Essence
Pervasive
& Distributed Intelligence

NETLAB
NETWORKED SYSTEMS RESEARCH LABORATORY

University of Glasgow | School of Computing Science

On the Optimality of Task Offloading in Mobile Edge Computing (MEC) Environments

IEEE Global Communications Conference
9-13 December 2019, Waikoloa, HI, USA
i.alghamdi.1@research.gla.ac.uk

Ibrahim Alghamdi, Christos Anagnostopoulos, Dimitrios P Pezaros [*University of Glasgow*]

Outline

- Introduction
  - Background
  - Motivation & Challenge
  - Related work and contribution
- Time-optimized offloading decision making
  - System Model
  - Problem Formulation
  - Maximizing the Probability of Offloading to the Best Server
  - Minimizing the Expected Total Delay of Task Offloading
- Performance evaluation
  - Data set
  - Performance Assessment in Single user scenario
  - Performance Assessment in Competitive Setting

Ibrahim Alghamdi, Christos Anagnostopoulos, Dimitrios P Pezaros [*University of Glasgow*]

Introduction: *New forms of mobile nodes*

Ibrahim Alghamdi, Christos Anagnostopoulos, Dimitrios P Pezaros [*University of Glasgow*]

Introduction: *the requirements of the emerging applications*

- Require higher computing/networking resources:
  - Latency-sensitive application (virtual reality)
  - Powerful CPUs (data analytics using machine learning)
  - Need more storages (sensing and collecting data)
- These requirements contradict with the mobile nodes capabilities.

Ibrahim Alghamdi, Christos Anagnostopoulos, Dimitrios P Pezaros [*University of Glasgow*]

Introduction: *Computation offloading*

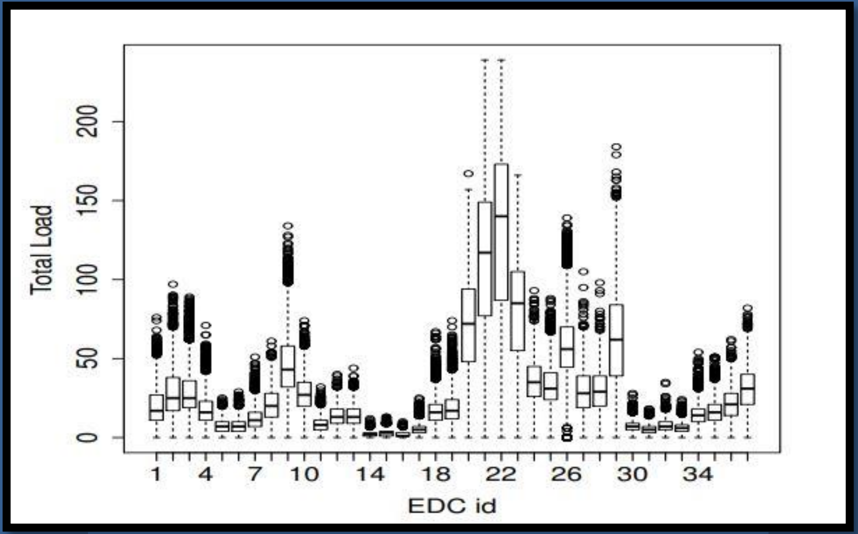
- Sending the computing task to an external server.
- The computation offloading reduces latency up to 88% and energy consumption of mobile devices up to 93%.[1]

[1] **J. Dolezal, Z. Becvar, and T. Zeman, "Performance evaluation of computation offloading from mobile device to the edge of mobile network,"** ***in CSCN*****. IEEE, 2016, pp. 1–7.**

Ibrahim Alghamdi, Christos Anagnostopoulos, Dimitrios P Pezaros [*University of Glasgow*]

Motivation

- The deployment of MEC servers.[2]
- MEC servers' load have large variation.[3]



