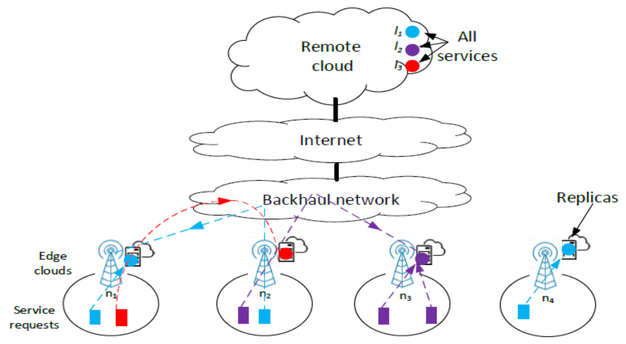
Service Placement and Request Scheduling for Data-intensive Applications in Edge Clouds

V. Farhadi (PSU), **F. Mehmeti** (PSU), **T. He** (PSU), **T. La Porta** (PSU), **H. Khamfroush** (University of Kentucky), **S. Wang** (IBM US), **K. Chan** (ARL), **S. Stein** (Southampton).

Objectives

- Serve requests from edge devices require fast response time in a dynamic environment
- Provide rapid, dynamic location and scheduling of resources to analytics tasks
- Accommodate realistic constraints on bandwidth and processing capabilities, and user mobility

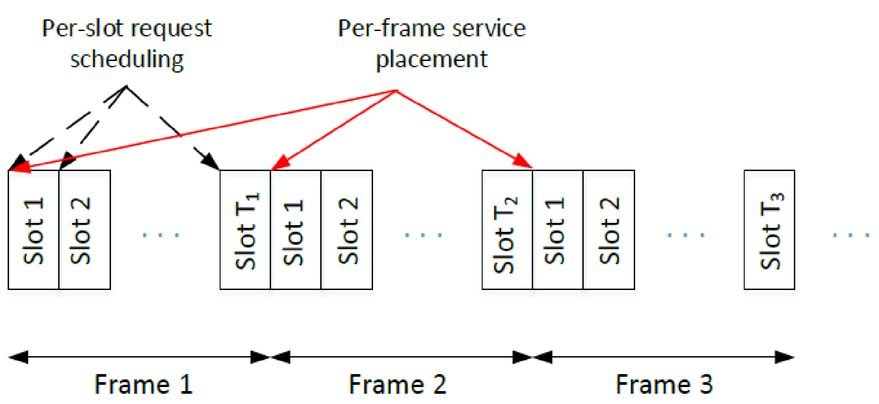


Technical Challenges

- Limited network connectivity between devices and edge cloud to upload data with requests
- Limited resources in the edge cloud with respect to memory, processing power and network connectivity
- Highly dynamic workloads and user mobility

Approaches

- Use edge cloud to process analytics requests submitted by users to provide low-latency service
- Assign resources to servers on a longer time-scale (frame), and schedule requests as they are received on a shorter time-scale (slots)
- Consider service placement prediction over multiple frames
- Use a greedy service placement algorithm, that under realistic conidiations is proven to have a guaranteed approximate performance related to the size of the service replicas
- Schedule tasks based on a Linear Program

