Load Balancing Under Strict Compatibility Constraints

Daan Rutten
Debankur Mukherjee
drutten@gatech.edu
debankur.mukherjee@isye.gatech.edu
Georgia Institute of Technology
Atlanta, Georgia, USA

ABSTRACT
Consider a system with \( N \) identical single-server queues and \( M(N) \) task types, where each server is able to process only a small subset of possible task types. Arriving tasks select \( d \geq 2 \) random compatible servers, and join the shortest queue among them. The compatibility constraints are captured by a fixed bipartite graph \( G_N \) between the servers and the task types. When \( G_N \) is complete bipartite, the meanfield approximation is accurate. However, such dense compatibility graphs are infeasible for large-scale implementation. We characterize a class of sparse compatibility graphs for which the meanfield approximation remains valid. For this, we introduce a novel notion, called proportional sparsity, and establish that systems with proportionally sparse compatibility graphs asymptotically match the performance of a fully flexible system. Furthermore, we show that proportionally sparse random compatibility graphs can be constructed, which reduce the server-degree almost by a factor \( N/\ln(N) \) compared to the complete bipartite compatibility graph.

CCS CONCEPTS
• Mathematics of computing → Markov processes; • Networks → Cloud computing; Network performance analysis; • Theory of computation → Random network models;

KEYWORDS
meanfield limit; power-of-d, stochastic coupling; load balancing on network; data locality; many-server asymptotics; queueing theory

ACM Reference format:
Daan Rutten and Debankur Mukherjee. 2021. Load Balancing Under Strict Compatibility Constraints. In Proceedings of Abstract Proceedings of the 2021 ACM SIGMETRICS / International Conference on Measurement and Modeling of Computer Systems, Virtual Event, China, June 14–18, 2021 (SIGMETRICS ’21 Abstracts), 2 pages.
https://doi.org/10.1145/3410220.3456275

A canonical model for large-scale systems, such as data centers and cloud networks, consists of a large number of parallel servers with dedicated queues. Tasks arrive into the system sequentially in time and are immediately and irrevocably assigned, using some efficient load balancing algorithm, to one of these queues, where they wait until executed. When each task is compatible with each server, the meanfield approximation is proven to be accurate. However, due to ever-increasing heterogeneity in the incoming traffic, large-scale systems typically suffer from stringent task-server compatibility constraints. Indeed, executing a task at a server requires some pre-stored data, and being able to serve all possible task types comes with an excessive storage capacity requirement \([5, 6]\) and an overwhelming implementation complexity \([2–4]\). Consequently, full flexibility in task allocation is not a luxury large-scale systems can afford. It is therefore important to understand the performance of load balancing algorithms under sparser compatibility constraints, where tasks of a particular type can only be served by a relatively small number of servers, naturally viewed as neighbors in a bipartite compatibility graph between the servers and the task types. Our goal in this paper is to characterize the class of sparse compatibility graphs for which the meanfield approximation remains valid.

Consider a system that consists of a set of dispatchers \( W_N \) and a set of servers \( V_N \), with \( |V_N| = N \) and \( |W_N| = M(N) \). Each dispatcher handles arrivals of a particular task type. Tasks of each type arrive as independent Poisson processes of rate \( \lambda N / M(N) \) for \( \lambda < 1 \). Each task requires exponentially distributed service time with unit mean. We will be looking at a scaling regime where both \( N, M(N) \to \infty \). The task-server compatibility is captured in terms of a bipartite graph \( G_N = (V_N, W_N, E_N) \) between \( V_N \) and \( W_N \), where \( E_N \subseteq V_N \times W_N \) is the set of edges. That is, a server \( v \in V_N \) shares an edge in \( G_N \) with a task type \( w \in W_N \), if task type \( w \) can be processed by server \( v \); see Figure 1 for an illustration of

Figure 1: A schematic overview of the system with task types \( W_N \), servers \( V_N \), and their compatibility relation.
the model. Following the JSQ(d) policy, when a task of type \( w \) arrives, \( d \) servers which share an edge with \( w \) are sampled uniformly at random and the task is routed to the shortest of the sampled queues. The quantity of interest is the global occupancy process \( q^N(t) = (q^N_1(t), q^N_2(t), \ldots) \), where \( q^N_i(t) \) denotes the fraction of servers with queue length at least \( i \) at time \( t \) in the \( N \)-th system. Note that the case when \( G_N \) is complete bipartite corresponds to the fully flexible system. Our focus is to identify the sparsest compatibility structures that preserve the performance benefits of a fully flexible system, asymptotically as \( N \to \infty \). In other words, we study the sparsity condition for the compatibility graph, which preserves the validity of the meanfield approximation. Specifically, our results can be categorized into two groups.

(1) Arbitrary deterministic compatibility graphs. We start by considering an arbitrary deterministic sequence of graphs \( \{G_N\}_{N \geq 1} \), indexed by the number of servers \( N \), and define a novel notion of expansion, which we call proportional sparsity. We show that if the sequence of compatibility graphs is proportionally sparse, then as \( N \to \infty \), on any finite time interval, the occupancy process \( q^N(\cdot) \) under the JSQ(d) policy converges to the same meanfield limit as the sequence of fully flexible systems. In fact, this process-level limit result extends to a broad class of load balancing algorithms, for which the assignment decision depends ‘smoothly’ on the empirical queue length distribution of the compatible servers. We call such algorithms Lipschitz continuous task assignment policies. An important step to prove the process-level limit is to show that for almost all dispatchers, the empirical queue length distribution observed in its neighborhood, is close to the empirical queue length distribution observed among all servers in the system. This allows us to construct a coupling between the constrained system and the fully flexible system and establish that the \( \ell_1 \)-distance between the global occupancy processes in two systems is small uniformly over any finite time interval.

