Staffing, Routing, and Payment to Trade Off Speed and Quality in Large Service Systems

Abstract

Most common queueing models used for service system design assume the servers work at fixed (possibly heterogeneous) rates. However, real-life service systems are staffed by people, and people may change their service speed in response to incentives. The delicacy is that the resulting service speed is jointly affected by staffing, routing, and payment decisions. Our objective in the paper is to find a joint staffing, routing, and payment policy that induces optimal service system performance.

We do this under the assumption that there is a trade-off between service speed and quality, and employees are paid based on both. The employees each selfishly choose their own service speed in order to maximize their own expected utility (which depends on the staffing through their busy time). The endogenous service rate assumption leads to a centralized control problem in which the system manager jointly optimizes over the staffing, routing, and service rate. By solving the centralized control problem under fluid scaling, we find four different economically optimal operating regimes—critically loaded, efficiency-driven, quality-driven, and intentional idling (in which there is simultaneous customer abandonment and server idling). Then, we show that a simple piece-rate payment scheme can be used to solve the associated decentralized control problem under fluid scaling.