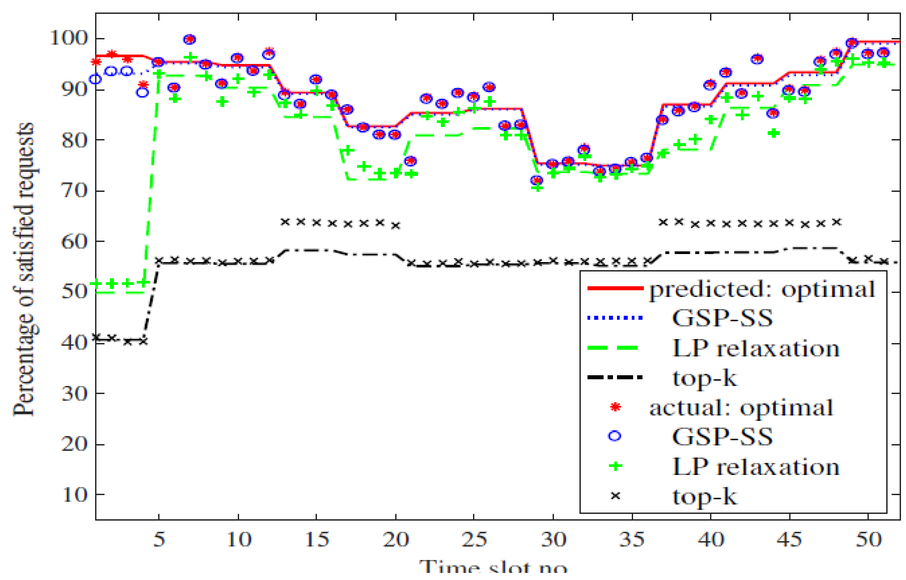


Military & Coalition Relevance

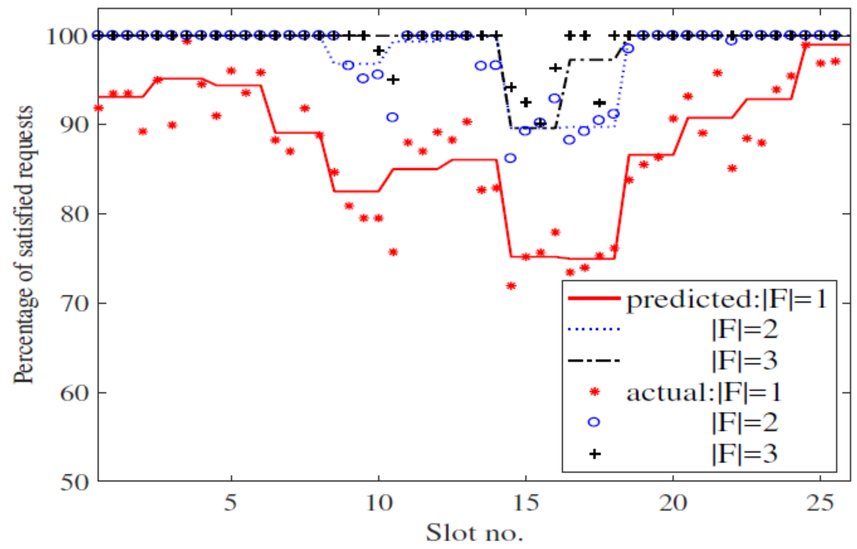
- Analytics requests from edge devices require fast response time in a dynamic environment
- Considers realistic constraints on bandwidth and processing capabilities, and user mobility
- Edge clouds with lower capabilities than core cloud services can be deployed in tactical networks

Results

- Cause of NP-Hardness is the presence of a reconfiguration budget and storage constraints
- Under certain reasonable assumptions the service placement objective function is monotone sub-modular and the constraints form a p-independent system leading to an approximation of 1/(1+p)
- The proposed algorithm achieves 90% of the optimal solution even when the approximation does not hold
- Exploiting multiple frames for placement improves performance and two-frame predictions are sufficient



- Single frame service placement
- Trace data for service requests and user mobility
- Difference between predicted & optimal performance due to shadow scheduling



- Multi-frame service placement
- Two frames provide most performance benefit

Summary & Future Work

- Our greedy service placement algorithm with optimal scheduling is suitable for tactical environments with user mobility and dynamic workloads
- Consider distributed approaches that support policy
- Consider elastic resources with deadlines
- Consider tasks that can be decomposed

Publication(s) & Impact

- V. Farhadi, F. Mehmeti, T.F. La Porta, T. He, H. Khamfroush, S. Wang, K.S. Chan, “Service Placement and Request Scheduling for Data-intensive Applications in Edge Clouds,” Proc. of IEEE INFOCOM, 2019.
- T. He, H. Khamfroush, S. Wang, S. Stein, T.F. La Porta, “It's Hard to Share: Joint Service Placement and Request Scheduling in Edge Clouds with Sharable and Non-sharable Resources,” Proc. of IEEE ICDCS, 2018.