Mobile Edge Cloud: Computation Scheduling and Caching

S M Shahrear Tanzil
Supervisor: Dr. Vikram Krishnamurthy
Committee Members: Dr. Lutz Lampe and Dr. Jane Wang

Department of Electrical and Computer Engineering
University of British Columbia, Vancouver, Canada

May 14, 2018

Mobile Edge Cloud

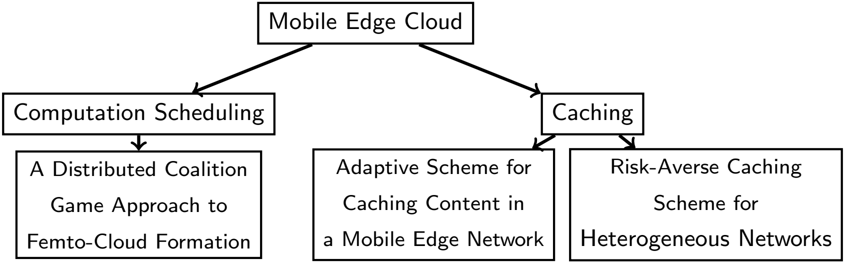

Figure 1: A schematic view of the contributions of the thesis.

Objective:

- Maximally exploiting edge network computation and caching resources
- Improved users' experience
- Improved network performance

Why Mobile Edge Cloud is Important?

- Computation
  - Major bottleneck in mobile cloud computing is latency[1]
  - Mobile edge computing is considered to be an important ingredients in 5G to support time critical applications[2]
- Caching
  - The expected data traffic served by cellular networks in 2020 is approximately 30.6 exabytes (10^{18}) per month[3]
  - Multiple downloading requests for a few popular video content account for the majority of the traffic[4]
  - Caching content within the wireless access network is considered to be an integral part of 5G to meet traffic demands with lower service latency[5]

[1] Barbera et al. (2013), in Proc. of the IEEE INFOCOM

[2] Ericsson (2017): 5G for latency-critical IoT applications

[3] Cisco white paper, 2016

[4] Jiang et al. (2016), IEEE Transactions on Mobile Computing

[5] Fadlallah et al. (2017), IEEE Communications Magazine

Computation and Caching at Mobile Edge Network

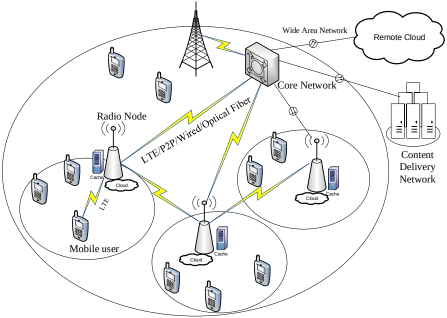

Figure 2: Network Architecture

- Computation and caching at edge network reduce network traffic between edge node and core network; and traffic between core network and cloud servers

Mobile Edge Cloud: Limitations

- Computation
  - Edge cloud resources are not maximally exploited[6]
  - Applicable for certain scenarios[7]
- Caching
  - Storage allocation and caching decisions are not addressed together[8]
  - Caching decisions and routing mechanisms are performed separately[9]
  - Content popularity prediction error is not considered[10]

[6] Munoz et al. (2014), IEEE Transactions on Vehicular Technology
[7] Jessica et al. (2015), in Proc. of the IEEE VTC
[8] Poularakis et al. (2016), IEEE Transactions on Wireless Communications
[9] Song et al. (2017), IEEE Transactions on Wireless Communications
[10] Bharath et al. (2016), IEEE Transactions on Communications

A Distributed Coalition Game Approach to Femto-Cloud Formation

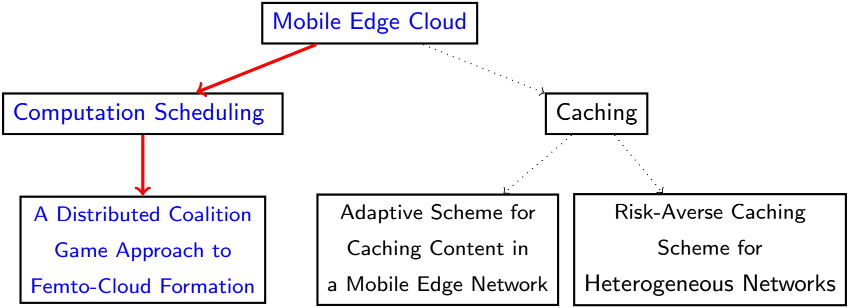

Figure 3: A schematic view of the contributions of the thesis.

