To Pool or Not to Pool: Queueing Design for Large-Scale Service Systems

Ping Cao
School of Management
University of Science and Technology of China

January 24, 2019

Abstract: There are two basic queue structures commonly adopted in service systems: the pooled structure where waiting customers are organized into a single queue served by a group of servers and the dedicated structure where each server has her own queue. Although the pooled structure, known to minimize the servers idle times, is widely used in large-scale service systems, this study reveals that the dedicated structure, along with the join-the-shortest-queue routing policy, could be more advantageous for improving some service levels, such as the probability of a customers waiting time being within a delay target. The servers additional idleness resulted from the dedicated structure will be negligible when the system has many servers. Using a fluid model substantiated by asymptotic analysis, we provide a performance comparison between the two structures for a moderately overloaded queueing system with customer abandonment. We intend to help service system designers answer the following questions: To achieve a specified service level, which queue structure will be more cost-effective? How many servers can be saved by converting one structure into the other? Aside from structure design, our results are also of practical value for performance analysis and staffing deployment.

Talk will take place from 2:30PM - 3:30PM, Watt Family Innovation Center, Room 203.

Bio: Dr. Ping Cao is currently an associate professor in the School of Management at University of Science and Technology of China. He received his Ph.D. in Operational Research at Academy of Mathematics and Systems Science, Chinese Academy of Science in 2011. His research interests include stochastic control, queueing theory, Markov decision process, and revenue management. His work has been published in journals such as SIAM Journal on Control and Optimization, IEEE Transactions on Automatic Control, European Journal of Operations Research and Queueing Systems.