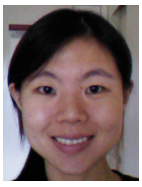
Cornell University $\mathbb{I}$*nformation,* $\mathbb{S}$*ystems, and* $\mathbb{N}$*etworks seminar*

Operating a Cloud: Optimal Switching of a Parallel Queue

Rhodes Hall 655: May 1, 2013 @ 12:00PM



ISN Seminar Speaker:

Xiaoxuan Zhang

IBM T. J. Watson Research Center

Abstract

We study the stochastic control of a cloud computing facility modeled by an M/M/∞ queue with holding, running and switching costs. The goal is to minimize average costs per unit time. The main result is that an average-optimal policy either always runs the system or is an (M, N)-policy defined by two thresholds M and N, such that the system is switched on upon an arrival epoch when the system size accumulates to N and it is switched off upon a departure epoch when the system size decreases to M. It is shown that this optimization problem can be reduced to a problem with a finite numbers of states and actions, and an average-optimal policy can be computed via linear programming. An example, in which the optimal (M, N)-policy outperforms the best (0, N)-policy, is provided. Thus, unlike the case of single-server queues studied in the literature, (0, N)-policies may not be average-optimal.

Biography

Xiaoxuan Zhang is a Postdoctoral researcher in Process and Energy Analytics group at Business Analytics and Mathematical Science Department of IBM T.J. Watson Research Center since January 2011. Her research interests are on stochastic control, stochastic optimization, and revenue management and pricing in retail supply chains and smart grids. She received her Ph.D. in Operations Research at Applied Mathematics and Statistics department from Stony Brook University in 2010, and a B.S. in Mathematics from Nanjing University, China in 2005.