Motivation

- Project TROPIC proposed Femto-Cloud architecture[11]
  - Excessive tasks are migrated to the remote cloud
    Limitation: Wide area network (WAN) latency
- How can we reduce WAN latency?
  Solution:
  - Maximally exploiting FAPs local resources
  - Form femto-clouds
- How to motivate FAP owners to share resources?
  Solution: Monetary incentives
- How should FAPs decide on formation of such femto-clouds in a distributed fashion?
  - Cooperative game theory is suitable for the framework

[11] TROPIC-D22, European Commission project

Main Contributions

- Femto-cloud formation problem is formulated as an incentive based coalition formation cooperative game with transferable utility, $U(\mathcal{C}) = U^r(\mathcal{C}) - U^c(\mathcal{C})$[12]
  where, $U^r(\mathcal{C})$ =revenue earned by the femto-cloud and $U^c(\mathcal{C})$= costs incurred by forming a femto-cloud
- The femto-cloud formation problem is then formulated as:

$$
\begin{aligned}
\max_{\mathcal{S}\in\mathcal{B}} \quad & \sum_{\mathcal{C}\in\mathcal{S}} \lfloor U(\mathcal{C}) \rfloor_{\Delta}, \\
\text{s.t.} \quad & r_k \in \mathcal{P}, \\
& \sum_{k\in\mathcal{C}} r_k = \lfloor U(\mathcal{C}) \rfloor_{\Delta}, \qquad \forall \mathcal{C} \in \mathcal{B}, \\
& \sum_{k\in\mathcal{C}'} r_k \geq \lfloor U(\mathcal{C}') \rfloor_{\Delta}, \quad \forall \mathcal{C}' \subseteq \mathcal{V}, \mathcal{C}' \neq \emptyset.
\end{aligned} \tag{1}
$$

where, $\mathcal{S}$: each femto-cloud structure; $\mathcal{B}$: set of all possible femto-cloud structures; $\mathcal{C}$: a coalition of FAPs; $\mathcal{V}$: set of FAPs; r : share of each FAP; and $\mathcal{P}$: demand of incentive.

- Present a distributed femto-cloud formation algorithm[13] that guarantees convergence to the solution of (1)

[12] $U(\mathcal{C}) = -U^c(\mathcal{C})$ for enterprise environment when minimization of latency is the main target
[13] Arnold et al. (2002), J. Econ. Behav. Organ

Main Contributions

- FAPs autonomously decide which femto-cloud to join, while maximizing their incentives
- Core of the game provides the solution to the femto-cloud formation problem
- Femto-clouds are formed for an interval
- Performance is evaluated for augmented reality application in femto-cloud using NS-3 simulator

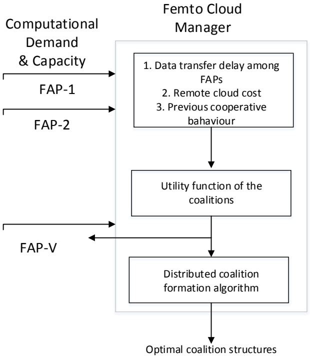

Figure 4: Schematic of femto-cloud formation

Result 1 Higher incentives to FAP owners

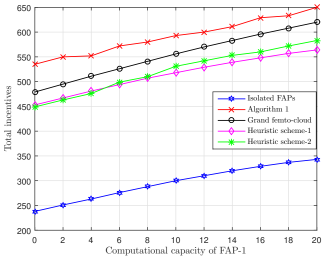

Figure 5: Computational capacity/resource of FAP-1 vs. femto-cloud incentive (Residential environment)

Result 2 Lower delay

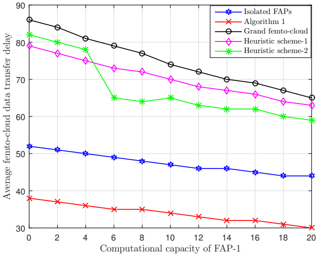

Figure 6: Computational capacity of FAP-1 vs. average data transfer delay per 2.2 Mb data in the femto-cloud (Enterprise environment)

