

Optimal Single-server and Multi-server Scheduling

Mor Harchol-Balter, Bruce J. Nelson Professor of Computer Science at Carnegie Mellon University

This talk discusses optimal stochastic scheduling in the context of a single-server queue (M/G/1) and an n-server system (M/G/n). We primarily focus on the case where the job size distribution is heavy-tailed, as is representative of computing workloads. We consider both the optimal scheduling policy for minimizing mean response time and for minimizing the asymptotic tail of response time.

For single-server systems, the optimal scheduling policies are now well understood, and we review these. However for multiserver systems, very little is known about optimal scheduling. In this talk, we will present *the first* results on tail-optimal scheduling for multi-server systems. Our central finding is that the multi-server case is intrinsically different from the single-server case. We find that policies that are strongly tail optimal for a single-server system, are only weakly tail optimal for a multi-server system. Furthermore, we derive new scheduling policies that are provably strongly tail optimal for multi-server systems, while being suboptimal for single-server systems.



Mor Harchol-Balter is the Bruce J. Nelson Professor of Computer Science at Carnegie Mellon. She received her Ph.D. from U.C. Berkeley in 1996, joined CMU in 1999, and served as the Head of the PhD program from 2008-2011. She is the SIG Chair for ACM SIGMETRICS as well as a Fellow of both ACM and IEEE. She has received several teaching awards, including the Herbert A. Simon Award and Spira Teaching Award. Mor's work focuses on designing new resource allocation policies, including load balancing policies, power management policies, and scheduling policies, for distributed systems. Mor is heavily involved in the SIGMETRICS / PERFORMANCE / INFORMS research community where her papers have received many awards (ValueTools 24, INFORMS George Nicholson Prize 22, SIGMETRICS 21, SIGMETRICS 19, PERFORMANCE 18, INFORMS

APS 18, EUROSYS 16, MASCOTS 16, MICRO 10, SIGMETRICS 03, ITC 03, SIGMETRICS 96). She is the author of two popular textbooks, both published by Cambridge University Press: *Performance Analysis and Design of Computer Systems: Queueing theory in action* (2013), which bridges Operations Research and Computer Science, and *Introduction to Probability for Computing* (2024). Mor is proudest of her many PhD students, almost all of whom are professors at top academic institutions.