Service Level Differentiation in Call Centers with Fully Flexible Servers

Mor Armony\textsuperscript{1} \hspace{1cm} Itay Gurvich\textsuperscript{2} \hspace{1cm} Avishai Mandelbaum\textsuperscript{3}

Abstract

We study large-scale service systems with multiple customer classes and many statistically identical servers. The following question is addressed: How many servers are required (staffing) and how does one match them with customers (control) in order to minimize staffing cost, subject to class level quality of service constraints? We tackle this question by characterizing scheduling and staffing schemes that are asymptotically optimal in the limit, as system load grows to infinity. The asymptotic regimes considered are consistent with the Efficiency Driven (ED), Quality Driven (QD) and Quality and Efficiency Driven (QED) regimes, first introduced in the context of a single class service system.

Our main findings are: a) Decoupling of staffing and control, namely (i) Staffing disregards the multi-class nature of the system and is analogous to the staffing of a single class system with the same aggregate demand and a single global quality of service constraint, and (ii) Class level service differentiation is obtained by using a simple Idle server based Threshold-Priority (ITP) control (with state-independent thresholds), b) Robustness of the staffing and control rules: Our proposed Single-Class Staffing (SCS) rule and ITP control are approximately optimal under various problem formulations and model assumptions. Particularly, although our solution is shown to be asymptotically optimal for large systems, we numerically demonstrate that it performs well also for relatively small systems.

\textsuperscript{1}Stern School of Business, New York University, marmony@stern.nyu.edu
\textsuperscript{2}Graduate School of Business, Columbia University, ig2126@columbia.edu
\textsuperscript{3}Faculty of Industrial Engineering and Management, Technion Institute of Technology, avim@ie.technion.ac.il