale with N (Theorem 3.8). This implies that, starting from the set of suitable states, observing the system at this mixing time will give us a good approximation of the steady state. In the above, the set of "suitable states" in particular includes the empty state. A crucial argument in the proof of Theorem 3.8 relies on a novel stochastic coupling. If one copy of the system starts from a state where the queue length at each server is at most the queue length of the corresponding server in another copy of the system, then there exists a stochastic coupling such that this ordering is maintained throughout for any sample path (Proposition 3.9). We believe that Proposition 3.9 and Theorem 3.8 hold for a large class of such monotone property, which may be of independent interest.

(*b*) *Process-level limit starting from i.i.d. queue lengths*. As the system quickly converges to the steady-state from any of the set of suitable states, it is sufficient to characterize the sample path of a subset of these states. Thus, we next characterize the asymptotics of the sample path of $\boldsymbol{q}^N(t)$ starting from a system with i.i.d. queue lengths. Let us introduce two quantities of the underlying graph:

$$\phi(G_N) := \max_{v \in V_N} \left| \frac{N}{M(N)} \sum_{w \in \mathcal{N}_v} \frac{1}{d_w^N} - 1 \right| \quad \text{and} \quad \gamma(G_N) := \frac{1}{M(N)} \sum_{w \in W_N} \frac{1}{d_w^N}. \tag{2.3}$$

Loosely speaking, $\phi(G_N)$ quantifies the extent to which the bipartite graph is regular and $\gamma(G_N)$ describes the average inverse degree of the task types. For example, if $d_w^N = d_{\text{task}}^N$ for all $w \in W_N$ and $d_v^N = d_{\text{server}}^N$ for all $v \in V_N$, then $\phi(G_N) = 0$ and $\gamma(G_N) = 1/d_{\text{task}}^N$ (see also Definition 3.5). We prove that the process-level limit remains close to the system of ODEs for the fully flexible system, in terms of the ℓ_2-distance, if $\phi(G_N)$ and $\gamma(G_N)$ are suitably small and the system is started in a state that has its queues "*sufficiently spread out*" (Theorem 3.10). This in particular includes states with i.i.d. queue lengths. Most importantly, *the result in Theorem 3.10 is nonasymptotic*.

(*c*) *Mean-field approximation*. Leveraging Theorem 3.10 and the mixing time bound, we determine the applicability of the mean-field approximation for any compatibility graph in terms of the local properties $\phi(G_N)$ and $\gamma(G_N)$. In particular, in Theorem 3.1 we provide a *finite N guarantee that, for any graph* G_N, the ℓ_2-distance between the steady-state and the fixed point of a system of ODEs is bounded by

$$\frac{c}{(\ln(1/\max\{\phi(G_N)^2, \gamma(G_N)\}))^{\alpha}} \tag{2.4}$$

for constants $c, \alpha > 0$ that depend only on λ, ρ_0 and d. In particular, if $\max\{\phi(G_N), \gamma(G_N)\} \to 0$ as $N \to \infty$, then the distribution of $\boldsymbol{q}^N(t)$, in steady state, converges weakly to the Dirac delta distribution at the fixed point of the ODE corresponding to the fully flexible system.

(*d*) *Implications for specific graph classes*. To show that the conditions on the graph sequence are satisfied by common graphs, we consider two sequences of *sparse* graphs for which the condition $\max\{\phi(G_N), \gamma(G_N)\} \to 0$ as $N \to \infty$ is satisfied.

First, let $(G_N)_{N \geq 1}$ be a sequence of random bipartite geometric graphs. From a high level, these graphs are obtained by placing the dispatchers and the servers at uniformly random locations and connecting a dispatcher and server by an edge if they are at most a fixed distance $r(N) > 0$ apart; see Section 3.2 for a precise definition. Recall that d_v^N and d_w^N denote the degree of $v \in V_N$ and $w \in W_N$, respectively. We prove that, if $r(N)$ is such that

$\liminf_{N\to\infty} \mathbb{E}[d_v^N]/\ln N = \infty$ and $\liminf_{N\to\infty} \mathbb{E}[d_w^N]/\max(\ln M(N), \ln N) = \infty$, then indeed $\max\{\phi(G_N), \gamma(G_N)\} \to 0$, and $\boldsymbol{q}^N(t)$ in steady-state becomes asymptotically indistinguishable from the fully flexible system (Corollary 3.4). Note that these conditions still ensure sparsity in that the degree of a server is nearly a factor $M(N)/\ln N$ smaller as compared to the complete bipartite graph where the degree is $M(N)$.

Second, the above convergence holds in much more generality for a sequence of regular bipartite graphs. That is, d_v^N is the same for all v and d_w^N is the same for all w *within each connected component of the graph*; see Section 3.2 for a precise definition. We prove that the convergence holds whenever $\gamma(G_N) \to 0$, which happens if for example, if $\min_{w\in W_N} d_w^N$ diverges (at any rate) as $N \to \infty$ (Corollary 3.6), and thus ensures sparsity. This includes *arbitrary deterministic graph sequences* and thus significantly broadens the applicability of the mean-field approximation.

3. Main results. In the following, all graphs will refer to bipartite graphs $G_N = (V_N, W_N, E_N)$ as described in the beginning of Section 2. We let $X_v(t)$ denote the queue length of a server $v \in V_N$ at time t. Let $Q_i^N(t) := \sum_{v\in V_N} \mathbb{1}\{X_v(t) \geq i\}$ denote the number of servers with queue length at least $i \in \mathbb{N}$ in the entire system. We will refer to these as *global* quantities. The *local* number of servers with queue length at least $i \in \mathbb{N}$, as seen from the perspective of a task type $w \in W_N$, is denoted by $Q_i^{N,w}(t) := \sum_{v\in\mathcal{N}_w} \mathbb{1}\{X_v(t) \geq i\}$. Define their scaled versions as $q_i^N(t) := Q_i^N(t)/N$ and $q_i^{N,w}(t) := Q_i^{N,w}(t)/d_w^N$. Note that $\{X_v(t) : v \in V_N\}$ is a Markov process, and the vector $(q_i^N(\infty))_{i\geq 1}$ will denote the corresponding steady-state functional of this Markov process.

Throughout this paper, the assumption of an exponential service time distribution is crucial. Many stochastic coupling arguments explicitly leverage the memoryless property of the exponential distribution. However, it is important to note that large-scale asymptotic analyses of load balancing systems with general service time distributions are scarce in the literature, even when the compatibility graph is complete bipartite (see [1, 2, 7, 8]). The analysis of systems with general service time distributions under nontrivial compatibility constraints is left as an interesting future research direction.

3.1. *Steady-state approximation for arbitrary graphs.* The JSQ(d) policy is known for its drastic delay-performance improvement over random routing. It is well known that on a complete bipartite graph with full flexibility, the steady-state quantity $q_i^N(\infty)$ approaches $q_i^* := \lambda^{\frac{d^i-1}{d-1}}$ as $N \to \infty$ [27, 39]. This is often referred to as "the power of two effect", meaning that the tail of the queue length distribution decays double-exponentially (in contrast to just exponentially for random routing). Recall the definitions of $\phi(G_N)$ and $\gamma(G_N)$ from (2.3). *For an arbitrary compatibility graph* G_N, a central result of this paper provides a finite N bound on the expected ℓ_2-distance between $(q_i^N(\infty))_{i\geq 1}$ and $(q_i^*)_{i\geq 1}$:

THEOREM 3.1. *Given any* G_N, *if* $\rho(G_N) \leq \rho_0 < 1$, *then* $\boldsymbol{X}(t) = (X_v(t))_{v\in V_N}$ *is ergodic. Moreover, if* $\max\{\phi(G_N), \gamma(G_N)\} \leq 1$, *then there exist constants* $c, \alpha > 0$ (*depending only on* λ, ρ_0 *and* d) *such that*

$$\sum_{i=1}^{\infty} \mathbb{E}\big[\big(q_i^N(\infty) - q_i^*\big)^2\big] \leq \frac{c}{(\ln(1/\max\{\phi(G_N)^2, \gamma(G_N)\}))^{\alpha}}, \tag{3.1}$$

where $q_i^* = \lambda^{\frac{d^i-1}{d-1}}$ *for* $i \in \mathbb{N}$.

Theorem 3.1 is proved in Section 4.4. For large N asymptotics, it also provides a rate of convergence, although we do not expect this rate to be tight for specific sequences of graphs, as the result holds for arbitrary graphs. The following is an immediate corollary.

COROLLARY 3.2. *Let $(G_N)_{N\geq 1}$ be a sequence of graphs with $\rho(G_N) \leq \rho_0 < 1$ for all $N \geq 1$ and assume $\max\{\phi(G_N), \gamma(G_N)\} \to 0$ as $N \to \infty$. Then $\sum_{i=1}^{\infty} \mathbb{E}[(q_i^N(\infty) - q_i^*)^2] \to 0$ as $N \to \infty$.*

REMARK 2. It is worthwhile to note that Theorem 3.1 extends to a bound on any ℓ_p-distance for $0 < p < \infty$. This follows by bounding the tail sum using Corollary 4.3 and bounding the finite remainder using Hölder's inequality.

3.2. *Convergence for specific graph sequences.* Let us now discuss two important classes of graph sequences that satisfy the mean-field approximation conditions in Corollary 3.2. We begin with a popular generative model for spatial graphs.

DEFINITION 3.3 (Random bipartite geometric graph). We say that G_N is a random bipartite geometric graph if G_N is constructed as follows. Let $r(N) > 0$ be fixed and A be the unit k-dimensional torus for $k \in \mathbb{N}$. We assign each $v \in V_N$ and $w \in W_N$ a location $x_v \in A$ and $x_w \in A$, respectively, independently and uniformly at random. Next, $(v, w) \in E_N$ if and only if $\|x_v - x_w\|_p \leq r(N)$ for $1 < p < \infty$.

We define random geometric graphs on a k-dimensional torus to avoid boundary effects. From a practical perspective, however, the boundary effects become negligible as $N \to \infty$. For the following theorem, recall that d_v^N and d_w^N denote the degrees of $v \in V_N$ and $w \in W_N$ in G_N, respectively.

COROLLARY 3.4. *Let $(G_N)_{N\geq 1}$ be a sequence of random bipartite geometric graphs, where $r(N)$ is chosen such that*

$$\liminf_{N\to\infty} \frac{\mathbb{E}[d_v^N]}{\ln N} = \infty, \qquad \liminf_{N\to\infty} \frac{\mathbb{E}[d_w^N]}{\max(\ln M(N), \ln N)} = \infty. \tag{3.2}$$

Then, almost surely for any realization of the graph sequence $(G_N)_{N\geq 1}$, $\boldsymbol{X}(t) = (X_v(t))_{v\in V_N}$ is ergodic for all N large enough and $\sum_{i=1}^{\infty} \mathbb{E}[(q_i^N(\infty) - q_i^*)^2] \to 0$ as $N \to \infty$, where $q_i^* = \lambda^{\frac{d^i-1}{d-1}}$ for $i \in \mathbb{N}$.

REMARK 3. The reason that "for all N large enough" is added in Corollary 3.4 is that, for any fixed N, with (small but) positive probability, the random graph may not satisfy the stability criterion. As $N \to \infty$, this probability becomes small and using the Borel–Cantelli lemma, we show that the stability criterion is satisfied almost surely for all N large enough. To be precise, the convergence statement for $(q_i^N(\infty))_{i\geq 1}$ should be interpreted for all N large enough where the ergodicity holds.

The proof relies on verifying the conditions of Corollary 3.2 using concentration of measure arguments and is given in Section 4.5. Next, we consider sequences of regular graphs, in which case, we can allow much more general sequences of graphs.

DEFINITION 3.5 (Regular bipartite graph). We say that G_N is a regular bipartite graph if $(v, w) \in E_N$ implies that $Nd_v^N = M(N)d_w^N$.

Note that the definition of a regular bipartite graph implies that the degrees of all servers and all dispatchers are the same *within every connected component*, and it allows the graph to have many connected components. The next theorem proves the convergence of the steady state for any such regular bipartite graphs with diverging minimum dispatcher degree.

COROLLARY 3.6. *Let $(G_N)_{N\geq 1}$ be a sequence of regular bipartite graphs, where $\gamma(G_N) \to 0$ as $N \to \infty$. Then, the queue length process is ergodic for all $N \geq 1$ and $\sum_{i=1}^{\infty} \mathbb{E}[(q_i^N(\infty) - q_i^*)^2] \to 0$ as $N \to \infty$, where $q_i^* = \lambda^{\frac{d^i-1}{d-1}}$ for $i \in \mathbb{N}$. The convergence holds in particular if $\min_{w\in W_N} d_w^N \to \infty$ as $N \to \infty$.*

PROOF. The proof is immediate by observing that, due to the regularity of G_N, we have

$$\phi(G_N) := \max_{v\in V_N} \left| \frac{N}{M(N)} \sum_{w\in\mathcal{N}_v} \frac{1}{d_w^N} - 1 \right| = \max_{v\in V_N} \left| \frac{N}{M(N)} \sum_{w\in\mathcal{N}_v} \frac{M(N)}{N d_v^N} - 1 \right| = 0. \tag{3.3}$$

Note also that $\rho(G_N) \leq \lambda(1+\phi(G_N)) = \lambda < 1$. Therefore, Corollary 3.2 completes the proof of the first part. The second part is proved by observing that $\gamma(G_N) \leq 1/(\min_{w\in W_N} d_w^N)$. □

The rest of the contributions will be pivotal in the proof of Theorem 3.1.

3.3. *Large-scale mixing-time bound.* A crucial step in identifying the steady-state distribution is to show that the distribution of $\boldsymbol{q}^N(t)$ becomes close to its steady state within a large, but finite time. We prove that, in appropriate sense, the Markov process mixes in polynomial time, independent of N, from any state that is stochastically dominated by the steady-state. We use the following notion of stochastic ordering.

DEFINITION 3.7 (Stochastic ordering). For $n \in \mathbb{N}$, let $\boldsymbol{X} = (X_1, \ldots, X_n)$ and $\boldsymbol{Y} = (Y_1, \ldots, Y_n)$ be two n-dimensional random variables. We write $\boldsymbol{X} \leq_{st} \boldsymbol{Y}$ if there exists a common probability space where $X_i \leq Y_i$ for all $i = 1, \ldots, n$, almost surely.

