(**GlobeCom 2020**)

Mobile Edge Computing (MEC) Network Control: Tradeoff Between Delay and Cost

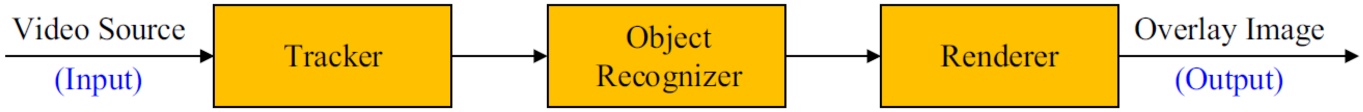
Yang Cai, Jaime Llorca, Antonia M. Tulino, Andreas F. Molisch

University of Southern California

New York University

Background

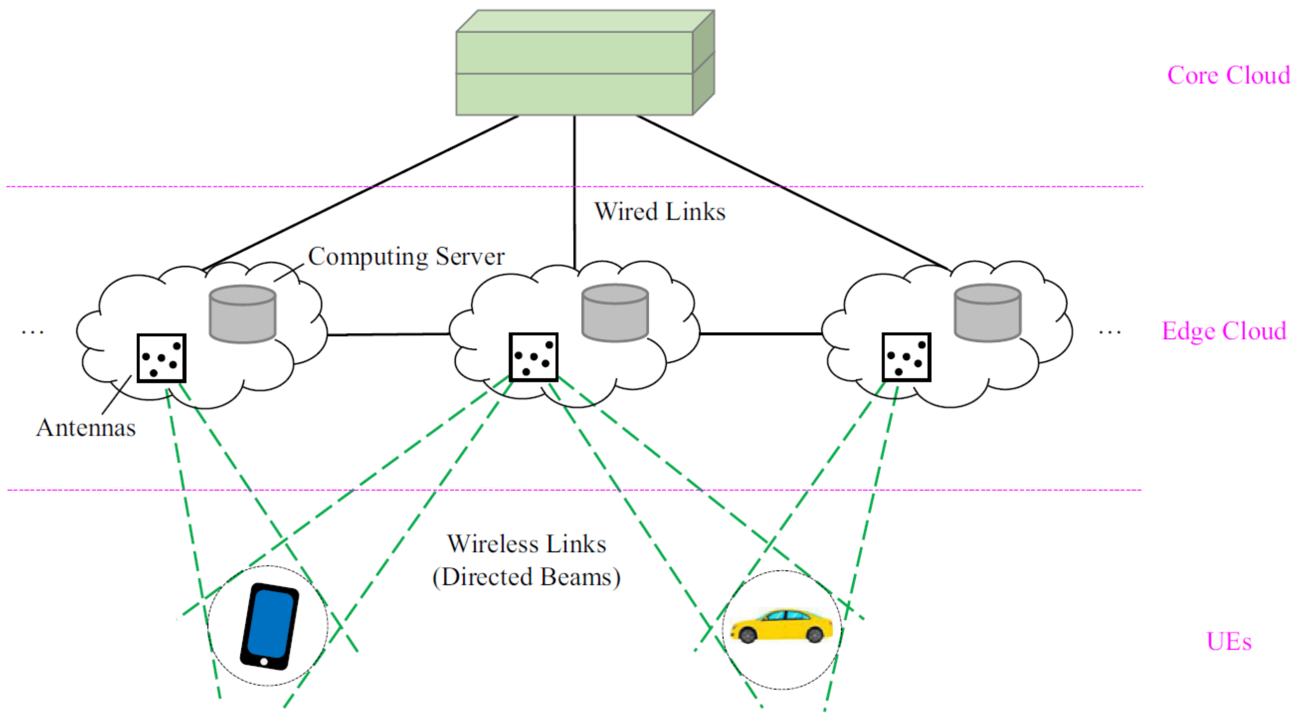
- Augmented Information (AgI) Services
  - Communication + Computation
  - The example of *augmented reality*



  - Not suitable to complete the computation tasks at user equipment (UE)
    - Why: restricted computation capability + limited power
    - How: offload the tasks to cloud networks