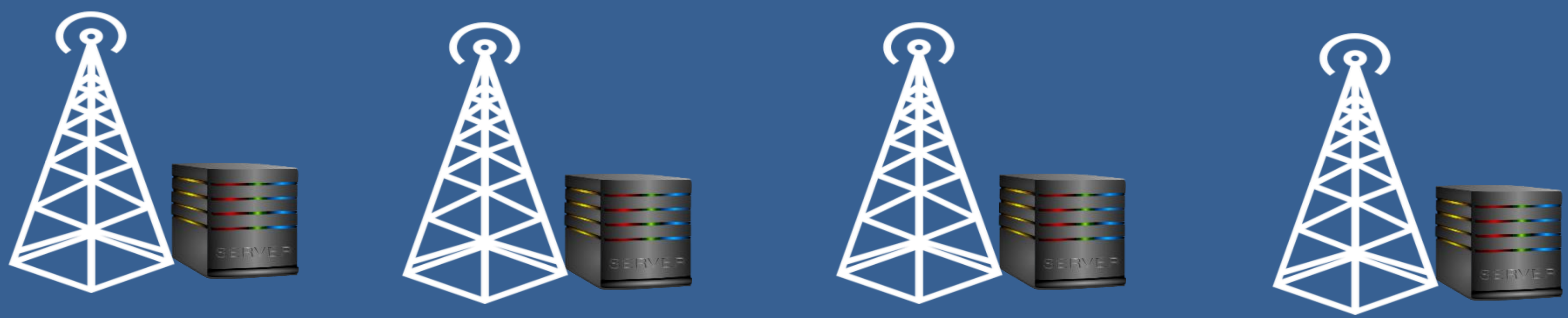
Workload in 37 EDCs according to the simulation in [3]

[2] M. Patel, B. Naughton, C. Chan, N. Sprecher, S. Abeta, A. Neal et al., "Mobile-edge computing introductory technical white paper," *White Paper, Mobile-edge Computing (MEC) industry initiative*, 2014.
[3] C. N. Le Tan, C. Klein, and E. Elmroth, "Location-aware load prediction in edge data centers," *in 2nd FMEC*. IEEE, 2017, pp. 25–31.

Ibrahim Alghamdi, Christos Anagnostopoulos, Dimitrios P Pezaros [*University of Glasgow*]

Challenge

- The decision of when and where to offload task/data?



Delay=45 ms   Delay=30 ms   Delay=20 ms   Delay=43 ms

Previous work (1)

- Previous works try to answer the questions:
  - Should a task be offloaded to external server?
  - If yes, should we do it in the cloud or to the edge?
  - In the edge, there is an assumption that the mobile node will have a set of options to select from.
- We consider a special case that might arise in the MEC environments and apply the concept of Optimal Stopping Theory.

Ibrahim Alghamdi, Christos Anagnostopoulos, Dimitrios P Pezaros [*University of Glasgow*]

Contribution

- This work departs from our previous works [4, 5]:
  - But different from our previous work, we propose a model for the realistic case where the number of servers is unknown.
  - we propose a model that maximizes the chance of offloading to the optimal server.

