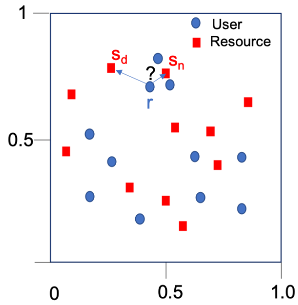
Resource Allocation in One-dimensional Distributed Service networks

Nitish K. Panigrahy (Umass), **Prithwish Basu** (Raytheon BBN), **Philippe Nain** (Inria), **Don Towsley** (Umass), **Ananthram Swami** (ARL), **Kevin S. Chan** (ARL) and **Kin K. Leung** (Imperial).

Objectives

Goal: Characterize cost of serving a request in a distributed data analytics network



Approach

- Consider distributed analytics network in one dimension
  - R: Set of users requesting service
    - λ: User Density
    - F_Y: Inter-user distance distribution; Variance: σ^2_Y
  - S: Set of resources providing service
    - μ: Resource Density
    - F_X: Inter-resource distance distribution; Variance: σ^2_X
    - C_j: Capacity of server j

Technical Challenges

- Map to queuing theoretic models
  - Exceptional service and accessible batches
  - Solution of new queueing theoretic models
- Heterogeneous resource capacities; coalition constraints

Military & Coalition Relevance

- Distributed analytics at tactical edge
  - User requests execution of analytic service
  - Resources: Agile code + Data

Unidirectional Matching

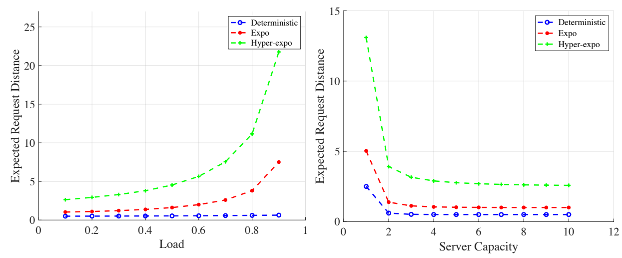
- Allocate users to available resource to its right
- Homogeneous resource capacity ($C_j = c$)
  - F_X ~ General and F_Y ~ Expo (*PRGS*)
    - » Analysis through supplementary variable technique
    - » $\mathbb{E}[D] = \frac{\mathbb{E}[NX]}{\lambda}$
    - » $\mathbb{E}[NX]$: Obtained by solving Chapman-Kolmogorov Equations
  - F_X ~ Expo and F_Y ~ General (*GRPS*)
    - » $\mathbb{E}[D] = \frac{C}{\lambda\mu(1-r_0^c)(1-r_0)} + \frac{1}{\mu}$
    - » C : Normalization constant
- Heterogeneous resource capacity
  - Analysis through generating function approach
  - $\mathbb{E}[D] = \frac{1}{\rho}\left[\frac{\mathbb{E}[NX]}{\mu} + \frac{\lambda}{2}\left(\sigma^2_X + \frac{1}{\mu^2}\right)\right]$
- Uncapacitated resource allocation
  - $\mathbb{E}[D] = \frac{\mu}{2}\left[\sigma^2_X + \frac{1}{\mu^2}\right]$ (*PRGS*)
  - $\mathbb{E}[D] = \frac{1}{\mu}$ (*GRPS*)

Bidirectional optimal matching

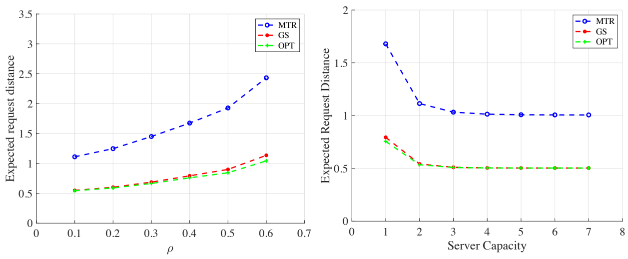
- Dynamic Programming Solution
  - $d(i,j) = \min\{|R[i] - S[j]| + d(i-1, j-1), d(i, j-1)\}$
  - Time Complexity: $O(R^2)$

Simulation Results

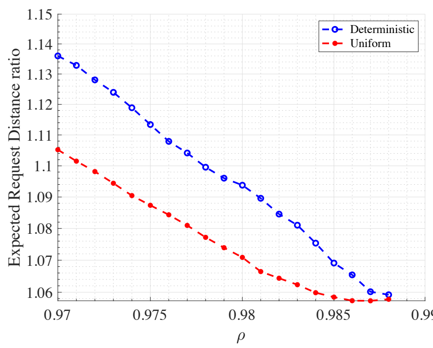
- Sensitivity Analysis:



- Comparison of allocation policies:



- Generally distributed users and resources:



Summary & Future Work

- Summary
  - Queueing models for resource allocation in spatial setting
  - Analyzed new queueing theoretic models for homogeneous and heterogeneous resource capacities
  - Proposed dynamic programming based solution for computing optimal bidirectional request distance
- Future Work
  - Extend work to 2-D setting
  - Cost models as functions of request distance

Publication(s) & Impact

- N. K. Panigrahy, P. Basu, P. Nain, D. Towsley, A. Swami, K. S. Chan and K. K. Leung, “Resource Allocation in One-dimensional Distributed Service Networks”, Accepted for publication in *IEEE MASCOTS*, Rennes, France, October 22-25, 2019.