To formalize the notion of large-scale mixing time, consider two copies of the system on the same graph G_N. For system k, with $k = 1, 2$, the queue length at server $v \in V_N$ is denoted by $X_v^{(k)}(t)$ and the fraction of servers with queue length at least $i \in \mathbb{N}$ is denoted by $q_i^{N,(k)}(t)$.

THEOREM 3.8. *Let G_N be a graph, $\rho(G_N) \leq \rho_0 < 1$ and $\boldsymbol{X}_0$ be a random variable on $\mathbb{N}^N$ such that $\boldsymbol{X}_0 \leq_{st} \boldsymbol{X}(\infty)$. Suppose $\boldsymbol{X}^{(1)}(0) \stackrel{d}{=} \boldsymbol{X}_0$ and $\boldsymbol{X}^{(2)}(0) \stackrel{d}{=} \boldsymbol{X}(\infty)$. Then there exist a joint probability space and constants $c_1, c_2 > 0$, $0 < \alpha \leq 1$ (depending only on ρ_0 and d) such that, for all $t \geq 0$,*

$$\sum_{i=1}^{\infty} \mathbb{E}\big[|q_i^{N,(2)}(t) - q_i^{N,(1)}(t)|\big] \leq \frac{1}{(c_1 + c_2 t)^{\alpha}}. \tag{3.4}$$

The proof is given in Section 4.2 and relies on the fact that the stochastic ordering is maintained throughout for all $t \geq 0$, as shown by the following proposition.

PROPOSITION 3.9. *Under the conditions of Theorem 3.8, there exists a joint probability space such that $X_v^{(1)}(t) \leq X_v^{(2)}(t)$ for all $v \in V_N$ and $t \geq 0$, almost surely, along any sample path.*

The proposition is proved in Section 4.2. The proof follows by an induction argument, where we show that the inequality is maintained for each arrival and departure epoch. At an arrival epoch, we use a monotonicity property of the JSQ(d) policy: if we sample the same d servers in both systems, then the task is routed to a server with a higher queue length in system 2 than in system 1. We relate this behavior to a property of the probabilistic assignment function of JSQ(d) (Lemma 4.4). The proposition generalizes to any assignment policy which satisfies such a monotonicity property.

3.4. *Process-level limit starting from i.i.d. queue lengths.* As our quantity of interest $\boldsymbol{q}^N(t)$ becomes arbitrarily close to the steady-state in finite time, it is sufficient to characterize the transient behavior of one sample path of the system. We prove that $\boldsymbol{q}^N(t)$ remains close to a system of ODEs if $\phi(G_N)$ and $\gamma(G_N)$ are small (recall (2.3) for their definition) and the queues in the starting state are "sufficiently spread out".

THEOREM 3.10. *Let G_N be a graph, $\rho(G_N) \le \rho_0$, and $\bar{\boldsymbol{q}}(t) = (\bar{q}_i(t))_{i\ge 1}$ be the unique solution to the system of ODEs*

$$\frac{d\bar{q}_i(t)}{dt} = \lambda\big(\bar{q}_{i-1}(t)^d - \bar{q}_i(t)^d\big) - \big(\bar{q}_i(t) - \bar{q}_{i+1}(t)\big) \quad \text{for } i \in \mathbb{N}. \tag{3.5}$$

Then, there exists a constant $c \ge 1$ (depending only on ρ_0 and d) such that, for all $t \ge 0$,

$$\begin{aligned}
&\mathbb{E}\left[\sup_{s\in[0,t]} \sum_{i=1}^{\infty} \big(q_i^N(s) - \bar{q}_i(s)\big)^2\right] \\
&\quad \le 2\phi(G_N)^2\left(\lambda t + \mathbb{E}\left[\sum_{i=1}^{\infty} q_i^N(0)^2\right]\right) \\
&\quad\quad + 12e^{ct^2}\left(t^2d^2\phi(G_N)^2 + \mathbb{E}\left[\sum_{i=1}^{\infty}\left(\frac{1}{M(N)}\sum_{w\in W_N} |q_i^{N,w}(0) - \bar{q}_i(0)|\right)^2\right]\right. \\
&\quad\quad \left. + 4t(\rho_0 d + 1)\gamma(G_N)\right).
\end{aligned} \tag{3.6}$$

Note that Theorem 3.10 is also a nonasymptotic result. An immediate corollary is the following.

COROLLARY 3.11. *Let $(G_N)_{N\ge 1}$ be a sequence of graphs, $\rho(G_N) \le \rho_0$ and $\max\{\phi(G_N), \gamma(G_N)\} \to 0$ as $N \to \infty$. Also, assume that $\boldsymbol{X}(0) = (X_v(0))_{v\in V_N}$ are i.i.d. and $\mathbb{P}(X_v(0) \ge i) = \bar{q}_i(0)$ where $\sum_{i=1}^{\infty} \bar{q}_i(0) < \infty$. Then, for any $t \ge 0$,*

$$\lim_{N\to\infty} \mathbb{E}\left[\sup_{s\in[0,t]} \sum_{i=1}^{\infty} \big(q_i^N(s) - \bar{q}_i(s)\big)^2\right] = 0, \tag{3.7}$$

where $(\bar{q}_i(t))_{i\ge 1}$ is as defined in (3.5).

The "sufficiently spread out" condition in Theorem 3.10 is imposed by the initial state quantity $\sum_{i=1}^{\infty}(\frac{1}{M(N)}\sum_{w\in W_N}|q_i^{N,w}(0) - \bar{q}_i(0)|)^2$. This term is small if $\boldsymbol{q}^{N,w}(0) \approx \bar{\boldsymbol{q}}(0)$ for most $w \in W_N$ and, hence, if the local queue length distribution from the perspective of each task type is approximately equal. In particular, this term is at most $\gamma(G_N)\sum_{i=1}^{\infty}\bar{q}_i(0)$ for i.i.d. queue lengths. Theorem 3.10 is proved in Section 4.3. The proof relies on tracking a sequence of martingales for each $w \in W_N$ and bounding the ℓ_2-distance to the ODE by their quadratic variation and quantities such as $\phi(G_N)$ and $\gamma(G_N)$ using Grönwall's inequality. In the proof, the quantity $\phi(G_N)$ is used in (4.34) and (4.42) and $\gamma(G_N)$ is used in (4.43).

REMARK 4. One should contrast Theorem 3.10 and Corollary 3.11 with the process-level limit result proved in Budhiraja et al. [9]. In this paper, the authors considered an undirected version of the model in the current paper. The model, as is, is not suitable for capturing the task-server compatibility constraints. An undirected graph would mean that if server i can process task type j, then server j must be able to process task type i. However, a generalization of the model in [9] to directed graphs can be viewed as a special case of our model:

when $M(N) = N$ and there is a perfect matching between the set of servers and the set of dispatchers (equivalently, a dedicated arrival stream per server). Although the undirected graph assumption is not crucial in [9], the $M(N) = N$ assumption plays a major role for the approach to work. In the current paper, $M(N)$ can grow at any rate (sub-/super-linearly) with N. As a result of the above structural differences, the queue length process in [9] is ergodic for *any* graph, whereas in our model, this is nontrivial.

Moreover, [9] establishes the process-level convergence if the initial queue lengths at the servers are i.i.d. from some distribution as in Corollary 3.11. The idea there is that, if the system starts from a state where the queue lengths at the servers are i.i.d., then any two queue lengths retain their stochastic independence on any finite time interval, asymptotically as $N \to \infty$. Consequently, the N-dimensional queue length vector can be coupled with an infinite-dimensional McKean–Vlasov process where any finite collection of coordinates are independent on any finite time interval. The assumption that the queue lengths are i.i.d. at time zero is crucial for this approach to go through. As a result, and as already remarked in the conclusion of [9], it is unclear how to prove convergence of the steady state. In addition to our main contribution on the convergence of steady states, Theorem 3.10 generalizes the process-level limit beyond the i.i.d. case. It identifies a structural condition on the initial state that ensures the same process-level limit of as the fully flexible system.

4. Proofs. Most of the results in this section are nonasymptotic and hold for any fixed N. Thus, throughout this section, we drop the dependence on N in the notation where possible, for the sake of brevity.

4.1. *Existence of steady-state and moment bound.* We first prove that the Markov process is positive recurrent and has a unique steady-state.

PROPOSITION 4.1. *If $\rho(G_N) \le \rho_0 < 1$, then the Markov process $\boldsymbol{X}(t) = (X_v(t))_{v \in V}$ is positive recurrent and there exists a unique steady-state of the process denoted as $\boldsymbol{X}(\infty)$.*

The proof relies on a Lyapunov argument. Let $V(t) := \sum_{i=1}^{\infty} \sum_{j=i}^{\infty} Q_j(t)$ be the Lyapunov function. If we show that the drift of $V(t)$ is strictly negative anywhere outside of a suitably chosen finite set of states, then this is sufficient for positive recurrence. As such, we compute the drift.

LEMMA 4.2. *Fix any $i \in \mathbb{N}$ and $t \ge 0$. Then,*

$$\frac{d}{dt}\mathbb{E}\left[\sum_{j=i}^{\infty} Q_j(t)\right] = \mathbb{E}\left[\frac{\lambda N}{M} \sum_{w \in W} q_{i-1}^{w}(t)^d - Q_i(t)\right]. \tag{4.1}$$

PROOF. To change the value of $\sum_{j=i}^{\infty} Q_j(t)$, a task must arrive to a server with queue length at least $i-1$ or a task must depart a server with queue length at least i.

Let us compute the probability that a task is assigned to a server with queue length at least $i-1$. At the epoch time of an arrival, a task adopts a type $w \in W$ uniformly at random. The task is routed to a server with queue length at least $i-1$ if and only if the system only samples servers with queue length at least $i-1$, which happens with probability $q_{i-1}^{w}(t-)^d$. This results in a probability of $\frac{1}{M}\sum_{w \in W} q_{i-1}^{w}(t-)^d$ to be routed to a server with queue length at least $i-1$.

Now, we compute the probability that a task departs a server with queue length at least i. At the epoch time of a potential departure, a server $v \in V$ is chosen uniformly at random.

A task departs a server with queue length at least i if and only if v has queue length at least i. This results in a probability of $q_i(t-)$ to depart a server with queue length at least i.

We describe the arrival and departure process as follows. Let $\mathcal{N}(t)$ be a Poisson process of rate $(\lambda+1)N$. An event of the process is either an arrival of type $w \in W$ with probability $\lambda/((\lambda+1)M)$ or a potential departure at server $v \in V$ with probability $1/((\lambda+1)N)$, independent of the past. Note that this is equivalent to the model description introduced before. Hence, for any $h > 0$,

$$\begin{aligned}
&\mathbb{E}\Bigg[\sum_{j=i}^{\infty} \Delta Q_j(t)\Big|\mathcal{F}_t\Bigg] \\
&\quad= \mathbb{E}\Bigg[\sum_{j=i}^{\infty} \Delta Q_j(t)\Big|\Delta\mathcal{N}(t)=1,\mathcal{F}_t\Bigg]\mathbb{P}\big(\Delta\mathcal{N}(t)=1\big) \\
&\quad\quad \pm \mathbb{E}\big[\Delta\mathcal{N}(t)|\Delta\mathcal{N}(t)\geq 2\big]\mathbb{P}\big(\Delta\mathcal{N}(t)\geq 2\big) \\
&\quad= \Bigg(\frac{\lambda}{\lambda+1}\frac{1}{M}\sum_{w\in W} q_{i-1}^{w}(t)^d - \frac{1}{\lambda+1}q_i(t)\Bigg)(\lambda+1)Nhe^{-(\lambda+1)Nh} \\
&\quad\quad \pm \big((\lambda+1)Nh+2\big)\big((\lambda+1)Nh\big)^2,
\end{aligned} \tag{4.2}$$

where $\Delta Q_j(t) := Q_j(t+h) - Q_j(t)$ and $\Delta\mathcal{N}(t) := \mathcal{N}(t+h) - \mathcal{N}(t)$. Here, we use the shorthand notation $\pm x$ to denote a term in $[-x, x]$. The equation above implies

$$\begin{aligned}
\frac{d}{dt}\mathbb{E}\Bigg[\sum_{j=i}^{\infty} Q_j(t)\Bigg] &= \lim_{h\downarrow 0}\frac{\mathbb{E}[\mathbb{E}[\sum_{j=i}^{\infty}(Q_j(t+h)-Q_j(t))|\mathcal{F}_t]]}{h} \\
&= \mathbb{E}\Bigg[\frac{\lambda N}{M}\sum_{w\in W} q_{i-1}^{w}(t)^d - Q_i(t)\Bigg],
\end{aligned} \tag{4.3}$$

which completes the proof of the lemma. □

PROOF OF PROPOSITION 4.1. Note that

$$\begin{aligned}
\frac{\lambda N}{M}\sum_{w\in W} q_i^w(t) &= \frac{\lambda N}{M}\sum_{w\in W}\sum_{\substack{v\in\mathcal{N}_w \\ X_v(t)\geq i}}\frac{1}{d_w} \\
&= \sum_{\substack{v\in V \\ X_v(t)\geq i}}\frac{\lambda N}{M}\sum_{w\in\mathcal{N}_v}\frac{1}{d_w} \leq \sum_{\substack{v\in V \\ X_v(t)\geq i}}\rho_0 = \rho_0 Q_i(t).
\end{aligned} \tag{4.4}$$