Background

- Mobile Edge Computing network
  - Computation resource -> end user



Background

- Related Problems
  - Task offloading
  - Packet routing and scheduling
  - Resource allocation

Each individual problem is difficult
Joint optimization is more complicated

- Performance Metrics: *average delay* and *resource cost*
  - In general, there is a tradeoff
    - Better delay -> data-center in proximity -> can be expensive
    - Cheap network location -> can be remote -> excessive delay
  - Goal of this work: to design a control policy that trades off the two metrics

System Model

- Cloud Network
  - Nodes: AP & UE
    - Computation resource choice $k_i(t)$: computation capability $C_{k_i(t)}$, config cost $s_{k_i(t)}$
  - Wired links between APs
    - Transmission resource choice $k_{ij}(t)$: transmission capability $C_{k_{ij}(t)}$, config cost $s_{k_{ij}(t)}$
  - Wireless links between AP and UE

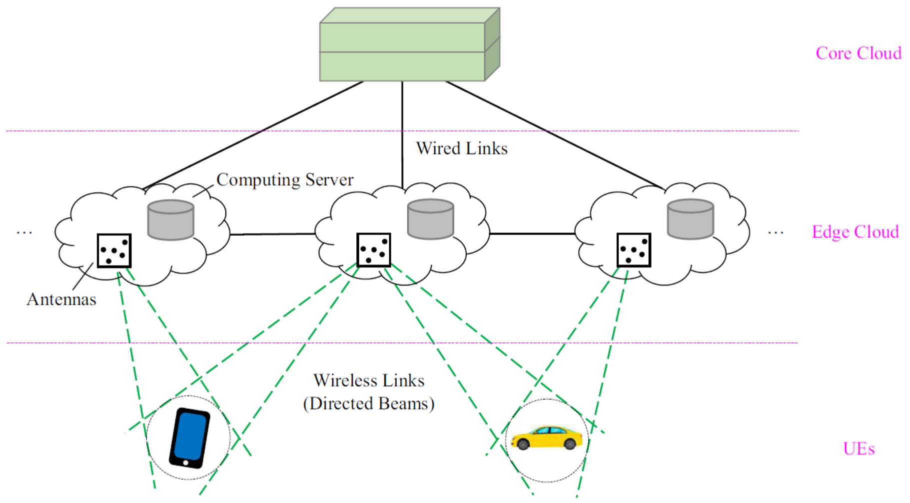
$$R_{ij}(t) = \left(\frac{B}{F}\right) x_{ij}(t) \log_2\left(1 + \frac{g_{ij}(t)p_{ij}(t)}{\sigma_{ij}^2}\right)$$

- Downlink (AP->UE) : beamforming
- Uplink (UE->AP): 1 to 1 communication

$$\sum_{j\in\tilde{\delta}_i^+} x_{ij}(t) \leq 1, \quad \forall i \in \mathcal{V}_{\mathrm{a}}$$

- Transmission power constraints

$$\sum_{j\in\tilde{\delta}_i^+} p_{ij}(t) \leq P_i, \quad \forall i \in \mathcal{V}$$



System Model

- Service Function Chain
  - AgI Service ϕ = Function 1 + … + Function m + … + Function M_ϕ

  - Parameter of function m: scaling factor $\xi_\phi^{(m)}$, workload $r_\phi^{(m)}$
  - Commodity (u, ϕ, m) (to distinguish the packets)
    - destination node u
    - requested service ϕ
    - current stage m

Queuing System

- Queues and Flow Variables
  - Queues $Q_i^{(u,\phi,m)}(t)$ for different commodities (u, ϕ, m)
  - Flow variable
    - Processing flow $\mu_{i,\text{pr}}^{(u,\phi,m)}(t)$
    - Transmission flow $\mu_{ij}^{(u,\phi,m)}(t)$
- Queuing dynamics

$$Q_i^{(u,\phi,m)}(t+1) \le \max\left\{0,\ Q_i^{(u,\phi,m)}(t) - \mu_{i,\text{pr}}^{(u,\phi,m)}(t) - \sum_{j\in\delta_i^+} \mu_{ij}^{(u,\phi,m)}(t)\right\} + \boxed{\mu_{\text{pr},i}^{(u,\phi,m)}(t)} + \sum_{j\in\delta_i^-} \mu_{ji}^{(u,\phi,m)}(t) + a_i^{(u,\phi,m)}(t)$$

