A FLUID MODEL FOR OVERLOADED QUEUES WITH SCORING-BASED PRIORITY RULES

Yichuan Ding, Sauder School of Business, University of British Columbia, BC, Canada, daniel.ding@sauder.ubc.ca
Peter Glynn, Department of Management Science and Engineering, Stanford University, CA, USA glynn@stanford.edu

We consider a queueing system with multitype customers and servers. Whenever a server attempts to deliver service, each customer is assigned a score that depends on customer type, duration of waiting time, and server type. Service is then provided to the customer with the highest score. We characterize the behavior of such a system using a fluid limit process, which has two important features: (1) the service rate in the transient state coincides with the max-flow of a parameterized network, so can be efficiently computed using the so-called GGT algorithm; (2) the service rate at the steady state coincides with the minimal-cost max-flow of a capacitated network, so can be computed within polynomial time. Thanks to these properties, we could compute the transient dynamics as well as the stationary state of the fluid limit process efficiently, and predict the performance of the system when a scoring policy has been implemented. As a byproduct, our method can determine whether a service network allows global first-come-first-serve (FCFS), a question raised by Talreja and Whitt(2008). We illustrate the application of our model in the context of cadaver kidney allocation. In particular, the fluid model we developed can predict the steady-state allocation outcome of the scoring policy proposed by the United Network of Organ Sharing (UNOS) in 2008.