Controlling Excessive Delays in Service Systems with Time-Varying Demand

Yunan Liu, NC State University

Abstract: Queueing theory is a field driven by applications. But unfortunately, there still remains a large gap between tractable theoretical studies and practical applications, such as call centers and health care systems, which have many realistic features (e.g., time-varying arrivals, customer abandonment, non-exponential distributions, and complicated network structures). In response to the challenge, we study a general G_t/GI/s_t+GI queueing model, which has a non-stationary non-Poisson arrival process (the G_t), non-exponential service times (the first GI), and allows customer abandonment according to a non-exponential patience distribution (the +GI). To bridge the gap between mathematical tractability and model applicability, we develop fundamental principles and optimal control policies for such a general queueing model.

Analytic formulas are developed to set the time-dependent number of servers in order to stabilize important service-level indicators, including: mean customer delay, probability of abandonment, and tail probability of delay (TPoD). Taking the TPoD for example: for any delay target w > 0 and probability target 0 < alpha < 1, we determine appropriate time-dependent staffing levels (the s_t) so that the time-varying probability that the waiting time exceeds a maximum acceptable value w is stabilized at alpha at all times. In addition, effective approximating formulas are provided for other important performance functions such as the probabilities of delay and abandonment, and the means of delay and queue length. Many-server heavy-traffic limit theorems in the efficiency-driven regime are developed to show that (i) the proposed staffing function achieves the goal asymptotically as the scale increases, and (ii) the proposed approximating formulas for other performance measures are asymptotically accurate as the scale increases. Extensive simulations show that both the staffing functions and the performance approximations are effective, even for smaller systems having an average of 3 servers.

Mini-bio:

Yunan Liu is an assistant professor at the Industrial and Systems Engineering Department and an associate faculty member of the Operations Research Center of North Carolina State University. His research interests include queueing theory, stochastic modeling, applied probability, simulation, and their applications in service systems including call centers, healthcare, and manufacturing systems. He received his M.S. and Ph.D. in Operations Research from Columbia University and B.S. in Electrical Engineering from Tsinghua University.