Let $V(t) := \sum_{i=1}^{\infty}\sum_{j=i}^{\infty} Q_j(t)$ and $\boldsymbol{X}(0) = \boldsymbol{x} \in \mathbb{N}^V$. Then,

$$\begin{aligned}
\frac{d}{dt}\mathbb{E}\big[V(t)\big] &= \frac{d}{dt}\sum_{i=1}^{\infty}\mathbb{E}\Bigg[\sum_{j=i}^{\infty} Q_j(t)\Bigg] = \sum_{i=1}^{\infty}\frac{d}{dt}\mathbb{E}\Bigg[\sum_{j=i}^{\infty} Q_j(t)\Bigg] \\
&= \sum_{i=1}^{\infty}\mathbb{E}\Bigg[\frac{\lambda N}{M}\sum_{w\in W} q_{i-1}^{w}(t)^d - Q_i(t)\Bigg] \leq \sum_{i=1}^{\infty}\mathbb{E}\Bigg[\frac{\lambda N}{M}\sum_{w\in W} q_{i-1}^{w}(t) - Q_i(t)\Bigg] \\
&\leq \sum_{i=1}^{\infty}\mathbb{E}\big[\rho_0 Q_{i-1}(t) - Q_i(t)\big] = -(1-\rho_0)\sum_{i=1}^{\infty}\mathbb{E}\big[Q_i(t)\big] + \rho_0 N,
\end{aligned} \tag{4.5}$$

where we use the monotone convergence theorem in the first equality, Lemma 4.2 in the third equality and (4.4) in the second inequality. Note that we exchange derivative and sum in the

second equality. To check whether this is allowed, we verify the conditions of Theorem 7.17 in [34]. Let $f_n(t) := \sum_{i=1}^n \mathbb{E}[\sum_{j=i}^\infty Q_j(t)]$ for $n \in \mathbb{N}$ and notice that $f_n(t)$ is differentiable on $[0,t]$ as a result of Lemma 4.2. Also, $f_n(0)$ converges since

$$f_n(0) = \sum_{i=1}^n \mathbb{E}\Bigg[\sum_{j=i}^\infty Q_j(0)\Bigg] = \sum_{i=1}^n \sum_{v\in V} (x_v - i + 1)^+ \to \sum_{i=1}^\infty \sum_{v\in V} (x_v - i + 1)^+, \tag{4.6}$$

as $n \to \infty$ due to monotone convergence theorem. Now,

$$\frac{d}{dt} f_n(t) = \sum_{i=1}^n \frac{d}{dt}\mathbb{E}\Bigg[\sum_{j=i}^\infty Q_j(t)\Bigg] = \sum_{i=1}^n \mathbb{E}\Bigg[\frac{\lambda N}{M} \sum_{w\in W} q_{i-1}^w(t)^d - Q_i(t)\Bigg], \tag{4.7}$$

by Lemma 4.2. We now check whether $\frac{d}{dt} f_n(t)$ converges uniformly on $[0,t]$. The sum $\sum_{i=1}^n \mathbb{E}[\frac{\lambda N}{M} \sum_{w\in W} q_{i-1}^w(t)^d]$ converges uniformly on $[0,t]$ since

$$\begin{aligned}
&\sup_{s\in[0,t]} \Bigg| \sum_{i=1}^\infty \mathbb{E}\Bigg[\frac{\lambda N}{M} \sum_{w\in W} q_{i-1}^w(s)^d\Bigg] - \sum_{i=1}^n \mathbb{E}\Bigg[\frac{\lambda N}{M} \sum_{w\in W} q_{i-1}^w(s)^d\Bigg] \Bigg| \\
&\quad = \sup_{s\in[0,t]} \sum_{i=n+1}^\infty \mathbb{E}\Bigg[\frac{\lambda N}{M} \sum_{w\in W} q_{i-1}^w(s)^d\Bigg] \le \sup_{s\in[0,t]} \sum_{i=n+1}^\infty \mathbb{E}\Bigg[\frac{\lambda N}{M} \sum_{w\in W} q_{i-1}^w(s)\Bigg] \\
&\quad \le \sup_{s\in[0,t]} \sum_{i=n+1}^\infty \mathbb{E}\big[\rho_0 Q_{i-1}(t)\big] \le \sup_{s\in[0,t]} \rho_0 N \mathbb{E}\Big[\Big(\max_{v\in V} X_v(s) - n + 1\Big)^+\Big] \\
&\quad \le \rho_0 N \mathbb{E}\Big[\Big(\max_{v\in V} x_v + \mathcal{N}_a(t) - n + 1\Big)^+\Big] \to 0,
\end{aligned} \tag{4.8}$$

as $n \to \infty$, where we use (4.4) in the second inequality and $\mathcal{N}_a$ is a Poisson process of rate λN. Note that the sum $\sum_{i=1}^n \mathbb{E}[Q_i(t)]$ converges uniformly on $[0,t]$ too, which follows from the second half of (4.8). Therefore, $\frac{d}{dt} f_n(t)$ converges uniformly on $[0,t]$. Then, by Theorem 7.17 in [34], we conclude that we may interchange derivative and sum in (4.5).

Let $S := \{\boldsymbol{x} \in \mathbb{N}^V : \sum_{i=1}^\infty \sum_{v\in V} \mathbb{1}\{x_v \ge i\} \le N/(1-\rho_0)\}$ and note that S is finite. Then,

$$\frac{d}{dt}\mathbb{E}\big[V(t)\big]\Big|_{t=0} \le -(1-\rho_0) \sum_{i=1}^\infty Q_i(0) + \rho_0 N \le -(1-\rho_0)N + N\mathbb{1}\{\boldsymbol{x} \in S\}. \tag{4.9}$$

Hence, by Theorem 4.2 in [26], the Markov process $\boldsymbol{X}(t)$ is positive recurrent and there exists a unique steady-state of the process denoted as $\boldsymbol{X}(\infty)$. □

As a consequence, a similar Lyapunov argument shows a moment bound on the steady-state.

COROLLARY 4.3. *If $\rho(G_N) \le \rho_0 < 1$, then $\mathbb{E}[q_i(\infty)] \le \rho_0^i$ for all $i \in \mathbb{N}$.*

PROOF. Fix any $i \in \mathbb{N}$ and $t \ge 0$. We let $\boldsymbol{X}(0) \overset{\mathrm{d}}{=} \boldsymbol{X}(\infty)$ such that $\boldsymbol{X}(t) \overset{\mathrm{d}}{=} \boldsymbol{X}(\infty)$. Then,

$$\begin{aligned}
0 &= \frac{d}{dt}\mathbb{E}\Bigg[\sum_{j=i}^\infty Q_j(t)\Bigg] = \mathbb{E}\Bigg[\frac{\lambda N}{M} \sum_{w\in W} q_{i-1}^w(t)^d - Q_i(t)\Bigg] \\
&\le \mathbb{E}\Bigg[\frac{\lambda N}{M} \sum_{w\in W} q_{i-1}^w(t) - Q_i(t)\Bigg] \le \mathbb{E}\big[\rho_0 Q_{i-1}(t) - Q_i(t)\big],
\end{aligned} \tag{4.10}$$

where we use Lemma 4.2 in the second equality and (4.4) in the second inequality. Hence, by induction, $\mathbb{E}[q_i(t)] \le \rho_0 \mathbb{E}[q_{i-1}(t)] \le \rho_0^i$, which completes the proof of the lemma. □

4.2. *Proof of the large-scale mixing-time bound.* The proofs of this section relies on the following stochastic ordering property of the load balancing policy as stated in Lemma 4.4. The lemma is proved in Appendix A. We prove this for the JSQ(d) policy. However, we believe that it is possible to generalize the mixing time bound to a large class of load balancing policies which satisfy an analogous monotonicity property.

LEMMA 4.4. *Fix any* $0 \le y_1 < x_1 \le 1$ *and* $0 \le y_2 < x_2 \le 1$ *such that* $x_1 \le x_2$ *and* $y_1 \le y_2$. *Then*,

$$\frac{x_1^d - y_1^d}{x_1 - y_1} \le \frac{x_2^d - y_2^d}{x_2 - y_2}. \tag{4.11}$$

PROOF OF PROPOSITION 3.9. We couple the arrival and potential departure epochs of the two systems such that any arrival of a task type $w \in W$ and any potential departure at a server $v \in V$ happen at the same time in both systems. We proceed to design a stochastic coupling that maintains the inequality $X_v^{(1)}(t) \le X_v^{(2)}(t)$ for all $v \in V$ on every arrival and potential departure epoch. At time $t = 0$, the inequality is maintained by the stochastic ordering assumption and by defining $\boldsymbol{X}^{(1)}(0)$ and $\boldsymbol{X}^{(2)}(0)$ on the suitable probability space.

Let $t \ge 0$ be a potential departure epoch at server $v \in V$ and assume that $X_{v'}^{(1)}(t-) \le X_{v'}^{(2)}(t-)$ for all $v' \in V$. Clearly, $X_v^{(1)}(t) \le X_v^{(2)}(t)$ also after the departure.

Now, let $t \ge 0$ be an arrival epoch of a task type $w \in W$ and assume that $X_{v'}^{(1)}(t-) \le X_{v'}^{(2)}(t-)$ for all $v' \in V$. Fix any $v \in \mathcal{N}_w$ with $i := X_v(t-)$ and let us compute the probability that the task is assigned to v. The task is routed to a server with queue length i if and only if the system only samples servers with queue length at least i and not only servers with queue length at least $i+1$, which happens with probability $q_i^w(t-)^d - q_{i+1}^w(t-)^d$. By symmetry, any server in $\mathcal{N}_w$ with queue length i has the same probability of receiving the task and there are a total of $Q_i^w(t-) - Q_{i+1}^w(t-)$ of such eligible servers. This results in a probability of

$$p_v^{(k)} := \frac{q^{(k),w}_{X_v^{(k)}(t-)}(t-)^d - q^{(k),w}_{X_v^{(k)}(t-)+1}(t-)^d}{Q^{(k),w}_{X_v^{(k)}(t-)}(t-) - Q^{(k),w}_{X_v^{(k)}(t-)+1}(t-)}, \tag{4.12}$$

of assigning the task to a server $v \in \mathcal{N}_w$ in system $k = 1, 2$. Let $\hat{p}_v := \min(p_v^{(1)}, p_v^{(2)})$ be the shared probability mass. For the sake of notation, assume that the servers in $\mathcal{N}_w$ are ordered and correspond to the integers $\{1, 2, \dots, d_w\}$. Let $U_t \in [0, 1]$ be a uniform random variable, independent of any other processes and independent across arrival epochs, and which is shared between the two systems. Then, in system k, assign the task to server $v \in \mathcal{N}_w$ if and only if

$$\begin{aligned} U_t \in &\left[\sum_{v'=1}^{v-1} \hat{p}_{v'}, \sum_{v'=1}^{v} \hat{p}_{v'}\right) \\ &\cup \left[\sum_{v' \in \mathcal{N}_w} \hat{p}_{v'} + \sum_{v'=1}^{v-1} \big(p_{v'}^{(k)} - \hat{p}_{v'}\big), \sum_{v' \in \mathcal{N}_w} \hat{p}_{v'} + \sum_{v'=1}^{v} \big(p_{v'}^{(k)} - \hat{p}_{v'}\big)\right). \end{aligned} \tag{4.13}$$

Note that the probability to assign to a server $v \in \mathcal{N}_w$ in system k is exactly equal to $p_v^{(k)}$.

To verify that the stochastic coupling maintains the ordering of queue lengths, note that if $U_t < \sum_{v' \in \mathcal{N}_w} \hat{p}_{v'}$, then the task is routed to the same server in both systems by the construction above. Thus, in this case, $X_{v'}^{(1)}(t) \le X_{v'}^{(2)}(t)$ for all $v' \in V$ also after the arrival.

Next, consider instead that $U_t \geq \sum_{v' \in \mathcal{N}_w} \hat{p}_{v'}$. Then, the task is routed to two different servers in both systems. Let $v \in \mathcal{N}_w$ be the server the task is routed to in system 1. Note that it does not matter to which server the task is routed to in system 2, since its queue length will only increase. By the construction above, it must hold that $p_v^{(1)} > \hat{p}_v = p_v^{(2)}$. We claim that this implies that $X_v^{(1)}(t-) < X_v^{(2)}(t-)$. To see why, suppose that $X_v^{(1)}(t-) = X_v^{(2)}(t-)$ instead. Note that

$$Q_i^{(1),w}(t-) = \sum_{v' \in \mathcal{N}_w} \mathbb{1}\{X_{v'}^{(1)}(t-) \geq i\} \leq \sum_{v' \in \mathcal{N}_w} \mathbb{1}\{X_{v'}^{(2)}(t-) \geq i\} = Q_i^{(2),w}(t-), \tag{4.14}$$

for all $i \in \mathbb{N}$ and $w \in W$ since $X_{v'}^{(1)}(t-) \leq X_{v'}^{(2)}(t-)$ for all $v' \in V$. Then, by Lemma 4.4,

$$\begin{aligned} p_v^{(1)} &= \frac{1}{d_w} \frac{q^{(1),w}_{X_v^{(1)}(t-)}(t-)^d - q^{(1),w}_{X_v^{(1)}(t-)+1}(t-)^d}{q^{(1),w}_{X_v^{(1)}(t-)}(t-) - q^{(1),w}_{X_v^{(1)}(t-)+1}(t-)} \\ &\leq \frac{1}{d_w} \frac{q^{(2),w}_{X_v^{(2)}(t-)}(t-)^d - q^{(2),w}_{X_v^{(2)}(t-)+1}(t-)^d}{q^{(2),w}_{X_v^{(2)}(t-)}(t-) - q^{(2),w}_{X_v^{(2)}(t-)+1}(t-)} = p_v^{(2)}, \end{aligned} \tag{4.15}$$

which is a contradiction. Hence, it must be that $X_v^{(1)}(t-) < X_v^{(2)}(t-)$ and therefore $X_{v'}^{(1)}(t) \leq X_{v'}^{(2)}(t)$ for all $v' \in V$ also after the arrival, which completes the proof of the proposition. □

