

# Dynamic routing in large-scale service systems with heterogeneous servers

Mor Armony  
Stern School of Business  
New York University  
44 West Fourth Street #8-62  
New York, NY 10012  
marmony@stern.nyu.edu  
Phone: 1 212 998 0291  
Fax: 1 212 995 4227

## Abstract

Motivated by modern call centers, we consider large-scale service systems with multiple server pools and a single customer class. For such systems, we propose a simple routing rule which asymptotically minimizes the steady-state queue length and virtual waiting time. The proposed routing scheme is FSF which assigns customers to the Fastest Servers First. The asymptotic regime considered is the Halfin-Whitt many-server heavy-traffic regime, which we refer to as the Quality and Efficiency Driven (QED) regime; it achieves high levels of both service quality and system efficiency by carefully balancing between the two. Additionally, expressions are provided for system limiting performance measures based on diffusion approximations. Our analysis shows that in the QED regime this heterogeneous server system outperforms its homogeneous server counterpart.