STAFFING SERVICE SYSTEMS WITH LOAD DEPENDENT SERVICE RATE

J. Dong, Columbia University, US, jd2736@columbia.edu
P. Feldman, University of California Berkeley, US, feldman@haas.berkeley.edu
G.B. Yom-Tov, Technion - Israel Institute of Technology, Israel, gality@tx.technion.ac.il

Most operations management literature assumes that service times are independent of the load of the system. However, empirical evidence suggests that the two are correlated. Several factors could contribute to the correlation. During heavily loaded intervals, fatigue may cause agents to slow down, while pressure may cause them to speed up. On the customer side, correlation between load and service time is well established. For example, patients' condition may worsen if treatment is delayed in health care facilities, resulting in longer stays. Speedup by itself will not worsen performance (measured by delays and abandonments), while slowdown may not only decrease customer service level but also increase agents' workload dramatically. Hence, we concentrate on the latter, and examine how the dependence between service rate and work load affects the operational performance of the system. We do that by developing and analyzing fluid and diffusion approximations of an Erlang-A model with load-dependent service times. We propose methods that help stabilize and improve system performance. We show that if load sensitivity is moderate, a specific correction to the square-root staffing formula is required to achieve good service levels in the QED regime, while if load sensitivity is large such a method is not adequate, and other interventions are preferable.