Abstract:

The configuration balancing problem with stochastic requests generalizes well-studied resource allocation problems such as load balancing and virtual circuit routing. There are given m resources and n requests; each request has multiple possible configurations, each of which increases the load of each resource by some amount. The goal is to select one configuration for each request to minimize the makespan: the load of the most-loaded resource. In the stochastic setting, the amount by which a configuration increases the resource load is uncertain until the configuration is chosen, but we are given a probability distribution.

We develop both offline and online algorithms for configuration balancing with stochastic requests. When the requests are known offline, we give a non-adaptive policy for configuration balancing with stochastic requests that $O(\frac{\log m}{\log \log m})$-approximates the optimal adaptive policy, which matches a known lower bound for the special case of load balancing on identical machines.

When requests arrive online in a list, we give a non-adaptive policy that is $O(\log m)$ competitive. Again, this result is asymptotically tight due to information-theoretic lower bounds for special cases (e.g., for load balancing on unrelated machines). Finally, we show how to leverage adaptivity in the special case of load balancing on related machines to obtain a constant-factor approximation offline and an $O(\log \log m)$-approximation online. A crucial technical ingredient in all of our results is a new structural characterization of the optimal adaptive policy that allows us to limit the correlations between its decisions.

This is joint work with Anupam Gupta, Nicole Megow, Ben Moseley, and Rudy Zhou.