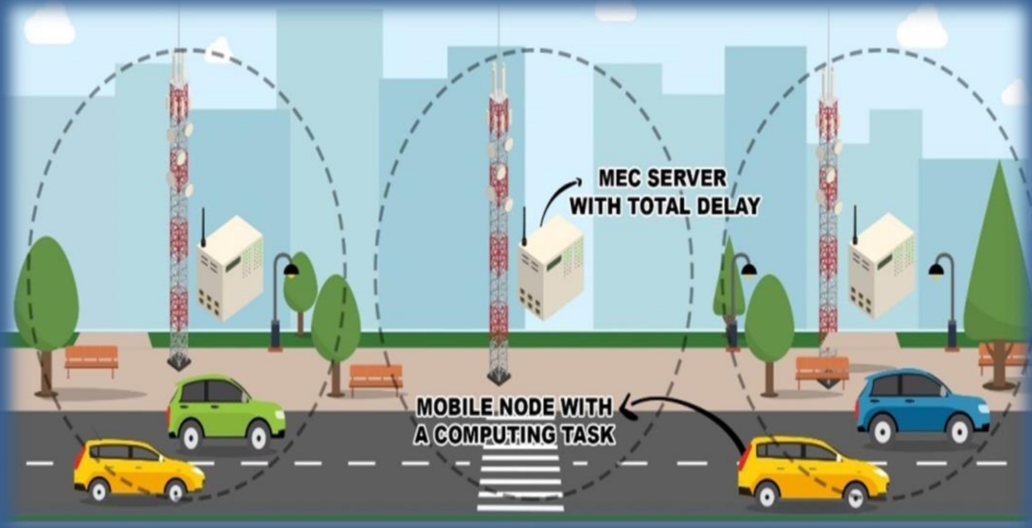
---

[4] I. A. I. Alghamdi, C. Anagnostopoulos, and D. Pezaros, "Timeoptimized task offloading decision making in mobile edge computing," in 11th IEEE Wireless Days, 2019[2] C. N. Le Tan, C. Klein, and E. Elmroth, "Location-aware load prediction in edge data centers," *in 2nd FMEC*. IEEE, 2017, pp. 25–31.
[5] I. A. I. Alghamdi, C. Anagnostopoulos, and D. Pezaros, "Delay-tolerantsequential decision making for task offloading in mobile edge computingenvironments,"Information, 2019.

Ibrahim Alghamdi, Christos Anagnostopoulos, Dimitrios P Pezaros [*University of Glasgow*]

Setting/system model?[1]

- MEC servers deployed along the user path with total delay X.
- Mobile node moves in 1D.
- Computing task to be offloaded to one of the MEC servers.
- The mobile node only knows about the current MEC (the one in the range of mobile node)



[6] K. Zhang, Y. Mao, S. Leng, Y. He, and Y. Zhang, "Mobile-edge computing for vehicular networks: A promising network paradigm with predictive off-loading," IEEE Vehicular Technology Magazine, vol. 12, no. 2, pp. 36–44, 2017.

Problem Statement

Problem 1: Maximizing the Probability of Offloading to the Best Server.

Problem 2: Minimizing the Expected Total Delay of Task Offloading.

- Specifically: find an stopping rules (offloading) that achieve the previous two goals.
- These two problems are modelled as an optimal stopping problem.

Ibrahim Alghamdi, Christos Anagnostopoulos, Dimitrios P Pezaros [*University of Glasgow*]

Maximizing the Probability of Offloading to the Best Server (1)

- Assumption:
  - We know the number of options servers/times.
  - No recalled is allowed.
- Goal:
  - Define an offloading policy/rule which maximizes the chance of choosing the best server w.r.t. The expected total delay.
  - Max (P_n^*)
- This is cast as a Best-Choice Problem (BCP) [7].

---

[7] T. S. Ferguson, "Optimal Stopping and Applications," http://www.math.ucla.edu/ tom/Stopping/Contents.html, March 2019.

Ibrahim Alghamdi, Christos Anagnostopoulos, Dimitrios P Pezaros [*University of Glasgow*]

Maximizing the Probability of Offloading to the Best Server (2)

- Let us call the t-th server *candidate*, if it is the best in terms of X_t, $t = 1, \dots, n$.
- Based on the BCP, the optimal policy is to reject the first $r_n - 1$ servers and then select the first candidate, if any, to offload the tasks.
- The value of r_n is defined as:
  - $r_n = \min\{r \geq 1: \frac{1}{r} + \frac{1}{r+1} + \dots + \frac{1}{n-1} \leq 1\}$ for $n \geq 2$. ..(1)
- Theorem 1. The optimal probability in selecting the best server in (1) is given by:
  - $P^*(r_n) = \frac{r_n - 1}{n} \sum_{k=r_n}^{n} \frac{1}{k-1}$ ...(2)
- In the case where there is a relatively high number of servers, the optimal probability is around 0.368 [7].