Flow received from computation $\mu_{\text{pr},i}^{(u,\phi,m+1)}(t) = \xi_\phi^{(m)} \mu_{i,\text{pr}}^{(u,\phi,m)}(t)$

Studied Problem

$$h_1(t) = \sum_{i\in\mathcal{V}} \left[s_{k_i(t)} + c_{\mathrm{pr},i} \sum_{(u,\phi,m)} r_\phi^{(m)} \mu_{i,\mathrm{pr}}^{(u,\phi,m)}(t) \right] + \sum_{(i,j)\in\mathcal{E}_\mathrm{b}} \left[s_{k_{ij}(t)} + c_{\mathrm{tr},ij} \sum_{(u,\phi,m)} \mu_{ij}^{(u,\phi,m)}(t) \right] + \sum_{i\in\mathcal{V}} c_{\mathrm{wt},i} \sum_{j\in\tilde{\delta}_i^+} x_{ij}(t) p_{ij}(t) \tau, \tag{9}$$

$$\min \quad \overline{h_1} \qquad \text{(Goal 1: resource cost)}$$

$$\text{s. t.} \quad \overline{h_2} = \boldsymbol{\kappa}^\mathrm{T} \overline{\{\boldsymbol{Q}(t)\}} < \infty, \qquad \text{(Goal 2: average delay)}$$

$$\boldsymbol{\mu}(t) \ge 0,$$

$$\mu_{\mathrm{pr},i}^{(u,\phi,m+1)}(t) = \xi_\phi^{(m)} \mu_{i,\mathrm{pr}}^{(u,\phi,m)}(t), \quad \forall i \in \mathcal{V},$$

Capacity constraint

$$\sum_{(u,\phi,m)} \mu_{i,\mathrm{pr}}^{(u,\phi,m)}(t) r_\phi^{(m)} \le C_{k_i(t)}, \quad \forall i \in \mathcal{V}$$

$$\sum_{(u,\phi,m)} \mu_{ij}^{(u,\phi,m)}(t) \le \begin{cases} C_{k_{ij}(t)} & \forall (i,j) \in \mathcal{E}_\mathrm{b} \\ \boxed{R_{ij}(t)}\tau & \forall (i,j) \in \mathcal{E}_\mathrm{a} \end{cases}$$

$$R_{ij}(t) = \left(\frac{B}{F}\right) x_{ij}(t) \log_2 \left(1 + \frac{g_{ij}(t) p_{ij}(t)}{\sigma_{ij}^2}\right)$$

$$\sum_{j\in\tilde{\delta}_i^+} p_{ij}(t) \le P_i, \quad \forall i \in \mathcal{V}$$

$$\sum_{j\in\tilde{\delta}_i^+} x_{ij}(t) \le 1, \quad \forall i \in \mathcal{V}_\mathrm{a}$$

Proposed Design

- Solve the problem by Lyapunov drift-plus-penalty (LDP) approach
  - Linear combination of drift and penalty weighted by parameter V

$$\mathrm{LDP} \triangleq \underbrace{[L(t+1) - L(t)]}_{\Delta(t)} + V h_1(t)$$

$$\tilde{\boldsymbol{Q}}(t) = \mathrm{diag}\{\boldsymbol{\kappa}\}\boldsymbol{Q}(t)$$

$$w_i^{(u,\phi,m)} = \left[\tilde{Q}_i^{(u,\phi,m)}(t) - \xi_\phi^{(m)}\tilde{Q}_i^{(u,\phi,m+1)}(t)\right]^+$$

