Abstract

We study the relationship between capacity and performance for a service firm with spatial operations, in the sense that requests arrive with origin-destination pairs. An example of such a system is a ride-hailing platform in which each customer arrives in the system with the need to travel from an origin to a destination. We propose a state-dependent queueing model that captures spatial frictions as well as spatial economies of scale through the service rate. In a classical M/M/n queueing model, the square root safety (SRS) staffing rule is known to balance server utilization and customer wait times. By contrast, we find that the SRS rule does not lead to such a balance in spatial systems. In a spatial environment, pickup times increase the load in the system; furthermore, they are an endogenous source of extra workload that leads the system to only operate efficiently if there is sufficient imbalance between supply and demand. In heavy traffic, we derive the mapping from load to operating regimes and establish implications on various metrics of interest. In particular, to obtain a balance of utilization and wait times, the service firm should use a higher safety factor, proportional to the offered load to the power of 2/3. We also discuss implications of these results for general systems.