In a call center, there is a natural trade-off between minimizing customer wait time and fairly dividing the workload amongst agents of different skill levels. The relevant control is the routing policy; that is, the decision concerning which agent should handle an arriving call when more than one agent is available. We formulate an optimization problem for a call center with multiple agent pools, that handle calls at a different speed, and a single customer class. The objective is to minimize steady-state expected customer wait time subject to a “fairness” constraint on the workload division.

The optimization problem we formulate is difficult to solve exactly. Therefore, we solve the diffusion control problem that arises in the many-server heavy-traffic QED limiting regime. The resulting routing policy is a threshold policy that prioritizes faster agents when the number of customers in the system exceeds some threshold level. We prove our proposed threshold routing policy is near-optimal as the number of agents increases, and the system load approaches its maximum processing capacity.

This is joint work with Amy Ward from the Marshall School of Business, University of Southern California.