PROOF OF THEOREM 3.8. We couple the two copies of the Markov process according to Proposition 3.9 such that $X_v^{(1)}(t) \leq X_v^{(2)}(t)$ for all $v \in V$ and $t \geq 0$, almost surely. This implies that $q_i^{(2),w}(t) \geq q_i^{(1),w}(t)$ for all $w \in W$ and $q_i^{(2)}(t) \geq q_i^{(1)}(t)$ for all $i \in \mathbb{N}$ and $t \geq 0$ by (4.14). Throughout, we will denote $\Delta_i(t) := q_i^{(2)}(t) - q_i^{(1)}(t)$. Let $\theta := \min(1/(2\rho_0 d), \rho_0)$ and define $V(t) := \sum_{i=1}^{\infty} \theta^i \sum_{j=i}^{\infty} \Delta_j(t)$. Then,

$$\begin{aligned} \frac{d}{dt}\mathbb{E}[V(t)] &= \frac{d}{dt} \sum_{i=1}^{\infty} \theta^i \mathbb{E}\left[\sum_{j=i}^{\infty} \Delta_j(t)\right] = \sum_{i=1}^{\infty} \theta^i \frac{d}{dt}\mathbb{E}\left[\sum_{j=i}^{\infty} \Delta_j(t)\right] \\ &= \sum_{i=1}^{\infty} \theta^i \mathbb{E}\left[\frac{\lambda}{M} \sum_{w \in W} \left(q_{i-1}^{(2),w}(t)^d - q_{i-1}^{(1),w}(t)^d\right) - \Delta_i(t)\right] \\ &\leq \sum_{i=1}^{\infty} \theta^i \mathbb{E}\left[\frac{\lambda d}{M} \sum_{w \in W} \left(q_{i-1}^{(2),w}(t) - q_{i-1}^{(1),w}(t)\right) - \Delta_i(t)\right] \\ &\leq \sum_{i=1}^{\infty} \theta^i (\theta \rho_0 d - 1)\mathbb{E}[\Delta_i(t)] \leq -\frac{1}{2} \sum_{i=1}^{\infty} \theta^i \mathbb{E}[\Delta_i(t)], \end{aligned} \tag{4.16}$$

where we use the monotone convergence theorem in the first equality, Lemma 4.2 in the third equality and the definition of θ in the third inequality. Note that we exchange derivative and sum in the second equality. To check whether this is allowed, we verify the conditions of Theorem 7.17 in [34]. Let $g_n(t) := \sum_{i=1}^{n} \theta^i \mathbb{E}[\sum_{j=i}^{\infty} \Delta_j(t)]$ for $n \in \mathbb{N}$ and notice that $g_n(t)$ is

differentiable on $[0,t]$ as a result of Lemma 4.2. Also, $g_n(0)$ converges since

$$\begin{aligned} g_n(0) &= \sum_{i=1}^{n} \theta^i \mathbb{E}\left[\sum_{j=i}^{\infty} \Delta_j(0)\right] \\ &= \sum_{i=1}^{n} \frac{\theta^i}{N} \sum_{v\in V} \big((X_v^{(2)}(0) - i + 1)^+ - (X_v^{(1)}(0) - i + 1)^+\big) \\ &\to \sum_{i=1}^{\infty} \frac{\theta^i}{N} \sum_{v\in V} \big((X_v^{(2)}(0) - i + 1)^+ - (X_v^{(1)}(0) - i + 1)^+\big), \end{aligned} \tag{4.17}$$

as $n \to \infty$. Now,

$$\begin{aligned} \frac{d}{dt} g_n(t) &= \sum_{i=1}^{n} \theta^i \frac{d}{dt} \mathbb{E}\left[\sum_{j=i}^{\infty} \Delta_j(t)\right] \\ &= \sum_{i=1}^{n} \theta^i \mathbb{E}\left[\frac{\lambda}{M} \sum_{w\in W} \big(q_{i-1}^{(2),w}(t)^d - q_{i-1}^{(1),w}(t)^d\big) - \frac{1}{N}\big(Q_i^{(2)}(t) - Q_i^{(1)}(t)\big)\right], \end{aligned} \tag{4.18}$$

by Lemma 4.2. The sums $\sum_{i=1}^{n} \theta^i \mathbb{E}[\frac{\lambda}{M} \sum_{w\in W} q_{i-1}^{(1),w}(t)^d]$, $\sum_{i=1}^{n} \theta^i \mathbb{E}[\frac{\lambda}{M} \sum_{w\in W} q_{i-1}^{(2),w}(t)^d]$, $\sum_{i=1}^{n} \frac{\theta^i}{N} \mathbb{E}[Q_i^{(1)}(t)]$ and $\sum_{i=1}^{n} \frac{\theta^i}{N} \mathbb{E}[Q_i^{(2)}(t)]$ each converge uniformly on $[0,t]$, which follows along the same lines as (4.8). Therefore, $\frac{d}{dt} g_n(t)$ converges uniformly on $[0,t]$. Then, by Theorem 7.17 in [34], we conclude that we may interchange derivative and sum in (4.16). The first inequality in (4.16) follows because, by the mean value theorem,

$$x^d - y^d = d\xi^{d-1}(x-y) \le d(x-y), \tag{4.19}$$

for all $1 \ge x \ge y \ge 0$ where $\xi \in [x,y]$. The second inequality in (4.16) follows because

$$\begin{aligned} \frac{\lambda N}{M} \sum_{w\in W} \big(q_i^{(2),w}(t) - q_i^{(1),w}(t)\big) &= \frac{\lambda N}{M} \sum_{w\in W} \sum_{\substack{v\in \mathcal{N}_w : X_v^{(2)}(t)\ge i \\ X_v^{(1)}(t)\le i-1}} \frac{1}{d_w} \\ &= \sum_{\substack{v\in V : X_v^{(2)}(t)\ge i \\ X_v^{(1)}(t)\le i-1}} \frac{\lambda N}{M} \sum_{w\in \mathcal{N}_v} \frac{1}{d_w} \\ &\le \sum_{\substack{v\in V : X_v^{(2)}(t)\ge i \\ X_v^{(1)}(t)\le i-1}} \rho_0 = \rho_0\big(Q_i^{(2)}(t) - Q_i^{(1)}(t)\big). \end{aligned} \tag{4.20}$$

Next, we find a lower bound on $\sum_{i=1}^{\infty} \theta^i \mathbb{E}[\Delta_i(t)]$ in terms of $\mathbb{E}[V(t)]$. Note that

$$\mathbb{E}\big[V(t)\big] = \mathbb{E}\left[\sum_{i=1}^{\infty}\sum_{j=i}^{\infty} \theta^i \Delta_j(t)\right] = \mathbb{E}\left[\sum_{j=1}^{\infty}\sum_{i=1}^{j} \theta^i \Delta_j(t)\right] = \mathbb{E}\left[\sum_{i=1}^{\infty} \frac{\theta(1-\theta^i)}{1-\theta} \Delta_i(t)\right], \tag{4.21}$$

and hence, again by the monotone convergence theorem,

$$\theta \sum_{i=1}^{\infty} \mathbb{E}\big[\Delta_i(t)\big] \le \mathbb{E}\big[V(t)\big] \le \frac{\theta}{1-\theta} \sum_{i=1}^{\infty} \mathbb{E}\big[\Delta_i(t)\big]. \tag{4.22}$$

Therefore, to find a lower bound on $\sum_{i=1}^{\infty} \theta^i \mathbb{E}[\Delta_i(t)]$ in terms of $\mathbb{E}[V(t)]$, it is sufficient to find a lower bound in terms of $\sum_{i=1}^{\infty} \mathbb{E}[\Delta_i(t)]$. Let $\eta := \sum_{i=1}^{\infty} \mathbb{E}[\Delta_i(t)]$. Note that $\mathbb{E}[\Delta_i(t)] \le$

$\mathbb{E}[q_i^{(2)}(t)] = \mathbb{E}[q_i^{(2)}(\infty)] \le \rho_0^i$ by Corollary 4.3. Thus, a lower bound on $\sum_{i=1}^{\infty} \theta^i \mathbb{E}[\Delta_i(t)]$ is given by the primal and dual pair

$$
\begin{aligned}
(P)\ \min_{\boldsymbol{x}} \quad & \sum_{i=1}^{\infty} \theta^i x_i \qquad (D)\ \max_{z,\boldsymbol{y}} \quad \eta z - \sum_{i=1}^{\infty} \rho_0^i y_i \\
\text{s.t.} \quad & \sum_{i=1}^{\infty} x_i = \eta \qquad \text{s.t.} \quad z - y_i \le \theta^i \\
& 0 \le x_i \le \rho_0^i \qquad\qquad y_i \ge 0.
\end{aligned} \tag{4.23}
$$

Fix any $i_0 \in \mathbb{N}$. A feasible solution to the dual is $y_i = 0$ for $i < i_0$, $y_i = \theta^{i_0} - \theta^i$ for $i \ge i_0$, and $z = \theta^{i_0}$. As any dual solution provides a lower bound to any primal solution by weak duality, it follows that

$$
\sum_{i=1}^{\infty} \theta^i \mathbb{E}\big[\Delta_i(t)\big] \ge \Big(\eta - \sum_{i=i_0}^{\infty} \rho_0^i\Big)\theta^{i_0} + \sum_{i=i_0}^{\infty} \rho_0^i \theta^i = \Big(\eta - \frac{\rho_0^{i_0}}{1-\rho_0}\Big)\theta^{i_0} + \frac{\rho_0^{i_0}\theta^{i_0}}{1-\rho_0\theta}. \tag{4.24}
$$

Now, let $i_0 := \lceil \ln((1-\rho_0)\eta)/\ln(\rho_0) \rceil$. Note that $i_0 \in \mathbb{N}$ since $\eta \le \rho_0/(1-\rho_0)$ and $\rho_0 < 1$. Then,

$$
\begin{aligned}
\sum_{i=1}^{\infty} \theta^i \mathbb{E}\big[\Delta_i(t)\big] &\ge \Big(\eta - \frac{\rho_0^{\ln((1-\rho_0)\eta)/\ln(\rho_0)}}{1-\rho_0}\Big)\theta^{i_0} + \frac{\rho_0^{i_0}\theta^{i_0}}{1-\rho_0\theta} \\
&= \Big(\eta - \frac{(1-\rho_0)\eta}{1-\rho_0}\Big)\theta^{i_0} + \frac{\rho_0^{i_0}\theta^{i_0}}{1-\rho_0\theta} = \frac{(\rho_0\theta)^{i_0}}{1-\rho_0\theta} \\
&\ge \frac{\rho_0\theta}{1-\rho_0\theta}(\rho_0\theta)^{\ln((1-\rho_0)\eta)/\ln(\rho_0)} = \frac{\rho_0\theta}{1-\rho_0\theta}\big((1-\rho_0)\eta\big)^{\ln(\rho_0\theta)/\ln(\rho_0)}.
\end{aligned} \tag{4.25}
$$

Let $\alpha := \ln(\theta)/\ln(\rho_0) \ge 1$. The equation above and (4.22) imply that

$$
\begin{aligned}
\sum_{i=1}^{\infty} \theta^i \mathbb{E}\big[\Delta_i(t)\big] &\ge \frac{\rho_0\theta}{1-\rho_0\theta}\big((1-\rho_0)\eta\big)^{1+\alpha} \\
&\ge \frac{\rho_0\theta}{1-\rho_0\theta}\Big(\frac{(1-\rho_0)(1-\theta)}{\theta}\mathbb{E}\big[V(t)\big]\Big)^{1+\alpha}.
\end{aligned} \tag{4.26}
$$

Thus, we have found a valid lower bound. We apply the lower bound to (4.16) to find $\frac{d}{dt}\mathbb{E}[V(t)] \le -c_1(\mathbb{E}[V(t)])^{1+\alpha}$, where $c_1 := \rho_0\theta((1-\rho_0)(1-\theta)/\theta)^{1+\alpha}/(2(1-\rho_0\theta)) > 0$. This implies that

$$
\mathbb{E}\big[V(t)\big] \le \frac{1}{((\mathbb{E}[V(0)])^{-\alpha} + c_1\alpha t)^{1/\alpha}} \le \frac{1}{(c_2^{-\alpha} + c_1\alpha t)^{1/\alpha}}, \tag{4.27}
$$

where $c_2 := \theta\rho_0/((1-\theta)(1-\rho_0)) > 0$ and we use the fact that

$$
\mathbb{E}\big[V(0)\big] \le \frac{\theta}{1-\theta}\sum_{i=1}^{\infty}\mathbb{E}\big[\Delta_i(0)\big] \le \frac{\theta}{1-\theta}\sum_{i=1}^{\infty}\rho_0^i = c_2, \tag{4.28}
$$

by (4.22) and Corollary 4.3 in the second inequality. Hence, by (4.22),

$$
\sum_{i=1}^{\infty}\mathbb{E}\big[\Delta_i(t)\big] \le \frac{\mathbb{E}[V(t)]}{\theta} \le \frac{1}{\theta(c_2^{-\alpha} + c_1\alpha t)^{1/\alpha}}, \tag{4.29}
$$

which completes the proof of the theorem. □

4.3. *Proof of the process-level limit.*

LEMMA 4.5. *Fix any $i \in \mathbb{N}$ and $w \in W$. The process*

$$M_i^w(t) := Q_i^w(t) - Q_i^w(0) - \int_0^t \left(\frac{\lambda N}{M} \sum_{\substack{v \in \mathcal{N}_w \\ X_v(s) = i-1}} \sum_{w' \in \mathcal{N}_v} \frac{q_{i-1}^{w'}(s)^d - q_i^{w'}(s)^d}{Q_{i-1}^{w'}(s) - Q_i^{w'}(s)} - \left(Q_i^w(s) - Q_{i+1}^w(s) \right) \right) \mathrm{d}s, \tag{4.30}$$

is a square-integrable martingale started at zero. Moreover, the quadratic variation $[M_i^w]_t$ satisfies

$$\mathbb{E}\left[\left[M_i^w \right]_t \right] = \mathbb{E}\left[\int_0^t \left(\frac{\lambda N}{M} \sum_{\substack{v \in \mathcal{N}_w \\ X_v(s) = i-1}} \sum_{w' \in \mathcal{N}_v} \frac{q_{i-1}^{w'}(s)^d - q_i^{w'}(s)^d}{Q_{i-1}^{w'}(s) - Q_i^{w'}(s)} + \left(Q_i^w(s) - Q_{i+1}^w(s) \right) \right) \mathrm{d}s \right]. \tag{4.31}$$

The proof of Lemma 4.5 is provided in Appendix B.

