Mean-field Analysis for Load Balancing on Spatial Graphs

Daan Rutten
drutten@gatech.edu
Georgia Institute of Technology
Atlanta, GA, United States

Debankur Mukherjee
debankur.mukherjee@isye.gatech.edu
Georgia Institute of Technology
Atlanta, GA, United States

ACM Reference Format:
Daan Rutten and Debankur Mukherjee. 2023. Mean-field Analysis for Load Balancing on Spatial Graphs. In *Abstract Proceedings of the 2023 ACM SIGMETRICS International Conference on Measurement and Modeling of Computer Systems (SIGMETRICS '23 Abstracts), June 19–23, 2023, Orlando, FL, USA.* ACM, New York, NY, USA, 2 pages. https://doi.org/10.1145/3578338.3593552

1 INTRODUCTION

A pivotal methodological tool behind the analysis of large-scale load balancing systems is *mean-field analysis*. The high-level idea is to represent the system state by aggregate quantities and characterize their rate of change as the system size grows large. An assumption for the above scheme to work is that the aggregate quantity is Markovian such that its rate of change can be expressed as a function of its current state. If the aggregate quantity is not Markovian, not only does this technique break down, the mean-field approximation may even turn out to be highly inaccurate.

In load balancing systems, if servers are exchangeable, then the aggregate quantity 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*, i.e., the fact that servers need to store resources to process tasks of a particular type locally and have only limited storage space. An emerging line of work thus considers a bipartite graph between task types and servers [2, 3, 5–7]. In this *compatibility* graph, an edge between a server and a task type represents the server's ability to process these tasks. In practice, storage capacity or geographical constraints force a server to process only a small subset of all task types, leading to sparse network topologies. This motivates the study of load balancing in systems with suitably *sparse* bipartite compatibility graphs.

The analysis of sparse systems poses significant challenges, mainly due to the fact that the aggregate quantity is no longer Markovian. One key question to understand here is: *Under what conditions on the (sparse) compatibility graph does the system behavior retain the performance benefits of the fully flexible system?* From a more foundational standpoint, this is equivalent to understanding how far the validity of the mean-field approximation can be extended to non-trivial graphs.

SIGMETRICS '23 Abstracts, June 19–23, 2023, Orlando, FL, USA

ACM ISBN 979-8-4007-0074-3/23/06.
https://doi.org/10.1145/3578338.3593552

2 MODEL DESCRIPTION

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 d_v^N be the number of compatible task types for a server $v \in V_N$ and d_w^N be the number of compatible servers for a task type $w \in W_N$. 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 its compatible servers 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.

We let $X_v(t)$ denote the queue length of a server $v \in V_N$ at time t. Let $q_i^N(t) := \sum_{v \in V_N} \mathbb{1}\{X_v(t) \geq i\}/N$ denote the fraction of servers with queue length at least $i \in \mathbb{N}$ in the entire system.

3 MAIN RESULTS

The criteria for ergodicity of the queue length process are known and have been developed, for example by Bramson [1] and Cardinaels et al. [3]. However, we work with a slightly stronger, but simplified condition on the graph as follows. Let

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

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 $d \geq 2$. 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.

We make contributions on four fronts: (a) establish bounds on a *large-scale mixing time* of the underlying Markov process; (b) quantify how much the transient behavior deviates from the mean-field ODE, starting from only empty queues, in terms of certain graph parameters; (c) 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) 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-scale systems are known to be hard to obtain, even without compatibility constraints [4]. First, as discussed in [4], 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, cannot be used either since these regeneration lengths are typically exponential in N. 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? We show that this large-scale mixing time is polynomial and does not scale with N.

Theorem 3.1. *Let $\boldsymbol{X}_0$ be a random variable such that $\boldsymbol{X}_0 \leq_{st} \boldsymbol{X}(\infty)$. Suppose that $\boldsymbol{X}^{(1)}(0) \stackrel{d}{=} \boldsymbol{X}_0$ and $\boldsymbol{X}^{(2)}(0) \stackrel{d}{=} \boldsymbol{X}^{(2)}(\infty)$. 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 $t \geq 0$,*

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

In the above, the mixing time bound holds in particular from the empty state, i.e., $\boldsymbol{X}^1(0) = \boldsymbol{0}$. A crucial argument in the proof of Theorem 3.1 relies on a novel stochastic coupling, which ensures that the monotonicity of the queue length in the starting state is maintained throughout for any sample path.

(b) Process-level limit starting from the empty state. Next, we characterize the asymptotics of the sample path of $\boldsymbol{q}^N(t)$ starting from a system with only empty queues. Let us introduce two quantities of the underlying graph:

$$\begin{aligned} \phi(G_N) &:= \max_{v \in V_N} \left| \frac{N}{M(N)} \sum_{w \in \mathcal{N}_v} \frac{1}{d_w^N} - 1 \right|, \\ \gamma(G_N) &:= \frac{1}{M(N)} \sum_{w \in W_N} \frac{1}{d_w^N}. \end{aligned} \tag{3}$$

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. 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.

