Abstract:
Maximizing computer system performance relies on careful resource management: how to best allocate resources among jobs. Effective resource allocation is most difficult in regimes with uncertainty. This talk examines three common types of uncertainty. We consider uncertainty in job sizes and ask how to optimally schedule jobs to minimize response time in such regimes. We next turn to uncertainty in the arrival rate and ask how we should adapt capacity provisioning and power management in data centers to handle unexpected load fluctuations. Finally, we consider uncertainty in the system state and look at how job replication can help curtail unpredictability. A common thread in this talk is stochastic performance modeling and the insights it illuminates.

Special Note:
https://cse.ucsd.edu/about/stochastic-resource-management-face-uncertainty