- Femto-clouds resulted in reduced latency and higher incentives to the FAP owners
- Our approach can capture a wide range of scenarios, e.g., residential, and enterprise femtocell environments

Adaptive Scheme for Caching Content in a Mobile Edge Network

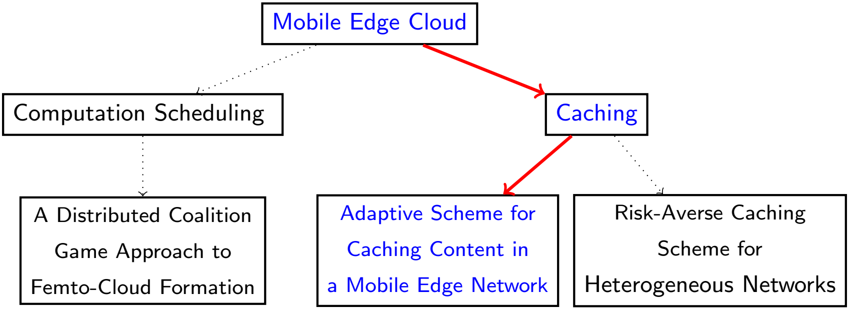

Figure 7: A schematic view of the contributions of the thesis.

Motivation

- Challenges
  - Which content are going to be popular?
    Solution: Estimate content popularity using learning methods
  - Given the content popularity how to cache content throughout the network?
    Solution: Linear programming/integer programming
  - Energy consumption due to caching should be addressed
    Solution: Consider storage/cache allocation in the network so that energy consumption remains below storage energy budget
- Objectives
  - Storage allocation and caching decisions should be performed simultaneously
  - Cache content before receiving any requests when the network load is minimal

Problem Formulation

- Problem is formulated as a mixed-integer linear programming (MILP)
- Objectives: minimize content downloading delay, initial file transferring cost and storage energy cost in the network
- Constraints: storage size and storage energy budget

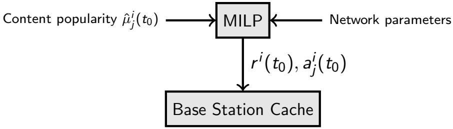

Figure 8: A schematic of the adaptive caching scheme.

$$\min \left(w_1 \sum_{j\in\mathcal{F}} \sum_{\substack{i\in\mathcal{V}\\ l\in\mathcal{V}}} f_j \hat{\mu}_j^l d^{il} b_j^{il} + w_2 \sum_{j\in\mathcal{F}} \sum_{i\in\mathcal{V}} f_j d^{gi} a_j^i + w_3 \sum_{i\in\mathcal{V}} r^i z_0 \right)$$

s.t.

$$\sum_{i\in\mathcal{V}} r^i s_0 \le S$$

$$\sum_{j\in\mathcal{F}} f_j a_j^i \le r^i s_0 \quad \forall i \in \mathcal{V} \quad s_0 \text{ depends on } z_0$$

Here, f : file size; $\hat{\mu}$: content popularity; d : downloading delay; z_0 : energy consumption index; a, b, r : decision variable; $\mathcal{V}$: set of radio nodes; $\mathcal{F}$: set of content, w_1, w_2, w_3 : weight of different factors; and S : cache size.

Risk-Averse Caching Scheme for Heterogeneous Networks

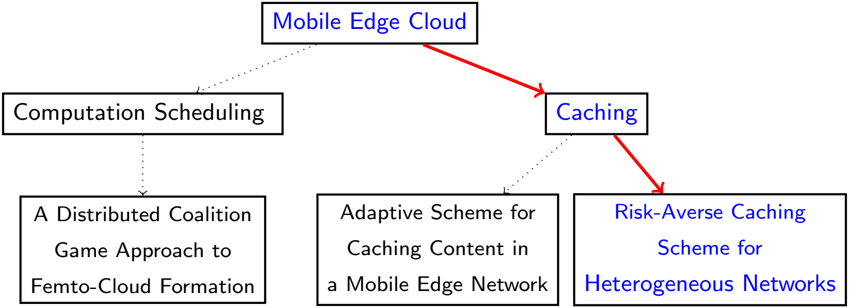

Figure 9: A schematic view of the contributions of the thesis.

