Joint Service Placement and Request Routing in Mobile Edge Networks with Storage, Computation, and Communication Constraints

Konstantinos Poularakis (Yale), **Ian Taylor** (Cardiff), **Leandros Tassiulas** (Yale)

Objectives

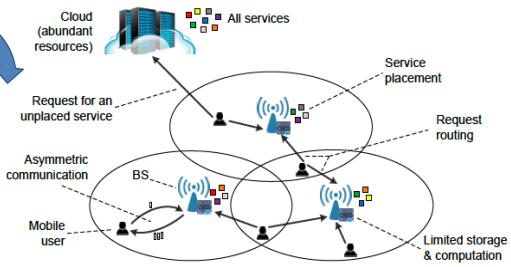
Modern tactical networks need to provide both computation and data intensive as well as latency sensitive services, often involving different coalition teams, that cannot be supported solely by existing **centralized** cloud platforms.

Distributed cloud architectures enable the execution of service tasks in edge servers, e.g., infrastructure nodes close to the mobile tactical units, therefore constituting a promising solution to expedite service delivery.

We address the problem of **jointly** optimizing the placement of services and the routing of requests to edge servers in a distributed cloud system with multidimensional (computation, storage, communication) constraints.

Technical Challenges

Services placed in Base Stations (BS)



Complex multi-cell environment:

- Ultra-dense BS deployments (~50 BSs per km^2)
- Users in range of multiple BSs (multiple routing options).

Multi-dimensional resource requirements of services:

- Not just allocation of **computation**
- But also placement of data (**storage**)
  - Database & visual recognition model for AR service

Bidirectional communication

- Both uplink and downlink bandwidth provisioning
  - Video streaming (only downlink)
  - Face Recognition (mainly uplink)
  - AR (both uplink and downlink)

Military & Coalition Relevance

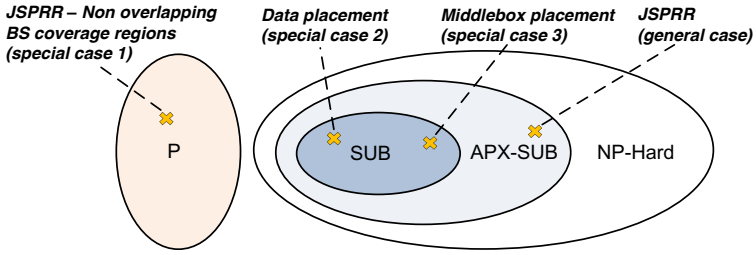
- The proposed methods of jointly optimizing service placement and request routing provide solutions to military scenarios, especially when multiple network paths exist among tactical units and it is critical to utilize network resources of different paths to meet the demand for low latency services.

Model & Complexity

Formulate the **Joint Service Placement and Request Routing (JSPRR)** as an integer programming problem.

NP-Hard as it generalizes the knapsack problem by adding multiple packing constraints.

Generalizes several problems in literature.



Randomized Rounding Approximation

Algorithm 1: Randomized Rounding algorithm

1 Solve the linear relaxation of JSPRR problem to obtain $(x^\dagger, y^\dagger)$ optimal solution
2 **for** $n \in \mathcal{N}$, $s \in \mathcal{S}$ **do**
3 Set $\widehat{x}_{ns} = 1$ with probability $x^\dagger_{ns}$ — *Placement variables rounding phase*
end
4 **for** $u \in \mathcal{U}$ **do**
5 Define $\mathcal{N}'_u = (n \in \mathcal{N}_u : \widehat{x}_{ns_u} > 0)$ and — *Candidate BSs for routing user u*
6 **if** $\mathcal{N}'_u = \emptyset$ **then**
7 set $\widehat{y}_{lu} = 1$ and $\widehat{y}_{nu} = 0 \ \forall n \in \mathcal{N}$ — *Route to the cloud*
else
8 set $\widehat{y}_{nu} = 1$, $n \in \mathcal{N}'_u$, with probability $\frac{y^\dagger_{nu}}{x^\dagger_{ns_u}}$, or
9 set $\widehat{y}_{lu} = 1$ with probability $\left[\frac{y^\dagger_{lu} - \prod_{n \in \mathcal{N}'_u}(1 - x^\dagger_{ns_u})}{1 - \prod_{n \in \mathcal{N}'_u}(1 - x_{ns_u})}\right]_+$ — *Route to a candidate BS or the cloud probabilistically*
end
end
10 Output $\widehat{x}, \widehat{y}$

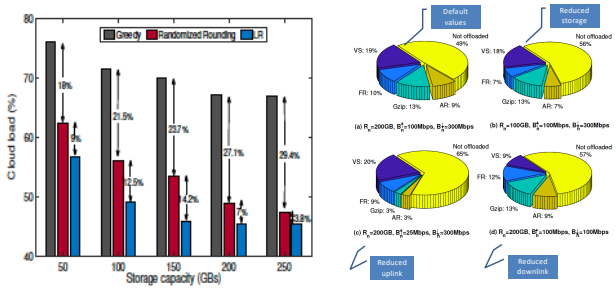
Guarantees on constraint violation

Theorem 1. The amount of data placed by the Randomized Rounding algorithm in BS n will not exceed its storage capacity by a factor of $\frac{3\ln(S)}{s_n} + 4$ with high probability.

Guarantees on performance

Theorem 3. The objective value returned by the Randomized Rounding algorithm is at most $\frac{2\ln(S)}{\xi} + 3$ times worse than optimal with high probability, where $\xi = \sum_u y_{lu}$.

Evaluation



✓ **Better than state-of-the-art Greedy placement schemes**

Summary & Future Work

- We studied the problem of finding the optimal placement and request routing policies for both data and computation intensive services. In future, we plan to explore dynamic versions of our algorithms that adapt their decisions over time in response to changes in demand and topology.

Publication(s) & Impact

- K. Poularakis, J. Llorca, A. Tulino, I. Taylor, L. Tassiulas, "Joint Service Placement and Request Routing in Multi-cell Mobile Edge Computing Networks", IEEE Infocom, 2019

Distributed Analytics and Information Science International Technology Alliance
Annual Fall Meeting, September 2019