PROOF OF THEOREM 3.10. Fix any $i \in \mathbb{N}$, $w \in W$ and $t \geq 0$ and let $d_i^w(t) := |q_i^w(t) - \bar{q}_i(t)|$. Then,

$$\begin{aligned} d_i^w(t) \leq{}& \lambda \int_0^t \left| \frac{1}{d_w} \frac{N}{M} \sum_{\substack{v \in \mathcal{N}_w \\ X_v(s) = i-1}} \sum_{w' \in \mathcal{N}_v} \frac{q_{i-1}^{w'}(s)^d - q_i^{w'}(s)^d}{Q_{i-1}^{w'}(s) - Q_i^{w'}(s)} - \left(\bar{q}_{i-1}(s)^d - \bar{q}_i(s)^d \right) \right| \mathrm{d}s \\ &+ \int_0^t \left| \left(q_i^w(s) - q_{i+1}^w(s) \right) - \left(\bar{q}_i(s) - \bar{q}_{i+1}(s) \right) \right| \mathrm{d}s + d_i^w(0) + \frac{|M_i^w(t)|}{d_w}, \end{aligned} \tag{4.32}$$

where $M_i^w(t)$ is a square-integrable martingale as defined in Lemma 4.5. We proceed by bounding the terms on the right-hand side. The term in the first integral in (4.32) is upper bounded by

$$\begin{aligned} &\left| \frac{1}{d_w} \frac{N}{M} \sum_{\substack{v \in \mathcal{N}_w \\ X_v(s) = i-1}} \sum_{w' \in \mathcal{N}_v} \frac{q_{i-1}^{w'}(s)^d - q_i^{w'}(s)^d}{Q_{i-1}^{w'}(s) - Q_i^{w'}(s)} - \left(\bar{q}_{i-1}(s)^d - \bar{q}_i(s)^d \right) \right| \\ &\leq \left| \frac{1}{d_w} \sum_{\substack{v \in \mathcal{N}_w \\ X_v(s) = i-1}} \sum_{w' \in \mathcal{N}_v} \frac{N}{M} \frac{q_{i-1}^{w'}(s)^d - q_i^{w'}(s)^d}{Q_{i-1}^{w'}(s) - Q_i^{w'}(s)} \right. \\ &\qquad \left. - \left(q_{i-1}^w(s) - q_i^w(s) \right) \frac{\bar{q}_{i-1}(s)^d - \bar{q}_i(s)^d}{\bar{q}_{i-1}(s) - \bar{q}_i(s)} \right| \\ &\qquad + \left| \left(q_{i-1}^w(s) - q_i^w(s) \right) - \left(\bar{q}_{i-1}(s) - \bar{q}_i(s) \right) \right| \frac{\bar{q}_{i-1}(s)^d - \bar{q}_i(s)^d}{\bar{q}_{i-1}(s) - \bar{q}_i(s)} \\ &\leq \frac{1}{d_w} \sum_{\substack{v \in \mathcal{N}_w \\ X_v(s) = i-1}} \left| \sum_{w' \in \mathcal{N}_v} \frac{N}{M} \frac{q_{i-1}^{w'}(s)^d - q_i^{w'}(s)^d}{Q_{i-1}^{w'}(s) - Q_i^{w'}(s)} - \frac{\bar{q}_{i-1}(s)^d - \bar{q}_i(s)^d}{\bar{q}_{i-1}(s) - \bar{q}_i(s)} \right| \\ &\qquad + d\left(d_{i-1}^w(s) + d_i^w(s) \right), \end{aligned} \tag{4.33}$$

where we use the triangle inequality in the first inequality and the mean value theorem in the second inequality (see (4.19)). The first term on the right-hand side above is further upper bounded by

$$
\begin{aligned}
&\frac{1}{d_w} \sum_{\substack{v\in\mathcal{N}_w\\ X_v(s)=i-1}} \Big| \sum_{w'\in\mathcal{N}_v} \frac{N}{M} \frac{q_{i-1}^{w'}(s)^d - q_i^{w'}(s)^d}{Q_{i-1}^{w'}(s) - Q_i^{w'}(s)} - \frac{\bar{q}_{i-1}(s)^d - \bar{q}_i(s)^d}{\bar{q}_{i-1}(s) - \bar{q}_i(s)} \Big| \\
&\quad \le \frac{1}{d_w} \sum_{\substack{v\in\mathcal{N}_w\\ X_v(s)=i-1}} \frac{N}{M} \sum_{w'\in\mathcal{N}_v} \frac{1}{d_{w'}} \Big| \frac{q_{i-1}^{w'}(s)^d - q_i^{w'}(s)^d}{q_{i-1}^{w'}(s) - q_i^{w'}(s)} - \frac{\bar{q}_{i-1}(s)^d - \bar{q}_i(s)^d}{\bar{q}_{i-1}(s) - \bar{q}_i(s)} \Big| \\
&\quad\quad + \frac{1}{d_w} \sum_{\substack{v\in\mathcal{N}_w\\ X_v(s)=i-1}} \Big| \frac{N}{M} \sum_{w'\in\mathcal{N}_v} \frac{1}{d_{w'}} - 1 \Big| \frac{\bar{q}_{i-1}(s)^d - \bar{q}_i(s)^d}{\bar{q}_{i-1}(s) - \bar{q}_i(s)} \\
&\quad \le \frac{1}{d_w} \sum_{v\in\mathcal{N}_w} \Big(\frac{N}{M} \sum_{w'\in\mathcal{N}_v} \frac{K}{d_{w'}} \big(d_{i-1}^{w'}(s) + d_i^{w'}(s)\big) + d\phi(G)\big(q_{i-1}^{w}(s) - q_i^{w}(s)\big) \Big),
\end{aligned}
\tag{4.34}
$$

where we use the triangle inequality in the first inequality and Lemma C.1 and the mean value theorem in the second inequality (see (4.19)). Note that the constant K stems from Lemma C.1 and only depends on d. Then, summing the first term on the right-hand side over $w \in W$,

$$
\begin{aligned}
&\sum_{w\in W} \frac{1}{d_w} \sum_{v\in\mathcal{N}_w} \frac{N}{M} \sum_{w'\in\mathcal{N}_v} \frac{K}{d_{w'}} \big(d_{i-1}^{w'}(s) + d_i^{w'}(s)\big) \\
&\quad = \sum_{w\in W} \frac{1}{d_w} \sum_{w'\in W} \frac{N}{M} \sum_{v\in\mathcal{N}_w\cap\mathcal{N}_{w'}} \frac{K}{d_{w'}} \big(d_{i-1}^{w'}(s) + d_i^{w'}(s)\big) \\
&\quad = \sum_{w'\in W} \frac{1}{d_{w'}} \sum_{w\in W} \frac{N}{M} \sum_{v\in\mathcal{N}_w\cap\mathcal{N}_{w'}} \frac{K}{d_w} \big(d_{i-1}^{w'}(s) + d_i^{w'}(s)\big) \\
&\quad = \sum_{w'\in W} \frac{1}{d_{w'}} \sum_{v\in\mathcal{N}_{w'}} \frac{N}{M} \sum_{w\in\mathcal{N}_v} \frac{K}{d_w} \big(d_{i-1}^{w'}(s) + d_i^{w'}(s)\big) \\
&\quad \le \sum_{w'\in W} \frac{1}{d_{w'}} \sum_{v\in\mathcal{N}_{w'}} \frac{\rho_0 K}{\lambda} \big(d_{i-1}^{w'}(s) + d_i^{w'}(s)\big) \\
&\quad = \sum_{w'\in W} \frac{\rho_0 K}{\lambda} \big(d_{i-1}^{w'}(s) + d_i^{w'}(s)\big),
\end{aligned}
\tag{4.35}
$$

where the inequality follows by the assumption that $\rho(G) \le \rho_0$ (recall (2.1) for the definition of $\rho(G)$). The term in the second integral in (4.32) is bounded by

$$
\big|\big(q_i^w(s) - q_{i+1}^w(s)\big) - \big(\bar{q}_i(s) - \bar{q}_{i+1}(s)\big)\big| \le d_i^w(s) + d_{i+1}^w(s). \tag{4.36}
$$

Therefore, putting the above together, by Jensen's inequality and the Cauchy–Schwarz inequality,

$$
\begin{aligned}
\Big(\sum_{w\in W} d_i^w(t)\Big)^2 \le \Big(\sum_{w\in W} \Big(\int_0^t \big((\rho_0 K + d)\big(d_{i-1}^w(s) + d_i^w(s)\big) \\
+ d\phi(G)\big(q_{i-1}^w(s) - q_i^w(s)\big)
\end{aligned}
$$

$$
\begin{aligned}
&+ \big(d_i^w(s) + d_{i+1}^w(s)\big)\big)\,\mathrm{d}s + d_i^w(0) + \frac{|M_i^w(t)|}{d_w}\Big)\Big)^2 \\
&\le 6\Big(\int_0^t \sum_{w\in W} c_1 d_{i-1}^w(s)\,\mathrm{d}s\Big)^2 + 6\Big(\int_0^t \sum_{w\in W} c_1 d_i^w(s)\,\mathrm{d}s\Big)^2 \\
&\quad + 6\Big(\int_0^t \sum_{w\in W} d_{i+1}^w(s)\,\mathrm{d}s\Big)^2 \\
&\quad + 6\Big(\int_0^t \sum_{w\in W} d\phi(G)\big(q_{i-1}^w(s) - q_i^w(s)\big)\,\mathrm{d}s\Big)^2 \\
&\quad + 6\Big(\sum_{w\in W} d_i^w(0)\Big)^2 \\
&\quad + 6\Big(\sum_{w\in W} \frac{|M_i^w(t)|}{d_w}\Big)^2 \\
&\le 6tc_1^2 \int_0^t \Big(\sum_{w\in W} d_{i-1}^w(s)\Big)^2 \mathrm{d}s + 6tc_1^2 \int_0^t \Big(\sum_{w\in W} d_i^w(s)\Big)^2 \mathrm{d}s \\
&\quad + 6t \int_0^t \Big(\sum_{w\in W} d_{i+1}^w(s)\Big)^2 \mathrm{d}s \\
&\quad + 6Mtd^2\phi(G)^2 \int_0^t \sum_{w\in W} \big(q_{i-1}^w(s) - q_i^w(s)\big)^2\,\mathrm{d}s \\
&\quad + 6\Big(\sum_{w\in W} d_i^w(0)\Big)^2 + 6M \sum_{w\in W} \frac{M_i^w(t)^2}{d_w^2},
\end{aligned}
\tag{4.37}
$$

where $c_1 := \rho_0 K + d + 1$. Then, by the monotone convergence theorem,

$$
\begin{aligned}
\sup_{s\in[0,t]} \sum_{i=1}^{\infty}\Big(\frac{1}{M}\sum_{w\in W} d_i^w(s)\Big)^2 &\le c_2 t \int_0^t \sup_{u\in[0,s]} \sum_{i=1}^{\infty}\Big(\frac{1}{M}\sum_{w\in W} d_i^w(u)\Big)^2 \mathrm{d}s \\
&\quad + 6t^2 d^2 \phi(G)^2 + 6\sum_{i=1}^{\infty}\Big(\frac{1}{M}\sum_{w\in W} d_i^w(0)\Big)^2 \\
&\quad + \frac{6}{M}\sum_{i=1}^{\infty}\sum_{w\in W} \sup_{s\in[0,t]} \frac{M_i^w(s)^2}{d_w^2},
\end{aligned}
\tag{4.38}
$$

where $c_2 := 6(2c_1^2 + 1)$ and we use that

$$
\sum_{i=1}^{\infty}\sum_{w\in W} \big(q_{i-1}^w(s) - q_i^w(s)\big)^2 \le \sum_{w\in W}\sum_{i=1}^{\infty}\big(q_{i-1}^w(s) - q_i^w(s)\big) = M. \tag{4.39}
$$

Hence, by Grönwall's inequality,

$$
\begin{aligned}
\sup_{s\in[0,t]} \sum_{i=1}^{\infty}\Big(\frac{1}{M}\sum_{w\in W} d_i^w(s)\Big)^2 &\le 6e^{c_2 t^2}\Big(t^2 d^2 \phi(G)^2 + \sum_{i=1}^{\infty}\Big(\frac{1}{M}\sum_{w\in W} d_i^w(0)\Big)^2 \\
&\quad + \sum_{i=1}^{\infty}\frac{1}{M}\sum_{w\in W} \sup_{s\in[0,t]} \frac{M_i^w(s)^2}{d_w^2}\Big).
\end{aligned}
\tag{4.40}
$$

This almost completes the proof of the theorem. Now, by Jensen's inequality,

$$\begin{aligned}
&\sum_{i=1}^{\infty}\big(q_i(t)-\bar{q}_i(t)\big)^2\\
&\quad\le 2\sum_{i=1}^{\infty}\Big(\frac{1}{M}\sum_{w\in W}q_i^w(t)-q_i(t)\Big)^2+2\sum_{i=1}^{\infty}\Big(\frac{1}{M}\sum_{w\in W}q_i^w(t)-\bar{q}_i(s)\Big)^2\\
&\quad\le 2\phi(G)^2\sum_{i=1}^{\infty}q_i(t)^2+2\sum_{i=1}^{\infty}\Big(\frac{1}{M}\sum_{w\in W}d_i^w(t)\Big)^2\\
&\quad\le 2\phi(G)^2\Big(\frac{N_a(t)}{N}+\sum_{i=1}^{\infty}q_i(0)^2\Big)+2\sum_{i=1}^{\infty}\Big(\frac{1}{M}\sum_{w\in W}d_i^w(t)\Big)^2,
\end{aligned}\tag{4.41}$$

where $N_a(t)$ denotes the number of arrivals until time t and the second inequality follows because

$$\begin{aligned}
&\Big|\frac{N}{M}\sum_{w\in W}q_i^w(t)-Q_i(t)\Big|\\
&\quad=\Big|\frac{N}{M}\sum_{w\in W}\sum_{\substack{v\in\mathcal{N}_w\\ X_v(t)\ge i}}\frac{1}{d_w}-Q_i(t)\Big|=\Big|\sum_{\substack{v\in V\\ X_v(t)\ge i}}\frac{N}{M}\sum_{w\in\mathcal{N}_v}\frac{1}{d_w}-\sum_{\substack{v\in V\\ X_v(t)\ge i}}1\Big|\\
&\quad\le\sum_{\substack{v\in V\\ X_v(t)\ge i}}\Big|\frac{N}{M}\sum_{w\in\mathcal{N}_v}\frac{1}{d_w}-1\Big|\le\sum_{\substack{v\in V\\ X_v(t)\ge i}}\phi(G)=\phi(G)Q_i(t).
\end{aligned}\tag{4.42}$$

Then, (4.40) and (4.41) together imply that

$$\begin{aligned}
&\mathbb{E}\Big[\sup_{s\in[0,t]}\sum_{i=1}^{\infty}\big(q_i(s)-\bar{q}_i(s)\big)^2\Big]\\
&\quad\le 2\phi(G)^2\mathbb{E}\Big[\frac{N_a(t)}{N}+\sum_{i=1}^{\infty}q_i(0)^2\Big]+2\mathbb{E}\Big[\sup_{s\in[0,t]}\sum_{i=1}^{\infty}\Big(\frac{1}{M}\sum_{w\in W}d_i^w(s)\Big)^2\Big]\\
&\quad\le 2\phi(G)^2\Big(\lambda t+\mathbb{E}\Big[\sum_{i=1}^{\infty}q_i(0)^2\Big]\Big)\\
&\qquad+12e^{c_2t^2}\Big(t^2d^2\phi(G)^2+\mathbb{E}\Big[\sum_{i=1}^{\infty}\Big(\frac{1}{M}\sum_{w\in W}d_i^w(0)\Big)^2\Big]+4t(\rho_0d+1)\gamma(G)\Big),
\end{aligned}\tag{4.43}$$

where we use Lemma C.2 in the second inequality. This completes the proof of the theorem. □

4.4. *Analysis of the steady-state.* The mixing time bound above shows that the system is close to the steady-state at a large, but finite time, starting from the empty state. The process-level limit characterizes this sample path and proves that the system remains close to an ODE. Together with a standard global convergence result, this implies that the steady-state is close to the fixed point of the system of ODEs.