[7] T. S. Ferguson, "Optimal Stopping and Applications," http://www.math.ucla.edu/ tom/Stopping/Contents.html, March 2019.

Ibrahim Alghamdi, Christos Anagnostopoulos, Dimitrios P Pezaros [*University of Glasgow*]

BCP based Optimal Task Offloading Policy

1. The node observes and reject the first n/e:
   - Ranks them immediately w.r.t. their total delay provided by each of them upon request.
2. The node offloads the task/data to the first t-th server with $t > [n/e]$ which is ranked as the relatively best server compared to the previously ones.

- *This rule is guaranteed to maximize the probability of offloading the task/data to the best server.*

[7] T. S. Ferguson, "Optimal Stopping and Applications," http://www.math.ucla.edu/ tom/Stopping/Contents.html, March 2019.

Ibrahim Alghamdi, Christos Anagnostopoulos, Dimitrios P Pezaros [*University of Glasgow*]

Minimizing the Expected Total Delay of Task Offloading

- Assumption:
  - We have an idea about the the load of the MEC servers, i.e. X.

- Goal:
  - We desire to find when to offload and which server that minimizes the total expected delay $\mathbb{E}[X]$.
  - The node pays c cost units per observation when it has not yet offloaded the task/data.

$$\boldsymbol{Y = X + ct} \;\ldots(3)$$

Ibrahim Alghamdi, Christos Anagnostopoulos, Dimitrios P Pezaros [*University of Glasgow*]

Cost-based Optimal Task Offloading Policy

$$\boldsymbol{Y = X + ct} \dots \textit{(3)}$$

- The node minimizes the expected cost in (3) by offloading at the first t-th server such that:

