Statistical Machine Learning Makes Automatic Control Practical for Internet Datacenters

Peter Bodik, Rean Griffith, Charles Sutton
Armando Fox, Michael Jordan, David Patterson
RAD Lab, UC Berkeley

Goal: optimal resource allocation

- Accurate, optimal resource allocation for web apps
  - monetary incentive for the cloud user
  - but can’t violate performance SLA
  - EC2 offers auto-scaling, but only reactive
    - e.g.: add more machines if CPU utilization > 70%

- Previous work: simple performance models trained offline
  - linear or simple queuing models
  - don’t adapt to changes in application performance
  - inaccurate model implies inaccurate control

- Approach
  - use statistical methods for accurate modeling and control

Statistical methods for modeling and control

- Accurate performance model
  - model end-to-end application latency and its variance
  - based on production data
  - [ACDC ’09] Active Exploration of Application Performance Regimes

- Adapting to changes in application performance
  - slowly adapt model to gradual changes
  - detect abrupt changes in performance using change-point detection
    - retrain model from scratch

- Control policy simulator
  - use simulator to evaluate different control policies
    - simulate application performance using its model
  - find control parameters that optimize particular cost function