PROOF OF THEOREM 3.1. Proposition 4.1 proves the first half of the theorem. To prove the second half, let $\bar{\boldsymbol{q}}(t)$ be the unique solution to the ODEs in Theorem 3.10, where $\bar{\boldsymbol{q}}(0)=0$.

Theorem 3.6 in [28] shows that there exist constants $c_1, c_2 > 0$ (depending only on λ) such that

$$\sum_{i=1}^{\infty}(\bar{q}_i(t) - q_i^*)^2 \le \sum_{i=1}^{\infty}|\bar{q}_i(t) - q_i^*| \le c_1 e^{-c_2 t} \le \frac{c_1}{1+c_2 t}. \tag{4.44}$$

Throughout, denote $\eta = \max\{\phi(G)^2, \gamma(G)\}$. Let $\boldsymbol{X}^{(1)}(t)$ and $\boldsymbol{X}^{(2)}(t)$ be two copies of the Markov process, where $\boldsymbol{X}^{(1)}(0) = 0$ and $\boldsymbol{X}^{(2)}(0) \overset{\mathrm{d}}{=} \boldsymbol{X}^{(2)}(\infty)$. Then, there exist constants $c_4, c_5, c_1', c_2', c_4' > 0$, $c_3 \ge 1$ and $0 < \alpha \le 1$ (depending only on λ, ρ_0 and d) such that, for all $t \ge 1$,

$$\begin{aligned}
\sum_{i=1}^{\infty}\mathbb{E}\big[(q_i^{(2)}(\infty) - q_i^*)^2\big] &= \sum_{i=1}^{\infty}\mathbb{E}\big[(q_i^{(2)}(t) - q_i^*)^2\big] \\
&\le 3\sum_{i=1}^{\infty}\mathbb{E}\big[(q_i^{(2)}(t) - q_i^{(1)}(t))^2\big] + 3\sum_{i=1}^{\infty}\mathbb{E}\big[(q_i^{(1)}(t) - \bar{q}_i(t))^2\big] \\
&\quad + 3\sum_{i=1}^{\infty}\mathbb{E}\big[(\bar{q}_i(t) - q_i^*)^2\big] \\
&\le \frac{3}{(c_4 + c_5 t)^{\alpha}} + 6\phi(G)^2\lambda t \\
&\quad + 36e^{c_3 t^2}\big(t^2 d^2 \phi(G)^2 + 4t(\rho_0 d + 1)\gamma(G)\big) + \frac{3c_1}{1+c_2 t} \\
&\le \frac{1}{(c_1' + c_2' t)^{\alpha}} + c_4' t^2 e^{c_3 t^2}\eta,
\end{aligned} \tag{4.45}$$

where we use Jensen's inequality in the first inequality and Theorem 3.8 and 3.10 in the second inequality. We consider two cases. If $\ln(1/\eta) \ge 2c_3$, then let $t = \sqrt{\ln(1/\eta)/(2c_3)} \ge 1$ such that

$$t^2 e^{c_3 t^2}\eta = \frac{\ln(1/\eta)\sqrt{\eta}}{2c_3} \le \frac{1}{c_3\sqrt{\ln(1/\eta)}} \le \frac{1}{(c_3\sqrt{\ln(1/\eta)})^{\alpha}}, \tag{4.46}$$

where we use that $\ln(1/x)\sqrt{x} \le 2/\sqrt{\ln(1/x)}$ for $0 \le x \le 1$ in the first inequality. Therefore,

$$\begin{aligned}
\sum_{i=1}^{\infty}\mathbb{E}\big[(q_i^{(2)}(\infty) - q_i^*)^2\big] &\le \frac{1}{(c_1' + c_2'\sqrt{\ln(1/\eta)})^{\alpha}} + \frac{c_4'}{(c_3\sqrt{\ln(1/\eta)})^{\alpha}} \\
&\le \frac{c_3^{\alpha} + c_2'^{\alpha} c_4'}{(c_2' c_3\sqrt{\ln(1/\eta)})^{\alpha}}.
\end{aligned} \tag{4.47}$$

If instead $\ln(1/\eta) < 2c_3$, then let $t = 1$ such that

$$\begin{aligned}
\sum_{i=1}^{\infty}\mathbb{E}\big[(q_i^{(2)}(\infty) - q_i^*)^2\big] &\le \frac{1}{(c_1' + c_2')^{\alpha}} + c_4' e^{c_3}\eta \\
&\le \frac{\sqrt{2c_3}}{(c_1' + c_2')^{\alpha}\sqrt{\ln(1/\eta)}} + \frac{c_4' e^{c_3}}{\sqrt{\ln(1/\eta)}} \\
&\le \frac{\sqrt{2c_3} + c_4' e^{c_3}(c_1' + c_2')^{\alpha}}{(c_1' + c_2')^{\alpha}\sqrt{\ln(1/\eta)}},
\end{aligned} \tag{4.48}$$

where we use that $x \le 1/\sqrt{\ln(1/x)}$ for $0 \le x \le 1$ in the second inequality. This completes the proof of the theorem. □

4.5. *Verification for random bipartite geometric graphs.*

PROOF OF COROLLARY 3.4. Fix any $v \in V_N$ and $0 < \varepsilon \le 1/2$. As each $w \in W_N$ is placed independently and uniformly at random, d_v^N is distributed as a binomial random variable. Therefore, a Chernoff bound (see, e.g., Corollary 2.3 in [16]) shows that

$$\mathbb{P}\big(|d_v^N - \mathbb{E}[d_v^N]| \ge \varepsilon \mathbb{E}[d_v^N]\big) \le 2\exp(-\varepsilon^2 \mathbb{E}[d_v^N]/3). \tag{4.49}$$

A similar Chernoff bound holds for $w \in W_N$. Let E_N denote the event that there exists a $v \in V_N$ such that $|d_v^N - \mathbb{E}[d_v^N]| \ge \varepsilon \mathbb{E}[d_v^N]$ or there exists a $w \in W_N$ such that $|d_w^N - \mathbb{E}[d_w^N]| \ge \varepsilon \mathbb{E}[d_w^N]$. Let N_1 be large enough such that $\varepsilon^2 \mathbb{E}[d_v^N]/3 \ge 3\ln N$, $\varepsilon^2 \mathbb{E}[d_w^N]/3 \ge 3\ln N$ and $\varepsilon^2 \mathbb{E}[d_w^N]/3 \ge 3\ln M$ for all $N \ge N_1$. Then,

$$\begin{aligned}
\mathbb{P}(E_N) &\le \sum_{v \in V_N} \mathbb{P}\big(|d_v^N - \mathbb{E}[d_v^N]| \ge \varepsilon \mathbb{E}[d_v^N]\big) + \sum_{w \in W_N} \mathbb{P}\big(|d_w^N - \mathbb{E}[d_w^N]| \ge \varepsilon \mathbb{E}[d_w^N]\big) \\
&\le 2N \exp(-\varepsilon^2 \mathbb{E}[d_v^N]/3) + 2M \exp(-\varepsilon^2 \mathbb{E}[d_w^N]/3) \\
&\le 2N \exp(-3\ln N) + 2M \exp\big(-\ln(M) - 2\ln(N)\big) = \frac{4}{N^2},
\end{aligned} \tag{4.50}$$

for all $N \ge N_1$. Hence, $\sum_{N=1}^{\infty} \mathbb{P}(E_N) < \infty$ and the Borel–Cantelli lemma shows that, almost surely, there exists $N_2 < \infty$ such that E_N does not occur for all $N \ge N_2$. This implies, in particular that,

$$\begin{aligned}
\frac{1-\varepsilon}{1+\varepsilon} &= \frac{N}{M(N)} \frac{(1-\varepsilon)\mathbb{E}[d_v^N]}{(1+\varepsilon)\mathbb{E}[d_w^N]} \le \frac{N}{M(N)} \frac{\min_{v \in V_N} d_v^N}{\max_{w \in W_N} d_w^N} \le \frac{N}{M(N)} \sum_{w \in \mathcal{N}_v} \frac{1}{d_w^N} \\
&\le \frac{N}{M(N)} \frac{\max_{v \in V_N} d_v^N}{\min_{w \in W_N} d_w^N} \le \frac{N}{M(N)} \frac{(1+\varepsilon)\mathbb{E}[d_v^N]}{(1-\varepsilon)\mathbb{E}[d_w^N]} = \frac{1+\varepsilon}{1-\varepsilon},
\end{aligned} \tag{4.51}$$

for all $N \ge N_2$ and therefore

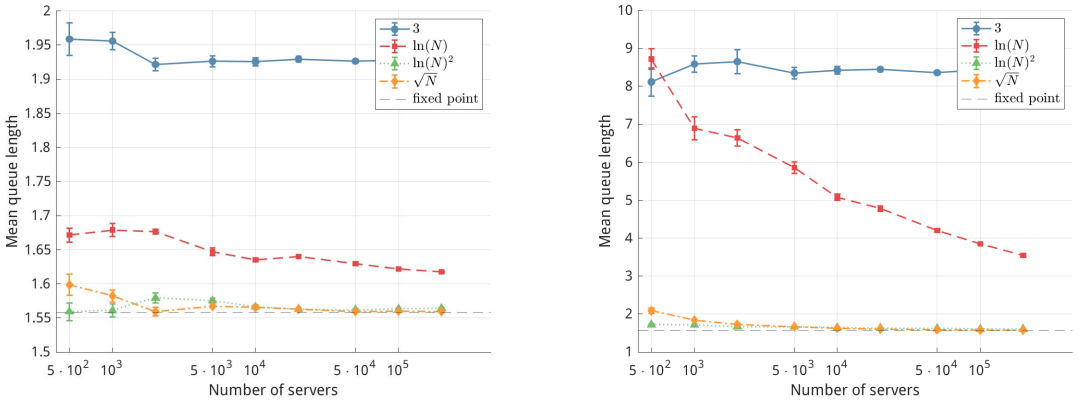
$$\phi(G_N) := \max_{v \in V_N} \left| \frac{N}{M(N)} \sum_{w \in W_N} \frac{1}{d_w^N} - 1 \right| \le \max\left(1 - \frac{1-\varepsilon}{1+\varepsilon}, \frac{1+\varepsilon}{1-\varepsilon} - 1\right) \le \frac{2\varepsilon}{1-\varepsilon} \le 4\varepsilon, \tag{4.52}$$

for all $N \ge N_2$. Also,

$$\gamma(G_N) := \frac{1}{M(N)} \sum_{w \in W_N} \frac{1}{d_w^N} \le \frac{1}{\min_{w \in W_N} d_w^N} \le \frac{1}{(1-\varepsilon)\mathbb{E}[d_w^N]} \le \frac{2}{\ln N}, \tag{4.53}$$

for all $N \ge N_2$. Note also that $\rho(G_N) \le \lambda(1 + \phi(G_N)) \le \lambda(1 + 4\varepsilon) < 1$ for all $N \ge N_2$ and ε small enough. Therefore, Theorem 3.1 completes the proof. □

5. Numerical experiments. We perform numerical experiments to complement the theoretical results. The experiments are in the scenario where $M(N) = N$, $d = 2$ and $\lambda = 0.8$. We simulate two types of graph sequences: random bipartite geometric graphs and random regular bipartite graphs. The random geometric graph is generated as described in its definition in Section 3.2 for $k = 2$ and $p = 2$. The random regular bipartite graph is generated by fixing a degree k upfront. Then, k half-edges are created at each server $v \in V_N$ and task-type $w \in W_N$. The half-edges at the servers are connected to the half-edges at the task types by sequentially picking two available half-edges at random, one at the server side and one at the task-type side and creating an edge between them. Although this may lead to multiple edges, the probability of this happening is negligible for large N.

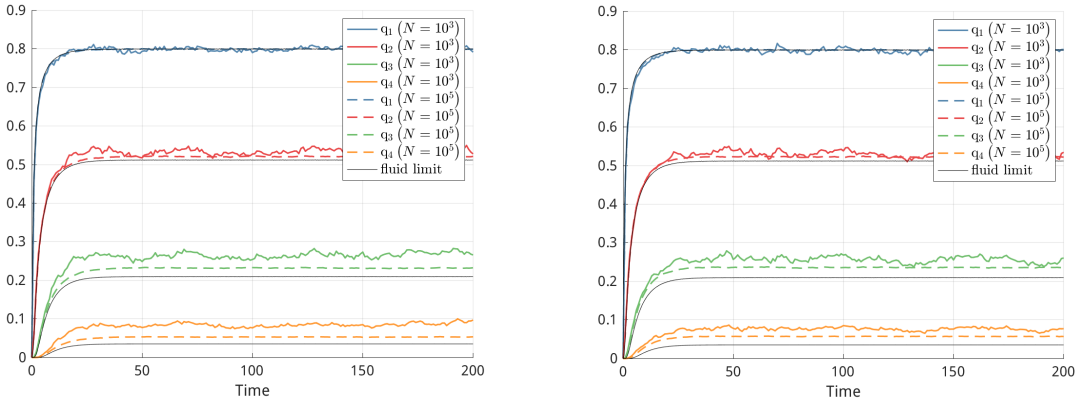


(a) Geometric graph (b) Regular graph

FIG. 2. *The mean queue length in steady-state for a random regular bipartite graph and a random bipartite geometric graph for various degrees compared to the fixed point of the fluid limit.*

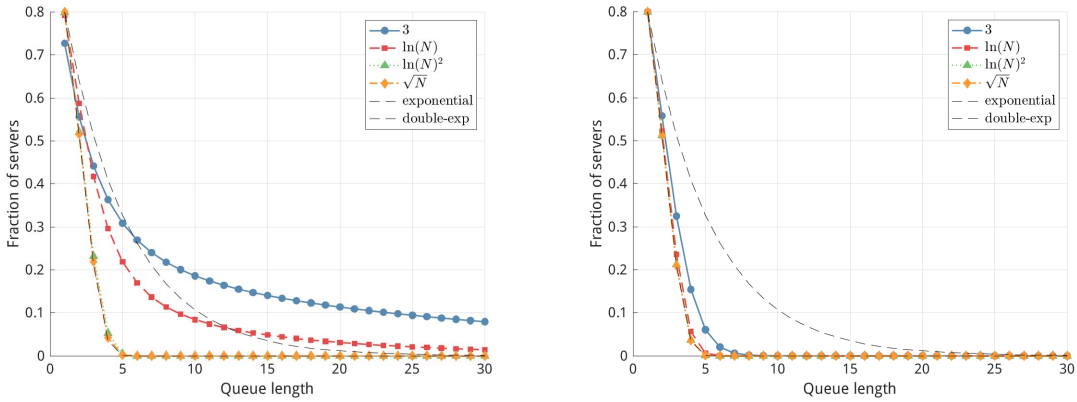
Mean queue length. Figure 2 shows the mean queue length in steady-state for various (average) degrees. For the random bipartite geometric graphs, the mean queue length converges to the fixed point as $N \to \infty$ for an average degree of $(\ln(N))^2$ and $\sqrt{N}$ as expected by our main results. The mean queue length also seems to converge for an average degree of $\ln(N)$, albeit slowly. A rate of $\ln(N)$ is the edge case of our main result and, even though the mean queue length seems to converge, the tail of the occupancy is not double exponential (see Figure 4). For the random regular bipartite graphs, the mean queue length converges to the fixed point as $N \to \infty$ for a degree of $\ln(N)$, $(\ln(N))^2$ and $\sqrt{N}$. The mean queue length does not converge for a constant degree of 3 in either case. Thus, the condition for the regular bipartite graph is both necessary and sufficient.