$$w_{ij}^{(u,\phi,m)} = \left[\tilde{Q}_i^{(u,\phi,m)}(t) - \tilde{Q}_j^{(u,\phi,m)}(t)\right]^+;$$

$$\le B_0 + \boldsymbol{\lambda}^{\mathrm{T}}\tilde{\boldsymbol{Q}}(t) - \sum_{(u,\phi,m)} \Big\{ \boxed{\sum_{i\in\mathcal{V}} \left[\left(w_i^{(u,\phi,m)} - V c_{\mathrm{pr},i}\, r_\phi^{(m)}\right)\mu_{i,\mathrm{pr}}^{(u,\phi,m)}(t) - V s_{k_i(t)}\right]}$$
$$+ \boxed{\sum_{(i,j)\in\mathcal{E}_\mathrm{b}} \left[\left(w_{ij}^{(u,\phi,m)} - V c_{\mathrm{tr},ij}\right)\mu_{ij}^{(u,\phi,m)}(t) - V s_{k_{ij}(t)}\right]}$$
$$+ \boxed{\sum_{(i,j)\in\mathcal{E}_\mathrm{a}} \left[w_{ij}^{(u,\phi,m)}\mu_{ij}^{(u,\phi,m)}(t) - V c_{\mathrm{wt},i}\, p_{ij}(t)\tau\right]} \Big\} \quad (14)$$

Processing	Max-weight
Wired Trans	Max-weight
Wireless Trans	cvx problem

Performance Analysis

- The delay performance

$$\overline{h_2} \leq \frac{B_0}{\epsilon} + \frac{\left[\overline{h_1}^*(\boldsymbol{\lambda} + \epsilon \mathbf{1}) - \overline{h_1}^*(\boldsymbol{\lambda})\right] V}{\epsilon} \quad \sim O(V)$$

- The cost performance

optimal cost

$$\overline{h_1} \leq \boxed{\overline{h_1}^*(\boldsymbol{\lambda})} + \frac{B_0}{V} \quad \sim O(1/V)$$

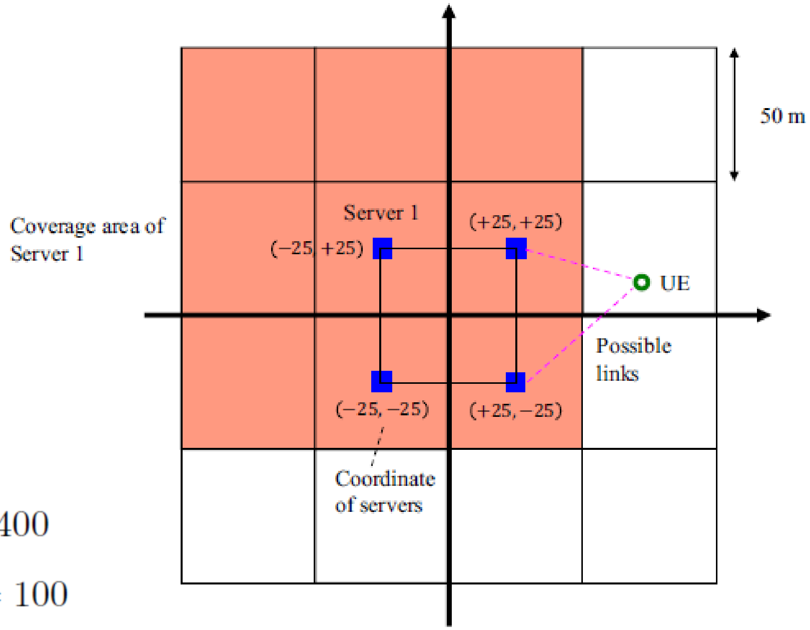
- The algorithm is fully distributed and efficient

Numerical Experiments

- Network Setup
  - 4 APs serving 100 UEs (random walk)
  - 3GPP urban microcell model
  - 100 MHz band allocated for each AP
- AgI services

$$\text{Service 1}: \ \xi_1^{(1)} = 1, \ \xi_1^{(2)} = 2; \ 1/r_1^{(1)} = 300, \ 1/r_1^{(2)} = 400$$

$$\text{Service 2}: \ \xi_2^{(1)} = \frac{1}{3}, \ \xi_2^{(2)} = \frac{1}{2}; \ 1/r_2^{(1)} = 200, \ 1/r_2^{(2)} = 100$$