$$t^* = \min\{t > 0 \colon X_t \leq V^*\} \dots (4)$$

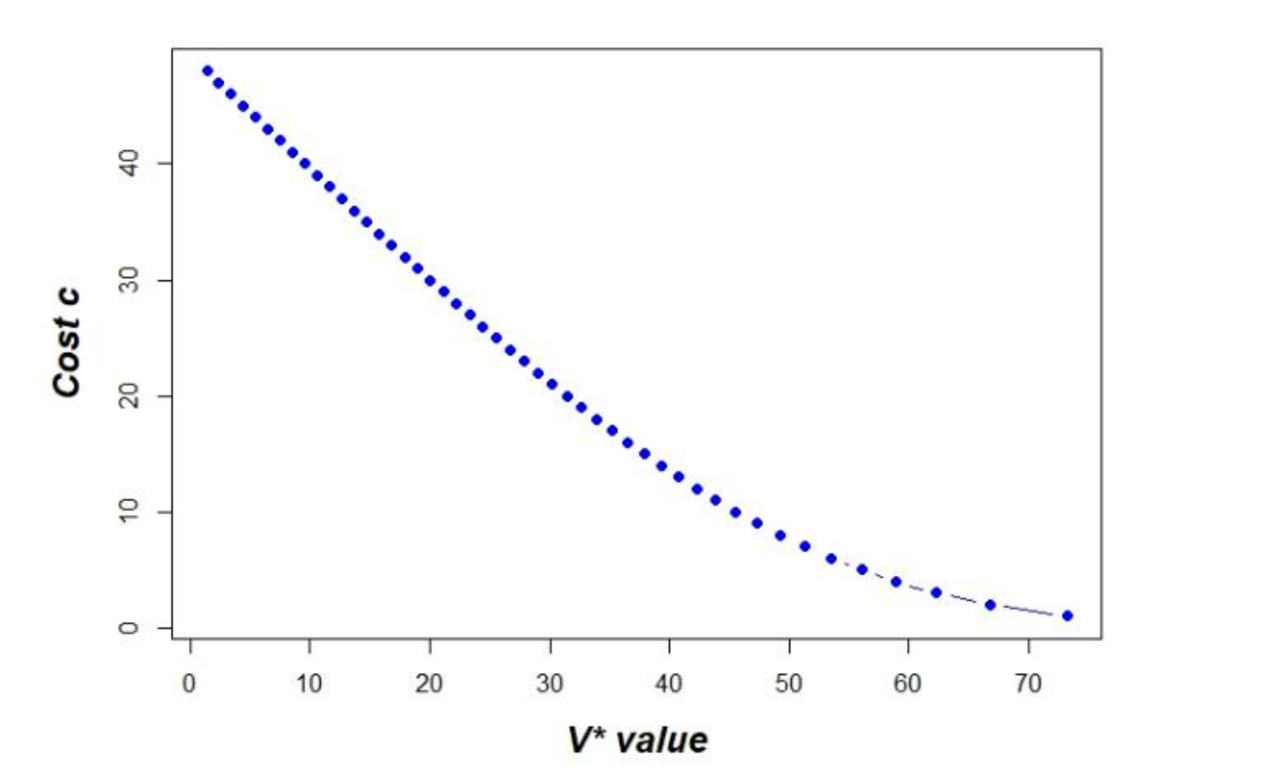
where the V^* *is the solution of:*

$$\int_{V^*}^{\infty} (x - V^*) dF(x) = c \dots (5)$$

- *where* F(x) *is the CDF of X.*

Ibrahim Alghamdi, Christos Anagnostopoulos, Dimitrios P Pezaros [*University of Glasgow*]

Cont'd



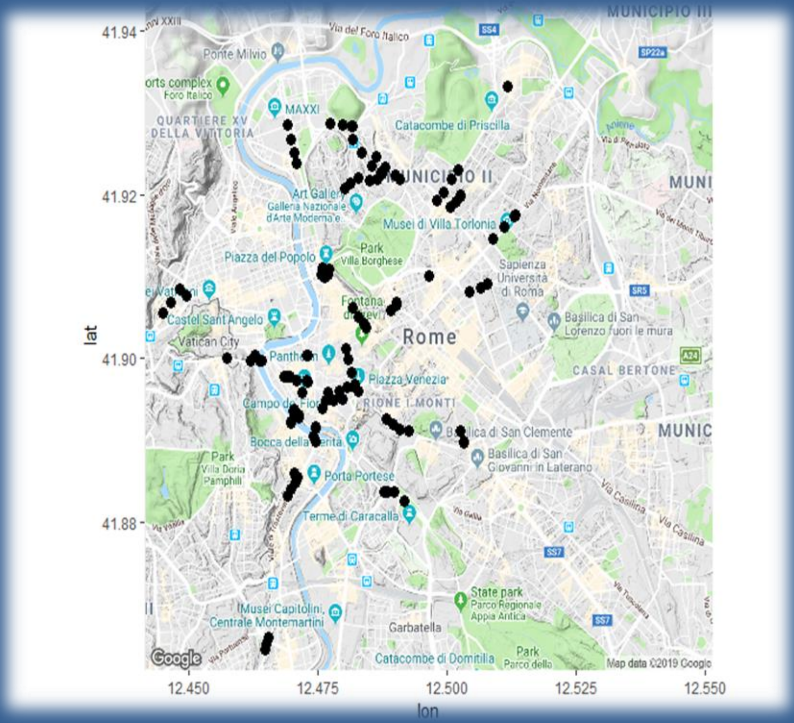
Ibrahim Alghamdi, Christos Anagnostopoulos, Dimitrios P Pezaros [*University of Glasgow*]

Performance Evaluation: data set

- We used the real dataset of taxi cabs' movements in Rome [8].

Table I: Data set used in the experiment

car id	Date	lat	long	Delay	Cell
156	"2014-02-0100:00:00.73"	41.88	12.48	80.61	4
156	"2014-02-0100:00:16.47"	41.88	12.48	62.97	4
156	"2014-02-0100:00:30.70"	41.88	12.48	4.53	4
156	"2014-02-0100:00:45.30"	41.88	12.49	4.37	4
187	"2014-02-0100:00:01.14"	41.92	12.46	70.17	1
187	"2014-02-0100:00:16.15"	41.92	12.46	66.59	1
187	"2014-02-0100:00:30.81"	41.92	12.47	31.65	4



[8] L.Bracciale,M.Bonola,P.Loreti,G.Bianchi,R.Amici,andA.Rabuffi, "CRAWDAD dataset roma/taxi (v. 2014-07-17)," Downloaded from https://crawdad.org/roma/taxi/20140717, Jul. 2014.