Process-level limit from the empty state. Figure 3 shows the transient behavior of the system for two values of N, starting from the empty state. As N increases the process remains close to the solution of ODEs, or the fluid limit, for both type of graphs. Note that the process still



(a) Geometric $(\ln(N)^2)$ (b) Regular $(\ln(N))$

FIG. 3. *The process-level limit of the occupancy process $(q_i(t))$ for a random bipartite geometric graph and a random regular bipartite graph and compared to the fluid limit, started from the empty state. The average degree is noted in parentheses.*



(a) Random geometric (b) Random regular

FIG. 4. *The occupancy in steady-state $(q_i(\infty))_{i\geq 1}$ for a random bipartite geometric graph and a random regular bipartite graph for various degrees compared to exponential and double-exponential tails for $N = 10^4$.*

deviates slightly from the fluid limit, especially for $q_3(t)$ and $q_4(t)$, since the average degree grows only logarithmic in N, which directly impacts the convergence rate as in Theorem 3.10.

Exponential or double-exponential tail. There are previous works that have asked whether a similar double-exponential tail of the queue lengths also holds for graphs of constant degree, such as a cycle [14]. Figure 4 shows the occupancy $q_i(\infty)$ in steady-state for various degrees and for $N = 10^5$. The figure compares the occupancy to an exponential tail of λ^i and a double exponential tail of $\lambda^{\frac{d^i-1}{d-1}}$. For the random bipartite geometric graph, the double exponential tail holds for an average degree of $(\ln(N))^2$ and $\sqrt{N}$ as expected by our main results. For an average degree of 3 or $\ln(N)$, the system does not have the double exponential tail, and the performance even appears to be worse than the exponential tail (or random routing on a complete graph). For the random regular bipartite graph, the double exponential tail seems to holds for any choice of degree, even for a constant degree of 3. However, in this case, the queue lengths have double exponential tail as $\lambda^{\frac{d^i-1}{d-1}}$ but with a slightly lower value of d. The question of whether it is possible to analytically characterize this value of d remains a very interesting direction for future work, even for specific regular graphs with constant degree.

APPENDIX A: PROOF OF LEMMA 4.4

PROOF OF LEMMA 4.4. Fix any $0 \leq y < x \leq 1$. Then,

$$\begin{aligned}\frac{x^d - y^d}{x-y} &= \frac{((x-y)+y)^d - y^d}{x-y} \\ &= \frac{\sum_{i=0}^{d}\binom{d}{i}(x-y)^i y^{d-i} - y^d}{x-y} = \sum_{i=1}^{d}\binom{d}{i}(x-y)^{i-1}y^{d-i}.\end{aligned} \tag{A.1}$$

Also,

$$\begin{aligned}\frac{\partial}{\partial y}\frac{x^d-y^d}{x-y} &= \frac{x^d-y^d}{(x-y)^2}-\frac{dy^{d-1}}{x-y} = \frac{((x-y)+y)^d-d(x-y)y^{d-1}-y^d}{(x-y)^2}\\ &= \frac{\sum_{i=0}^{d}\binom{d}{i}(x-y)^i y^{d-i}-d(x-y)y^{d-1}-y^d}{(x-y)^2}\\ &= \sum_{i=2}^{d}\binom{d}{i}(x-y)^{i-2}y^{d-i}\geq 0.\end{aligned} \tag{A.2}$$

Then, by the mean value theorem, there exists $\xi\in[y_1,y_2]$ such that

$$\begin{aligned}\frac{x_1^d-y_1^d}{x_1-y_1} &= \sum_{i=1}^{d}\binom{d}{i}y_1^{d-i}(x_1-y_1)^{i-1}\leq\sum_{i=1}^{d}\binom{d}{i}y_1^{d-i}(x_2-y_1)^{i-1}=\frac{x_2^d-y_1^d}{x_2-y_1}\\ &= \frac{x_2^d-y_2^d}{x_2-y_2}-\frac{\partial}{\partial y}\frac{x_2^d-y^d}{x_2-y}\bigg|_{y=\xi}(y_2-y_1)\leq\frac{x_2^d-y_2^d}{x_2-y_2},\end{aligned} \tag{A.3}$$

which completes the proof of the lemma. □

APPENDIX B: PROOF OF LEMMA 4.5

PROOF OF LEMMA 4.5. Fix any $t\geq 0$. To change the value of $Q_i^w(t)$, a task must arrive to a server $v\in\mathcal{N}_w$ with queue length $i-1$ or a task must depart a server $v\in\mathcal{N}_w$ with queue length i.

Fix any $v\in\mathcal{N}_w$ with $X_v(t-)=i-1$ and let us compute the probability that a task is assigned to v. At the epoch time of an arrival, a task adopts a task type $w'\in W$ uniformly at random. The task is then routed to a server with queue length $i-1$ if and only if the system only samples servers with queue length at least $i-1$ and not only servers with queue length at least i, which happens with probability $q_{i-1}^{w'}(t-)^d-q_i^{w'}(t-)^d$. By symmetry, any server in $\mathcal{N}_{w'}$ with queue length $i-1$ has the same probability of receiving the task and there are a total of $Q_{i-1}^{w'}(t-)-Q_i^{w'}(t-)$ of such eligible server. This results in a probability of

$$\frac{1}{M}\sum_{w'\in\mathcal{N}_v}\frac{q_{i-1}^{w'}(t-)^d-q_i^{w'}(t-)^d}{Q_{i-1}^{w'}(t-)-Q_i^{w'}(t-)}. \tag{B.1}$$

Now, fix any $v\in\mathcal{N}_w$ with $X_v(t-)=i$ and let us compute the probability that a task departs v. At the epoch time of a potential departure, a server $v'\in V$ is chosen uniformly at random and a task departs if v' has at least one task in its queue. This results in a probability of $1/N$.

We describe the arrival and departure process as follows. Let $\mathcal{N}(t)$ be a Poisson process of rate $(\lambda+1)N$. An event of the process is either an arrival of type $w\in W$ with probability $\lambda/((\lambda+1)M)$ or a potential departure at server $v\in V$ with probability $1/((\lambda+1)N)$, independent of the past. Note that this is equivalent to the model description introduced before. Hence, for any $h>0$,

$$\begin{aligned}\mathbb{E}\big[\Delta Q_i^w(t)|\mathcal{F}_t\big] &= \mathbb{E}\big[\Delta Q_i^w(t)|\Delta\mathcal{N}(t)=1,\mathcal{F}_t\big]\mathbb{P}\big(\Delta\mathcal{N}(t)=1\big)\\ &\quad\pm\mathbb{E}\big[\Delta\mathcal{N}(t)|\Delta\mathcal{N}(t)\geq 2\big]\mathbb{P}\big(\Delta\mathcal{N}(t)\geq 2\big)\\ &= \Bigg(\frac{\lambda}{\lambda+1}\sum_{\substack{v\in\mathcal{N}_w\\ X_v(t)=i-1}}\frac{1}{M}\sum_{w'\in\mathcal{N}_v}\frac{q_{i-1}^{w'}(t)^d-q_i^{w'}(t)^d}{Q_{i-1}^{w'}(t)-Q_i^{w'}(t)}-\frac{1}{\lambda+1}\sum_{\substack{v\in\mathcal{N}_w\\ X_v(t)=i}}\frac{1}{N}\Bigg)\\ &\quad\cdot(\lambda+1)Nhe^{-(\lambda+1)Nh}\pm\big((\lambda+1)Nh+2\big)\big((\lambda+1)Nh\big)^2,\end{aligned} \tag{B.2}$$

where $\Delta Q_i^w(t) := Q_i^w(t+h) - Q_i^w(t)$ and $\Delta\mathcal{N}(t) := \mathcal{N}(t+h) - \mathcal{N}(t)$. Here, we use the shorthand notation $\pm x$ to denote a term in $[-x, x]$. Fix any $0 \le s \le t$. The equation above implies that

$$\begin{aligned}\frac{d}{dt}\mathbb{E}\big[Q_i^w(t)|\mathcal{F}_s\big] &= \lim_{h\downarrow 0}\frac{\mathbb{E}[\mathbb{E}[Q_i^w(t+h) - Q_i^w(t)|\mathcal{F}_t]|\mathcal{F}_s]}{h}\\ &= \mathbb{E}\Bigg[\frac{\lambda N}{M}\sum_{\substack{v\in\mathcal{N}_w\\ X_v(t)=i-1}}\sum_{w'\in\mathcal{N}_v}\frac{q_{i-1}^{w'}(t)^d - q_i^{w'}(t)^d}{Q_{i-1}^{w'}(t) - Q_i^{w'}(t)}\\ &\qquad - \big(Q_i^w(t) - Q_{i+1}^w(t)\big)\Big|\mathcal{F}_s\Bigg],\end{aligned} \tag{B.3}$$

and hence, by the second fundamental theorem of calculus and Fubini's theorem,

$$\begin{aligned}\mathbb{E}\big[Q_i^w(t) - Q_i^w(s)|\mathcal{F}_s\big] &= \int_s^t \frac{d}{du}\mathbb{E}\big[Q_i^w(u)|\mathcal{F}_s\big]\,\mathrm{d}u\\ &= \mathbb{E}\Bigg[\int_s^t \frac{\lambda N}{M}\sum_{\substack{v\in\mathcal{N}_w\\ X_v(u)=i-1}}\sum_{w'\in\mathcal{N}_v}\frac{q_{i-1}^{w'}(u)^d - q_i^{w'}(u)^d}{Q_{i-1}^{w'}(u) - Q_i^{w'}(u)}\\ &\qquad - \big(Q_i^w(u) - Q_{i+1}^w(u)\big)\,\mathrm{d}u\Big|\mathcal{F}_s\Bigg],\end{aligned} \tag{B.4}$$

which proves that $\mathbb{E}[M_i^w(t)|\mathcal{F}_s] = M_i^w(s)$. Also,

$$\begin{aligned}\big|M_i^w(t)\big| &\le \big|Q_i^w(t) - Q_i^w(0)\big|\\ &\quad + \int_0^t \frac{\lambda N}{M}\sum_{\substack{v\in\mathcal{N}_w\\ X_v(s)=i-1}}\sum_{w'\in\mathcal{N}_v}\frac{q_{i-1}^{w'}(s)^d - q_i^{w'}(s)^d}{Q_{i-1}^{w'}(s) - Q_i^{w'}(s)} + \big(Q_i^w(s) - Q_{i+1}^w(s)\big)\,\mathrm{d}s\\ &\le d_w + \Bigg(\frac{\lambda N}{M}\sum_{v\in\mathcal{N}_w}\sum_{w'\in\mathcal{N}_v}\frac{d}{d_{w'}} + d_w\Bigg)t < \infty,\end{aligned} \tag{B.5}$$

by the mean-value theorem. This implies, in particular, that $M_i^w(t)$ is a square-integrable martingale.

We proceed by computing the quadratic variation of $M_i^w(t)$. As $Q_i^w(0)$ is a constant and the integral term is a continuous, finite variation process, it follows that $[M_i^w]_t = [Q_i^w]_t$. Furthermore, since $Q_i^w(t)$ is a finite variation process that is right-continuous with left limits, it follows that $[Q_i^w]_t = \sum_{k=1}^n (Q_i^w(t_k) - Q_i^w(t_k-))^2$, where $t_1, t_2, \ldots, t_n$ are the (random) jump times of the process. Now, recall that the jumps of $Q_i^w(t)$ are always equal to one and hence $[Q_i^w]_t$ must simply count the total number of jumps. Thus, a similar computation along the lines of (B.2) and (B.3) yields

$$\frac{d}{dt}\mathbb{E}\big[\big[Q_i^w\big]_t\big] = \mathbb{E}\Bigg[\frac{\lambda N}{M}\sum_{\substack{v\in\mathcal{N}_w\\ X_v(t)=i-1}}\sum_{w'\in\mathcal{N}_v}\frac{q_{i-1}^{w'}(t)^d - q_i^{w'}(t)^d}{Q_{i-1}^{w'}(t) - Q_i^{w'}(t)} + \big(Q_i^w(t) - Q_i^w(t)\big)\Bigg]. \tag{B.6}$$

Then, applying the second fundamental theorem of calculus and Fubini's theorem as done in (B.4) concludes the proof of the lemma. □

APPENDIX C: AUXILIARY LEMMAS

LEMMA C.1. *Fix any* $0 \le y_1 < x_1 \le 1$ *and* $0 \le y_2 < x_2 \le 1$. *Then, there exists* $K > 0$ (*depending only on* d) *such that*

$$\left|\frac{x_1^d - y_1^d}{x_1 - y_1} - \frac{x_2^d - y_2^d}{x_2 - y_2}\right| \le K\big(|x_1 - x_2| + |y_1 - y_2|\big). \tag{C.1}$$

PROOF. Fix any $0 \le y < x \le 1$. Then,

$$\begin{aligned}\frac{x^d - y^d}{x - y} &= \frac{((x-y)+y)^d - y^d}{x - y} = \frac{\sum_{i=0}^d \binom{d}{i}(x-y)^i y^{d-i} - y^d}{x - y} \\ &= \sum_{i=1}^d \binom{d}{i}(x-y)^{i-1} y^{d-i}.\end{aligned} \tag{C.2}$$

