Multi-Server Queues with Time-Varying Arrival Rates

Ward Whitt, Columbia University

Public Abstract

The research objective of this project is to develop methods to describe and control the level of congestion (e.g., customer waiting times and queue lengths) in multi-server stochastic queueing models with time-varying arrival rates. The models will also include the important features of customer abandonment, non-exponential service and patience distributions, and network structure. The research will focus on models in which the system alternates between periods of overloading and underloading, as seen when the staffing levels cannot be adjusted flexibly in response to rapidly changing arrival rates. This research focuses on approximating deterministic time-varying fluid models, yielding approximations for the time-dependent mean values of queue lengths and waiting times as a function of the model parameters and the initial conditions. Then the research will consider refined Gaussian approximations for the full probability distributions. The approximations will be substantiated by simulation experiments and stochastic process limits.

If successful, the results of this research will provide the basis for new methods and tools to design and manage large-scale service systems, such as healthcare systems and customer contact centers, for which there is a recognized need for new methods to improve capacity planning and operational control. The models are also relevant for many other applications, e.g., web server farms processing web-page requests, public housing authorities providing apartments to low-income tenants and financial back offices processing loan applications. Successful results can contribute to better ways to select capacity levels in design decisions and better ways to select staffing levels and to route and schedule customers in operational decisions. Nonstandard methods for treating time-varying arrivals are especially important when service times are relatively long, as in healthcare.