Dispatching to Incentivize Fast Service in Multi-server Queues

Sherwin Doroudi
California Institute of Technology

Ragavendran Gopalakrishnan and Adam Wierman
Carnegie Mellon University

MOTIVATION
Classic scheduling theory fixes arrival and service rates. But in reality, incentives are often involved:
- People choose whether or not to join a queue.
- Servers choose their service rate.
Examples: Call centers, Enterprise data centers

DOES IT MATTER?
- Fastest Server First (FSF) dispatching is optimal for fixed service rates, but incentivizes slow service.
- The opposite, SSF, incentivizes fast service, but there are oscillations (no equilibrium).

RESULTS (2-server, rate-based policies)
- For a given value of k, if a symmetric Nash equilibrium exists, it is unique, and the corresponding response time is increasing in k.
- There exists a bounded interval for k outside which no symmetric Nash equilibria exist.
- A symmetric Nash equilibrium always exists for k = 0, -1, -2.

RESULT (idle-time-based policies)
- The class of idle-time-based policies is a special case of the class of rate-based policies with k = 0.

FUTURE WORK
(beyond 2 servers)
Challenge: In an m-server model, no closed form for idle-time exists for a general m.
Approximate Analysis: When m is finite but large, mean field analysis could be applied.

Support provided by: Social and Information Sciences Laboratory (SISL) and National Science Foundation (NSF)