Ibrahim Alghamdi, Christos Anagnostopoulos, Dimitrios P Pezaros [*University of Glasgow*]

Data set: example

Car id	Date	Lat	Long	Delay	Cell(server)
156	2014-02-01 00:00:00.73	41.88	12.48	80.61	4
156	”2014-02-01 00:00:16.47”	41.88	12.48	62.97	4
156	”2014-02-01 00:00:30.70”	41.88	12.48	4.53	4
156	”2014-02-01 00:00:45.30”	41.88	12.48	4.37	4

Ibrahim Alghamdi, Christos Anagnostopoulos, Dimitrios P Pezaros [*University of Glasgow*]

Performance Assessment in Single User Scenario (1)

- BCP
- COT
- HS
  - Based on a threshold obtained by solve finite horizon OST discussed in [4, 5]
- Random.
- *p*-model with different probability *p*
- ***The optimal.***

---

[4] I. A. I. Alghamdi, C. Anagnostopoulos, and D. Pezaros, "Timeoptimized task offloading decision making in mobile edge computing," in 11th IEEE Wireless Days, 2019[2] C. N. Le Tan, C. Klein, and E. Elmroth, "Location-aware load prediction in edge data centers," *in 2nd FMEC*. IEEE, 2017, pp. 25–31.

[5] I. A. I. Alghamdi, C. Anagnostopoulos, and D. Pezaros, "Delay-tolerantsequential decision making for task offloading in mobile edge computingenvironments,"Information, 2019.

Performance Assessment in Single user scenario (2)