Motivation & Problem Formulation

- Risk-neutral caching scheme
- Includes routing mechanism in the caching decision
- Objectives: minimize content downloading delay
- Constraints: link load and storage size

$$C^* \in \arg\min_{C,k,\delta,r} \left\{ \sum_{f=1}^{F} \sum_{d \in \mathcal{V}_d} \sum_{i,j \in \mathcal{V}} \hat{y}_{df} A_{ijf} \delta_{ijdf} \right\}$$

s.t.

$$\sum_{i \in \mathcal{V}} \delta_{sidf} - \delta_{isdf} = k_{sdf}, \quad \sum_{i \in \mathcal{V}} \delta_{didf} - \delta_{iddf} = -1,$$

$$\sum_{f=1}^{F} s_f c_{sf} \le S_s, \quad \sum_{s=1}^{V} c_{sf} \ge 1, \quad \sum_{f=1}^{F} \sum_{d \in \mathcal{V}_d} \delta_{ijdf} \le T_{ij}$$

$$r_{sdf} \le k_{sd}^{f}, \quad r_{sdf} \le c_{sf}, \quad r_{sdf} \ge k_{sdf} + c_{sf} - 1$$

- Risk-averse caching scheme
- Uncertainty associated with the predicted content requests
- Probabilistic guarantees can be made on the network operating characteristics

$$C^* \in \arg\min_{C,k,\delta,r,c} \left\{ c + \frac{1}{K(1-\alpha)} \sum_{\xi=1}^{K} \xi_k \right\}$$

s.t.

$$\xi_k \ge \sum_{f=1}^{F} \sum_{d \in \mathcal{V}_d} \sum_{i,j \in \mathcal{V}} \hat{y}_{df} A_{ijf} \delta_{ijdf} - c,$$

other constraints

Here, s_f : file size; $\hat{y}$: content requests; A : edge weight; α : confidence level; C, k, δ, r, c : decision variable; $\mathcal{V}$: set of radio nodes; $\mathcal{V}_d$: users' connected nodes; $\mathcal{F}$: set of content; T : link loads, and S_s : cache size; K : number of samples.

- Performance is evaluated using real world Youtube data and NS-3 simulator

Result 1 Network traffic comparison

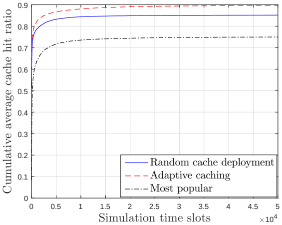

Figure 10: Cumulative average cache hit ratio in the network vs. number of time slots (Network performance: Adaptive Caching)

Result 2 Downloading delay comparison

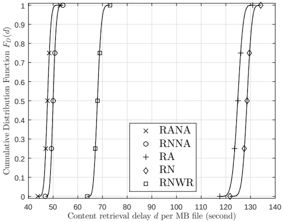

Figure 11: The cumulative distribution function $F_D(d)$ of the content retrieval delay for different caching schemes.

- The presented caching method resulted in reduced downloading delay

Summary

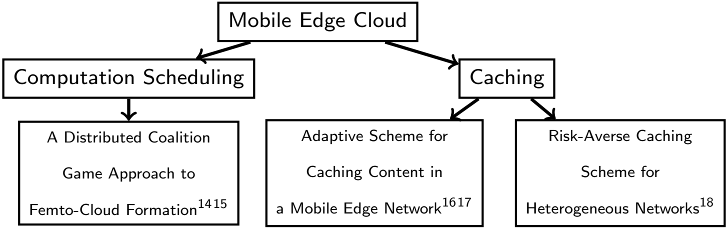

Figure 12: A schematic view of the contributions of the thesis.

---

[14] A Distributed Coalition Game Approach to Femto-Cloud Formation, IEEE Transactions on Cloud Computing (accepted)

[15] Femto-Cloud Formation: A Coalitional Game-Theoretic Approach, IEEE Global Communication Conference, San Diego, CA, 2015, pp. 1-6

[16] Adaptive Scheme for Caching YouTube Content in a Cellular Network: Machine Learning Approach, IEEE Access, Vol. 5, pp. 5870-5881, 2017

[17] Systems and Methods for Caching (A non-provisional patent application has been filed by Huawei)

[18] Risk-Averse Caching Policies for YouTube Content in Femtocell Networks using Density Forecasting, IEEE Transactions on Cloud Computing (minor revision)


Thank
You!!!
www.thebodytransformation.com