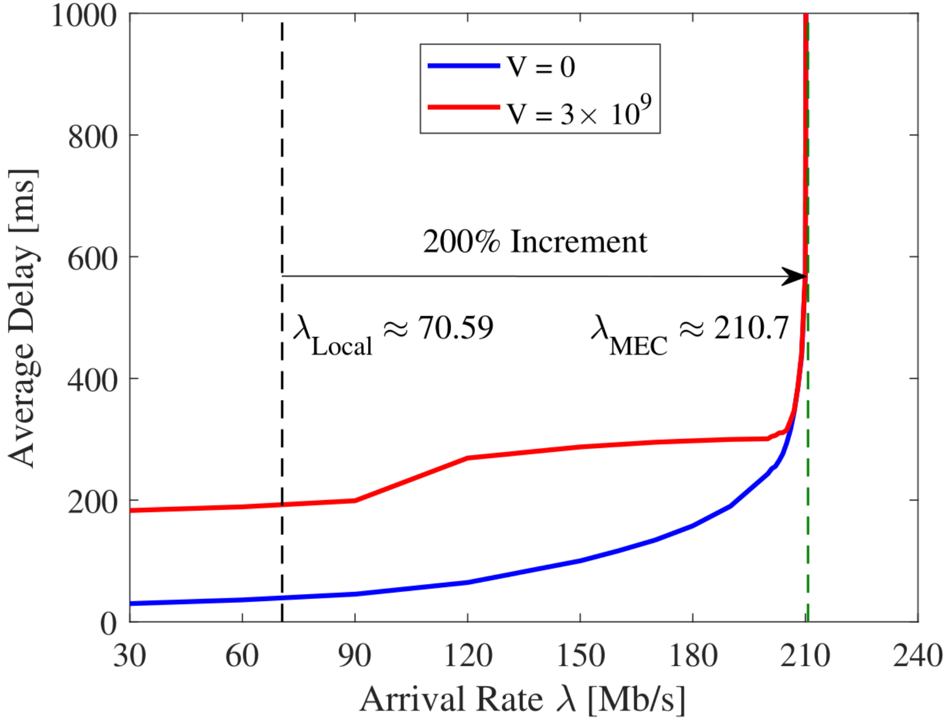


AVAILABLE RESOURCES AND COSTS OF THE MEC NETWORK (ON THE BASIS OF SECOND)

	User $i \in \mathcal{V}_\text{a}$	Edge Server $i \in \mathcal{V}_\text{b}$
Computation	$\mathcal{K}_i = \{0, 1\}$, $C_{k_i} = k_i$ CPUs, $s_{k_i} = 5k_i$, $c_{\text{pr},i} = 1$ /CPU	$\mathcal{K}_i = \{0, \cdots, 10\}$, $C_{k_i} = 5k_i$ CPUs, $s_{k_i} = 5k_i$, $c_{\text{pr},i} = .2$ /CPU
Wired Links	No wired transmission between users	$\mathcal{K}_{ij} = \{0, \cdots, 5\}$, $C_{k_{ij}} = 10k_{ij}$ Gbps, $s_{k_{ij}} = k_{ij}$, $c_{\text{tr},ij} = 1$ /Gb
Wireless Links	$P_i = 200\,\text{mW}$, $c_{\text{wt},i} = 1$ /W	$P_i = 10\,\text{W}$, $c_{\text{wt},i} = .2$ /W