Theorem 3.2. *Let $\boldsymbol{X}(0) = 0$ and $\bar{\boldsymbol{q}}(t) = (\bar{q}_i(t))_{i \geq 1}$ be the unique solution to the system of ODEs*

$$\frac{d\bar{q}_i(t)}{dt} = \lambda \left(\bar{q}_{i-1}(t)^d - \bar{q}_i(t)^d \right) - (\bar{q}_i(t) - \bar{q}_{i+1}(t)) \text{ for } i \in \mathbb{N}. \tag{4}$$

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

$$\begin{aligned} \mathbb{E}\Big[\sup_{s \in [0,t]} \sum_{i=1}^{\infty} \left(q_i^N(s) - \bar{q}_i(s) \right)^2 \Big] &\leq 2\lambda t \phi(G_N)^2 \\ &+ 12 e^{ct^2} \Big(t^2 d^2 \phi(G_N)^2 + 4t(\rho_0 d + 1)\gamma(G_N) \Big). \end{aligned} \tag{5}$$

(c) Mean-field approximation. Leveraging the mixing time bound of Theorem 3.1 and the process-level limit of Theorem 3.2, we determine the applicability of the mean-field approximation for any compatibility graph in terms of $\phi(G_N)$ and $\gamma(G_N)$.

Theorem 3.3. *Given any G_N, if $\max\{\phi(G_N), \gamma(G_N)\} < 1$, then there exist constants $c, \alpha > 0$ (depending only on λ, ρ_0 and d) such that*

$$\sum_{i=1}^{\infty} \mathbb{E}\left[\left(q_i^N(\infty) - q_i^* \right)^2 \right] \leq \frac{c}{\ln\left(1/\max\{\phi(G_N)^2, \gamma(G_N)\}\right)^{\alpha}}, \tag{6}$$

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

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 spatial distance $r(N) > 0$ apart. We prove that, if $r(N)$ is such that

$$\liminf_{N \to \infty} \frac{\mathbb{E}\left[d_v^N\right]}{\ln N} = \infty \text{ and } \liminf_{N \to \infty} \frac{\mathbb{E}\left[d_w^N\right]}{\max(\ln M(N), \ln N)} = \infty, \tag{7}$$

then indeed $\max\{\phi(G_N), \gamma(G_N)\} \to 0$ a.s., and $\boldsymbol{q}^N(t)$ in steady state becomes asymptotically indistinguishable from the fully flexible system. 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 much more generally 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*. We prove that the convergence holds whenever $\gamma(G_N) \to 0$, which happens for example if $\min_{w \in W_N} d_w^N$ diverges (at any rate) as $N \to \infty$, and thus ensures sparsity. This includes *arbitrary deterministic graph sequences* and thus significantly broadens the applicability of the mean-field approximation significantly.

ACKNOWLEDGMENTS

The work was partially supported by the NSF grant CIF-2113027.

REFERENCES

[1] Maury Bramson. 2011. Stability of join the shortest queue networks. *Ann. Appl. Probab.* 21, 4 (2011), 1568–1625. https://doi.org/10.1214/10-AAP726
[2] Amarjit Budhiraja, Debankur Mukherjee, and Ruoyu Wu. 2019. Supermarket model on graphs. *Ann. Appl. Probab.* 29, 3 (2019), 1740–1777. https://doi.org/10.1214/18-AAP1437
[3] Ellen Cardinaels, Sem C Borst, and Johan S H van Leeuwaarden. 2019. Job assignment in large-scale service systems with affinity relations. *Queueing Syst.* 93, 3-4 (2019), 227–268. https://doi.org/10.1007/s11134-019-09633-y
[4] Malwina J. Luczak and Colin McDiarmid. 2006. On the maximum queue length in the supermarket model. *Ann. Probab.* 34, 2 (2006), 493–527. https://doi.org/10.1214/00911790500000710
[5] Debankur Mukherjee, Sem C. Borst, and Johan S. H. Van Leeuwaarden. 2018. Asymptotically optimal load balancing topologies. *Proc. ACM Meas. Anal. Comput. Syst.* 2, 1 (2018), 1–29. https://doi.org/10.1145/3179417
[6] Daan Rutten and Debankur Mukherjee. 2022. Load balancing under strict compatibility constraints. *Math. Oper. Res.* (2022). https://doi.org/10.1287/moor.2022.1258
[7] Wentao Weng, Xingyu Zhou, and R. Srikant. 2020. Optimal load balancing with locality constraints. *Proc. ACM Meas. Anal. Comput. Syst.* 4, 3 (2020), 1–37. https://doi.org/10.1145/3428330