Note that $|x^d - y^d| \le d|x - y|$ by the mean value theorem (see (4.19)). Therefore,

$$\begin{aligned}&\left|\frac{x_1^d - y_1^d}{x_1 - y_1} - \frac{x_2^d - y_2^d}{x_2 - y_2}\right| \\ &\quad\le \sum_{i=1}^d \binom{d}{i}\left|(x_1 - y_1)^{i-1} y_1^{d-i} - (x_2 - y_2)^{i-1} y_2^{d-i}\right| \\ &\quad= \sum_{i=1}^d \binom{d}{i}\left|(x_1 - y_1)^{i-1} y_1^{d-i} - (x_1 - y_1)^{i-1} y_2^{d-i}\right. \\ &\qquad\left. + (x_1 - y_1)^{i-1} y_2^{d-i} - (x_2 - y_2)^{i-1} y_2^{d-i}\right| \\ &\quad\le \sum_{i=1}^d \binom{d}{i}\big((d-i)|y_1 - y_2| + (i-1)\big|(x_1 - y_1) - (x_2 - y_2)\big|\big) \\ &\quad\le \sum_{i=1}^d \binom{d}{i}\big((i-1)|x_1 - x_2| + (d-1)|y_1 - y_2|\big) \\ &\quad\le \big(2^d - 1\big)(d-1)\big(|x_1 - x_2| + |y_1 - y_2|\big),\end{aligned} \tag{C.3}$$

which completes the proof of the lemma. □

LEMMA C.2. *Let* $M_i^w(t)$ *be as defined in Lemma* 4.5. *Then, for all* $t \ge 0$,

$$\mathbb{E}\left[\sum_{i=1}^\infty \frac{1}{M}\sum_{w \in W} \sup_{s \in [0,t]} \frac{M_i^w(s)^2}{d_w^2}\right] \le 4t(\rho_0 d + 1)\gamma(G). \tag{C.4}$$

PROOF. Fix any $i \in \mathbb{N}$, $w \in W$ and $t \geq 0$. Note that $M_i^w(t)$ is a square-integrable martingale and therefore, by Doob's martingale inequality,

$$\begin{aligned}
\mathbb{E}\Big[\sup_{s\in[0,t]} M_i^w(s)^2\Big] &\leq 4\mathbb{E}\big[M_i^w(t)^2\big] = 4\mathbb{E}\big[\big[M_i^w\big]_t\big] \\
&= 4\mathbb{E}\Bigg[\int_0^t \frac{\lambda M}{N} \sum_{\substack{v\in\mathcal{N}_w \\ X_v(s)=i-1}} \sum_{w'\in\mathcal{N}_v} \frac{q_{i-1}^{w'}(s)^d - q_i^{w'}(s)^d}{Q_{i-1}^{w'}(s) - Q_i^{w'}(s)} \\
&\qquad + \big(Q_i^w(s) - Q_{i+1}^w(s)\big)\,\mathrm{d}s\Bigg] \qquad \text{(C.5)} \\
&\leq 4\mathbb{E}\Bigg[\int_0^t \sum_{\substack{v\in\mathcal{N}_w \\ X_v(s)=i-1}} \frac{\lambda M}{N} \sum_{w'\in\mathcal{N}_v} \frac{d}{d_{w'}} + \big(Q_i^w(s) - Q_{i+1}^w(s)\big)\,\mathrm{d}s\Bigg] \\
&\leq 4\mathbb{E}\Bigg[\int_0^t \rho_0 d\big(Q_{i-1}^w(s) - Q_i^w(s)\big) + \big(Q_i^w(s) - Q_{i+1}^w(s)\big)\,\mathrm{d}s\Bigg],
\end{aligned}$$

where we use Lemma 4.5 in the second equality and the mean-value theorem in the second inequality. The equation above implies, by the monotone convergence theorem and Fubini's theorem,

$$\begin{aligned}
&\mathbb{E}\Bigg[\sum_{i=1}^{\infty} \frac{1}{M} \sum_{w\in W} \sup_{s\in[0,t]} \frac{M_i^w(s)^2}{d_w^2}\Bigg] \\
&\leq 4\int_0^t \frac{1}{M} \sum_{w\in W} \frac{1}{d_w^2} \sum_{i=1}^{\infty} \mathbb{E}\big[\rho_0 d\big(Q_{i-1}^w(s) - Q_i^w(s)\big) + \big(Q_i^w(s) - Q_{i+1}^w(s)\big)\big]\,\mathrm{d}s \qquad \text{(C.6)} \\
&\leq 4\int_0^t \frac{1}{M} \sum_{w\in W} \frac{\rho_0 d + 1}{d_w}\,\mathrm{d}s = 4t(\rho_0 d + 1)\gamma(G),
\end{aligned}$$

which completes the proof of the lemma. □

Funding. The work was supported in part by NSF Grants CIF-2113027 and CPS-2240982.

REFERENCES

[1] AGARWAL, P. and RAMANAN, K. (2020). Invariant states of hydrodynamic limits of randomized load balancing networks. arXiv preprint. Available at arXiv:2008.08510.

[2] AGHAJANI, R. and RAMANAN, K. (2019). The hydrodynamic limit of a randomized load balancing network. *Ann. Appl. Probab.* **29** 2114–2174. MR3984253 https://doi.org/10.1214/18-AAP1444

[3] ANTON, E., AYESTA, U., JONCKHEERE, M. and VERLOOP, I. M. (2020). Improving the performance of heterogeneous data centers through redundancy. *Proc. ACM Meas. Anal. Comput. Syst.* **4** 1–29.

[4] ANTON, E., AYESTA, U., JONCKHEERE, M. and VERLOOP, I. M. (2021). On the stability of redundancy models. *Oper. Res.* **69** 1540–1565. MR4330629

[5] BARBOUR, A. D. (1980). Density dependent Markov population processes. In *Biological Growth and Spread* (*Proc. Conf., Heidelberg*, 1979) (H. Rost and P. Tautu, eds.). *Lecture Notes in Biomathematics* **38** 36–49. Springer, Berlin. MR0609344

[6] BRAMSON, M. (2011). Stability of join the shortest queue networks. *Ann. Appl. Probab.* **21** 1568–1625. MR2857457 https://doi.org/10.1214/10-AAP726

[7] BRAMSON, M., LU, Y. and PRABHAKAR, B. (2012). Asymptotic independence of queues under randomized load balancing. *Queueing Syst.* **71** 247–292. MR2943660 https://doi.org/10.1007/s11134-012-9311-0

[8] BRAMSON, M., LU, Y. and PRABHAKAR, B. (2013). Decay of tails at equilibrium for FIFO join the shortest queue networks. *Ann. Appl. Probab.* **23** 1841–1878. MR3114919 https://doi.org/10.1214/12-AAP888

[9] BUDHIRAJA, A., MUKHERJEE, D. and WU, R. (2019). Supermarket model on graphs. *Ann. Appl. Probab.* **29** 1740–1777. MR3914555 https://doi.org/10.1214/18-AAP1437

[10] CARDINAELS, E., BORST, S. C. and VAN LEEUWAARDEN, J. S. H. (2019). Job assignment in large-scale service systems with affinity relations. *Queueing Syst.* **93** 227–268. MR4032926 https://doi.org/10.1007/s11134-019-09633-y

[11] CHOUDHURY, T., JOSHI, G., WANG, W. and SHAKKOTTAI, S. (2021). Job dispatching policies for queueing systems with unknown service rates. In *Proceedings of the Twenty-Second International Symposium on Theory, Algorithmic Foundations, and Protocol Design for Mobile Networks and Mobile Computing* 181–190.

[12] COMTE, C. (2019). Dynamic load balancing with tokens. *Comput. Commun.* **144** 76–88.

[13] DOWN, D., MEYN, S. P. and TWEEDIE, R. L. (1995). Exponential and uniform ergodicity of Markov processes. *Ann. Probab.* **23** 1671–1691. MR1379163

[14] GAST, N. (2015). The power of two choices on graphs: The pair-approximation is accurate. In *Proc. MAMA Workshop* 2015 69–71. https://doi.org/10.1145/2825236.2825263

[15] GUPTA, V. and WALTON, N. (2019). Load balancing in the nondegenerate slowdown regime. *Oper. Res.* **67** 281–294. MR3919870 https://doi.org/10.1287/opre.2018.1768

[16] JANSON, S., ŁUCZAK, T. and RUCINSKI, A. (2000). *Random Graphs. Wiley-Interscience Series in Discrete Mathematics and Optimization*. Wiley-Interscience, New York. MR1782847 https://doi.org/10.1002/9781118032718

[17] JONCKHEERE, M., MOYAL, P., RAMÍREZ, C. and SOPRANO-LOTO, N. (2022). Generalized max-weight policies in stochastic matching. *Stoch. Syst.* **13** 40–58. MR4574835 https://doi.org/10.1287/stsy.2022.0098

[18] KURTZ, T. G. (1970). Solutions of ordinary differential equations as limits of pure jump Markov processes. *J. Appl. Probab.* **7** 49–58. MR0254917 https://doi.org/10.2307/3212147

[19] KURTZ, T. G. (1971). Limit theorems for sequences of jump Markov processes approximating ordinary differential processes. *J. Appl. Probab.* **8** 344–356. MR0287609 https://doi.org/10.1017/s002190020003535x

[20] KURTZ, T. G. (1978). Strong approximation theorems for density dependent Markov chains. *Stochastic Process. Appl.* **6** 223–240. MR0464414 https://doi.org/10.1016/0304-4149(78)90020-0

[21] LU, X., KONG, F., YIN, J., LIU, X., YU, H. and FAN, G. (2015). Geographical job scheduling in data centers with heterogeneous demands and servers. In 2015 *IEEE 8th International Conference on Cloud Computing* 413–420. IEEE, New York, NY, USA.

[22] LU, Y., XIE, Q., KLIOT, G., GELLER, A., LARUS, J. R. and GREENBERG, A. (2011). Join-Idle-Queue: A novel load balancing algorithm for dynamically scalable web services. *Perform. Eval.* **68** 1056–1071.

[23] LUCZAK, M. J. and MCDIARMID, C. (2006). On the maximum queue length in the supermarket model. *Ann. Probab.* **34** 493–527. MR2223949 https://doi.org/10.1214/00911790500000710

[24] MASWOOD, M. M. S., NASIM, R., KASSLER, A. J. and MEDHI, D. (2018). Cost-efficient resource scheduling under QoS constraints for geo-distributed data centers. In *NOMS* 2018—2018 *IEEE/IFIP Network Operations and Management Symposium* 1–9. IEEE, Taipei, Taiwan.

[25] MCDONALD, D. R. and TURNER, S. R. E. (2000). Comparing load balancing algorithms for distributed queueing networks. In *Analysis of Communication Networks*: *Call Centres, Traffic and Performance* (*Toronto, ON*, 1998). *Fields Inst. Commun.* **28** 109–133. Amer. Math. Soc., Providence, RI. MR1788713 https://doi.org/10.1016/s0370-2693(00)00162-3

[26] MEYN, S. P. and TWEEDIE, R. L. (1993). Stability of Markovian processes. III. Foster–Lyapunov criteria for continuous-time processes. *Adv. in Appl. Probab.* **25** 518–548. MR1234295 https://doi.org/10.2307/1427522

[27] MITZENMACHER, M. (1996). The power of two choices in randomized load balancing. PhD thesis, Univ. California, Berkeley. MR2695522

[28] MITZENMACHER, M. (2001). The power of two choices in randomized load balancing. *IEEE Trans. Parallel Distrib. Syst.* **12** 1094–1104. https://doi.org/10.1109/71.963420

[29] MUKHERJEE, D., BORST, S. C. and VAN LEEUWAARDEN, J. S. H. (2018). Asymptotically optimal load balancing topologies. *Proc. ACM Meas. Anal. Comput. Syst.* **2** 1–29. https://doi.org/10.1145/3179417

[30] NORMAN, M. F. (1972). *Markov Processes and Learning Models. Mathematics in Science and Engineering* **84**. Academic Press, New York. MR0423546

[31] NORMAN, M. F. (1974). A central limit theorem for Markov processes that move by small steps. *Ann. Probab.* **2** 1065–1074. MR0368150 https://doi.org/10.1214/aop/1176996498

[32] PANIGRAHY, N. K., VASANTAM, T., BASU, P., TOWSLEY, D., SWAMI, A. and LEUNG, K. K. (2022). On the analysis and evaluation of proximity based load balancing policies. *ACM Trans. Model. Perform. Eval. Comput. Syst.* https://doi.org/10.1145/3549933

[33] PENROSE, M. (2003). *Random Geometric Graphs*. *Oxford Studies in Probability* **5**. Oxford Univ. Press, Oxford. MR1986198 https://doi.org/10.1093/acprof:oso/9780198506263.001.0001

[34] RUDIN, W. (1964). *Principles of Mathematical Analysis*, 2nd ed. McGraw-Hill, New York. MR0166310

[35] RUTTEN, D. and MUKHERJEE, D. (2022). Load balancing under strict compatibility constraints. *Math. Oper. Res.* **48** 227–256. MR4567285

[36] TSITSIKLIS, J. N. and XU, K. (2017). Flexible queueing architectures. *Oper. Res.* **65** 1398–1413. MR3710053 https://doi.org/10.1287/opre.2017.1620

[37] VAN DER BOOR, M., BORST, S. C., VAN LEEUWAARDEN, J. S. H. and MUKHERJEE, D. (2022). Scalable load balancing in networked systems: A survey of recent advances. *SIAM Rev.* **64** 554–622. MR4461562 https://doi.org/10.1137/20M1323746

[38] VAN DER BOOR, M. and COMTE, C. (2021). Load balancing in heterogeneous server clusters: Insights from a product-form queueing model. In 2021 *IEEE/ACM* 29*th International Symposium on Quality of Service* (*IWQOS*) 1–10. IEEE, Tokyo, Japan.

[39] VVEDENSKAYA, N. D., DOBRUSHIN, R. L. and KARPELEVICH, F. I. (1996). A queueing system with a choice of the shorter of two queues—an asymptotic approach. *Problemy Peredachi Informatsii* **32** 20–34. MR1384927

[40] WENG, W. and WANG, W. (2020). Achieving zero asymptotic queueing delay for parallel jobs. *Proc. ACM Meas. Anal. Comput. Syst.* **4** 1–36.

[41] WENG, W., ZHOU, X. and SRIKANT, R. (2020). Optimal load balancing with locality constraints. *Proc. ACM Meas. Anal. Comput. Syst.* **4** 1–37. https://doi.org/10.1145/3428330