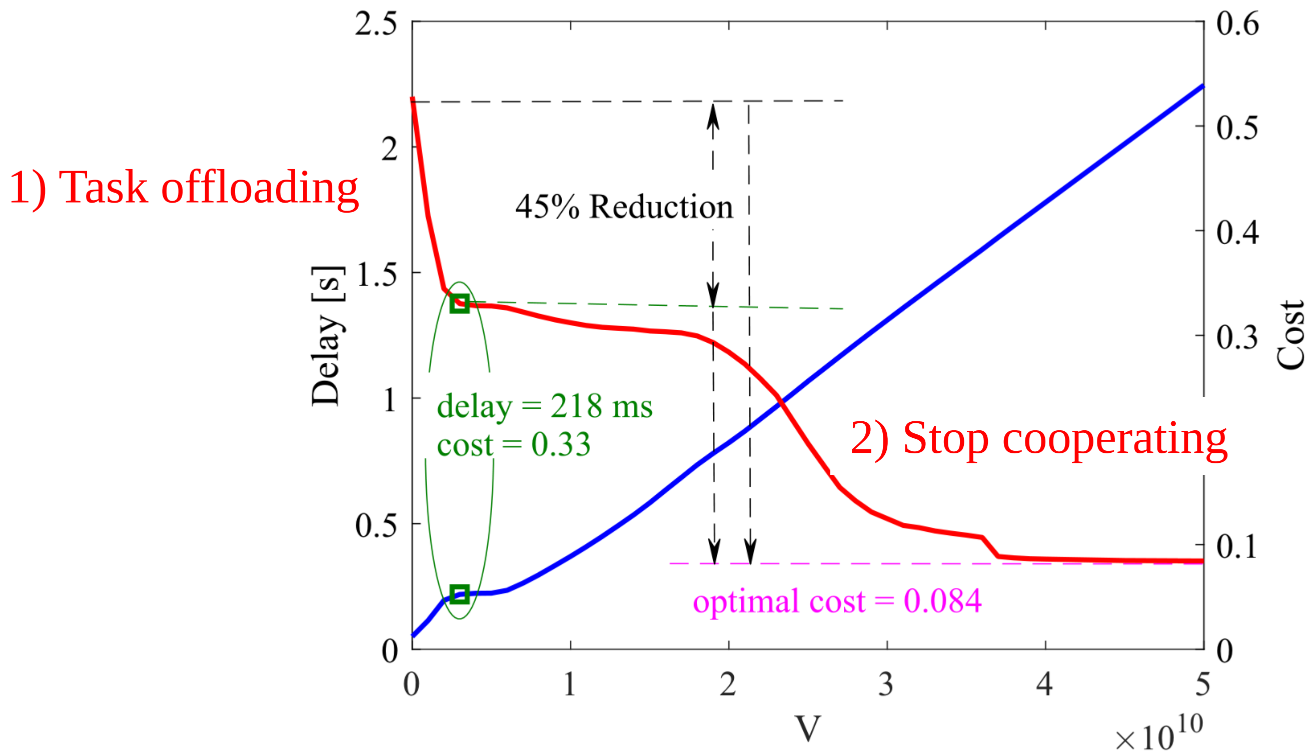
Numerical Experiments

- Stable region
  - The maximum arrival rate of requests that the considered network can support



Numerical Experiments

- Delay-Cost Tradeoff



Conclusions

- MEC can aid the delivery of real-time AgI services requested by end users, which can significantly improve the stable region
- The developed LDP-based algorithm can trade off the delay and cost performance, i.e., achieving near-optimal resource cost with guaranteed average delay performance
- The developed LDP-based algorithm is efficient and fully distributed

Q & A

- Thanks for joining the talk
- References

  - Y. Cai, J. Llorca, A. M. Tulino, and A. F. Molisch, "Mobile edge computing network control: Tradeoff between delay and cost," *arXiv*.
  - H. Feng, J. Llorca, A. M. Tulino, and A. F. Molisch, "Optimal dynamic cloud network control," *IEEE/ACM Trans. Netw.*.
  - H. Feng, J. Llorca, A. M. Tulino, and A. F. Molisch,, "Optimal control of wireless computing networks," *IEEE Trans.Wireless Commun.*.

- Any questions, comments, please contact via

  - Email: yangcai@usc.edu
  - Website: https://wides.usc.edu/