For the interchange of limits and hence the convergence of steady state, two more key ingredients that we need are ergodicity of the prelimit system (for each fixed \( N \)) and the tightness of steady states in an appropriate sense. Note that if \( G_N \) is not complete bipartite, the occupancy process \( q^N(\cdot) \) is no longer Markovian. Consequently, one needs to be careful in defining its time asymptotics and hence, the interchange of limits. For ergodicity of the underlying Markov process, we need the graph sequence to satisfy a certain subcriticality condition that was first introduced in [1]. The tightness, however, is technically more challenging. In particular, we need to show that the sequence of steady state occupancy is tight with respect to a certain weighted \( \ell_1 \) norm. For this, we can construct a coupling of Lyapunov functions, which provide uniform tail bounds on the steady state of the global occupancy process.

Combining the above results, we conclude that if a sequence of graphs is proportionally sparse and satisfies the subcriticality condition, then both finite-time dynamics and steady state behavior of the empirical queue length process coincide with that of a fully flexible system, asymptotically as \( N \to \infty \). It is worth highlighting that in the above interchange of limits, we do not impose any restrictions on how the number of task types \( M(N) \) scales with \( N \). This includes the two popular scenarios \( M(N) = \text{constant} \) and \( M(N) = N \) as special cases.

(2) Random compatibility graphs. The results for deterministic graph sequence provide us all the theoretical framework needed to analyze these systems. Next, we exploit these results to construct random sparse compatibility graphs with desired performance benefits. In the context of data-file placement or content replication in large-scale systems, the degree of a server in the compatibility graph can be thought to be roughly proportional to the storage capacity requirement of that server. It is also considered to be a measure of complexity of the network. To this end, we consider two cases.

First, suppose that the servers are constrained to have degrees exactly equal to \( c(N) \). In this case construct \( G_N \) by selecting \( c(N) \) task types for each server, independently uniformly at random, without replacement, from the set of all task types. For such a randomly constructed compatibility graph, we establish that the empirical queue length distribution of the system has the same asymptotic law as the fully flexible system, both process-level and in steady state, if \( c(N) \gg M(N) \ln(N)/N \) and \( c(N) \gg 1 \).

Second, we consider a system that allows for inhomogeneous levels of flexibility for different task types. In this case, the compatibility graph is constructed by selecting each edge incident to a task-type \( w \in W_{\infty} \) with probability \( p_{w}(N) \), independently of other edges. Thus, task type \( w \) will have an average degree \( N p_{w}(N) \). In this case, we show that the empirical queue length distribution of the system has the same asymptotic law as the fully flexible system, both process-level and in steady state, if \( \min_{w \in W_{\infty}} p_{w}(N) \) and the \( \ell_2 \) norm of the inverse probability vector \( (1/p_{w}(N))_{w \in W_{\infty}} \) satisfy suitable growth conditions.

To prove the results for random instances, we verify using concentration of measure arguments, that the graph sequence satisfies both the proportional sparsity and the subcriticality conditions, under the respective growth rate conditions as \( N \to \infty \).

Extensive simulation experiments are conducted to corroborate the theoretical results. A full version of this paper is available at https://arxiv.org/abs/2008.07562.

**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] Aatish Mihara, Joseph L. Hellerstein, Walid al抖m, and Chaita R Das. 2010. Towards characterizing cloud backend workloads: insights from google compute clusters. *ACM SIGMETRICS Performance Evaluation Review* 37, 4 (2010), 34–41. https://doi.org/10.1145/1773394.1774300

[3] Charles Reiss, Alexey Tumanov, Gregory R. Ganger, Randy H. Katz, and Michael A. Kozuch. 2012. Heterogeneity and dynamics of clouds at scale: Google trace analysis. In *Proceedings of the Third ACM Symposium on Cloud Computing (SoCC ’12).* Association for Computing Machinery, Article 7, 13 pages. https://doi.org/10.1145/2391229.2391236

[4] John N. Tsitsiklis and Kuang Xu. 2017. Flexible queueing architectures. *Oper. Res.* 65, 5 (2017), 1398–1413. https://doi.org/10.1287/opre.2017.1620

[5] Weina Wang, Kai Zhu, Lei Yang, Jian Tan, and Li Zhang. 2016. MapTask scheduling in MapReduce with data locality: Throughput and heavy-traffic optimality. *IEEE/ACM Trans. Netw.* 24, 1 (2016), 190–203. https://doi.org/10.1109/TNET.2014.2362745

[6] Qianmin Xie, Ali Yekkehkhany, and Yi Lu. 2016. Scheduling with multi-level data locality: Throughput and heavy-traffic optimality. In *Proc. INFOCOM ’16.* 1–9. https://doi.org/10.1109/INFOCOM.2016.7524416