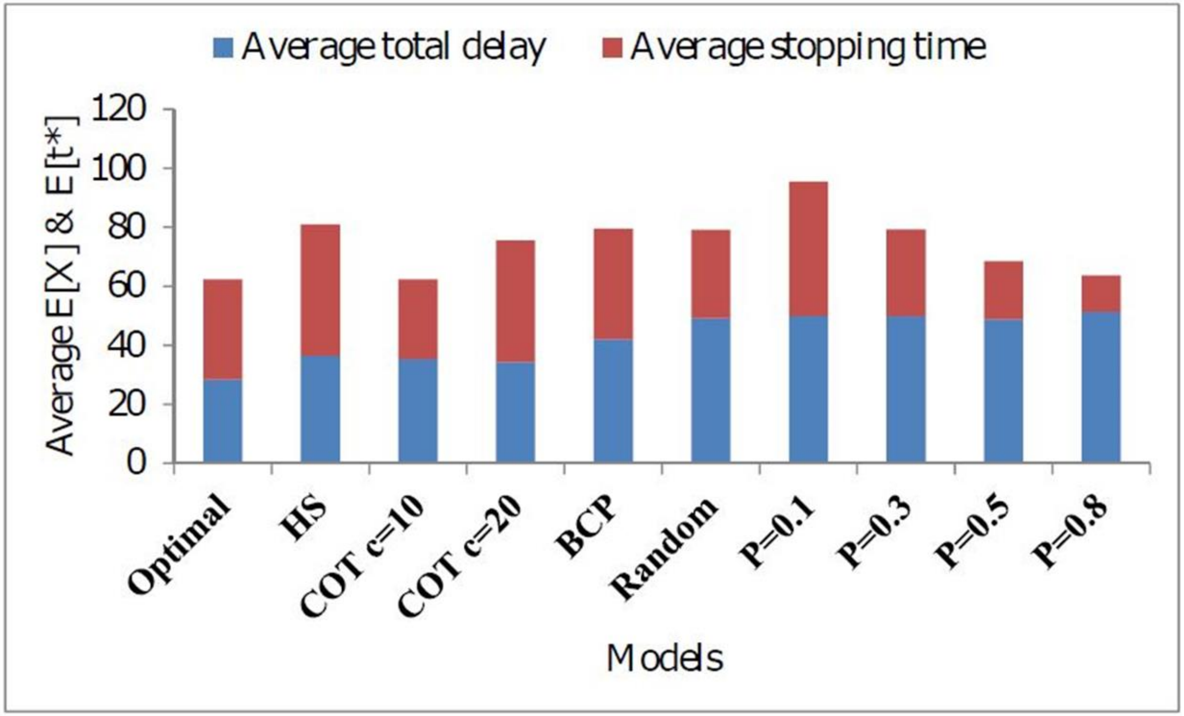


Figure 3: Average total delay $\mathbb{E}[X]$ and average stopping time $\mathbb{E}[t^*]$ in a single user setting.

Ibrahim Alghamdi, Christos Anagnostopoulos, Dimitrios P Pezaros [*University of Glasgow*]

Cont’d

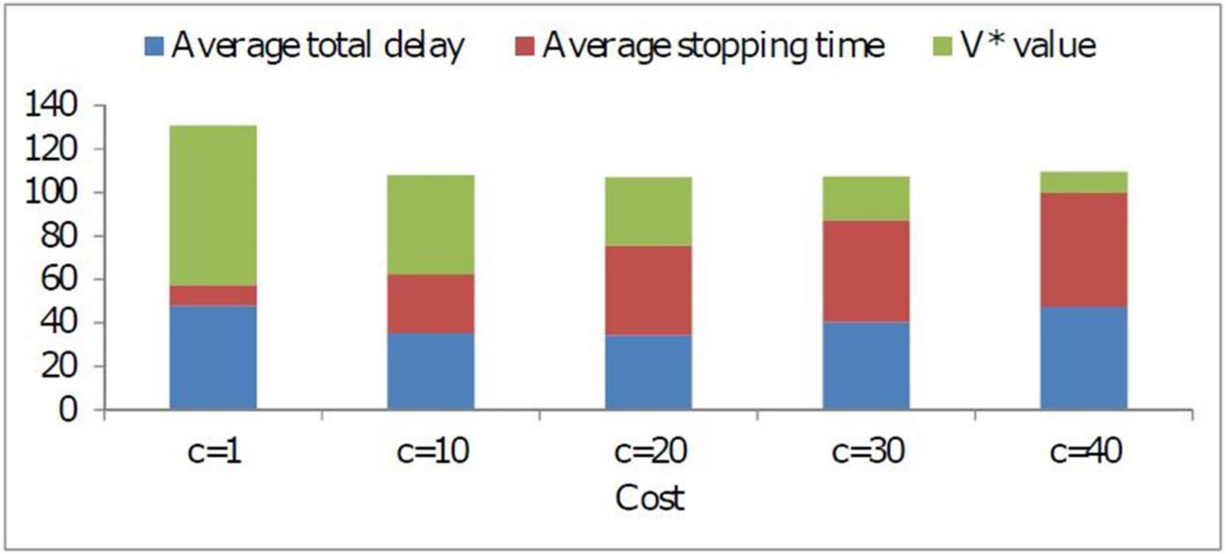


Figure 4: Average total delay $\mathbb{E}[X]$, average stopping time $\mathbb{E}[t^*]$, and optimal decision thresholds V^* in the COT model with different costs c.

Performance Assessment in Competitive Setting (1)

- When we have similar expected stopping times (many users offload to the same server)
- We used the *simmer* discrete simulator in R environment [9].
- We evaluated all models in terms of the average Waiting Time Ratio (WTR).
- We look for lower WTR.

---

[9] **I. Ucar, B. Smeets, and A. Azcorra, "simmer: Discrete-event simulation for r," arXiv:1705.09746, 2017.**

Performance Assessment in Competitive Setting (2)

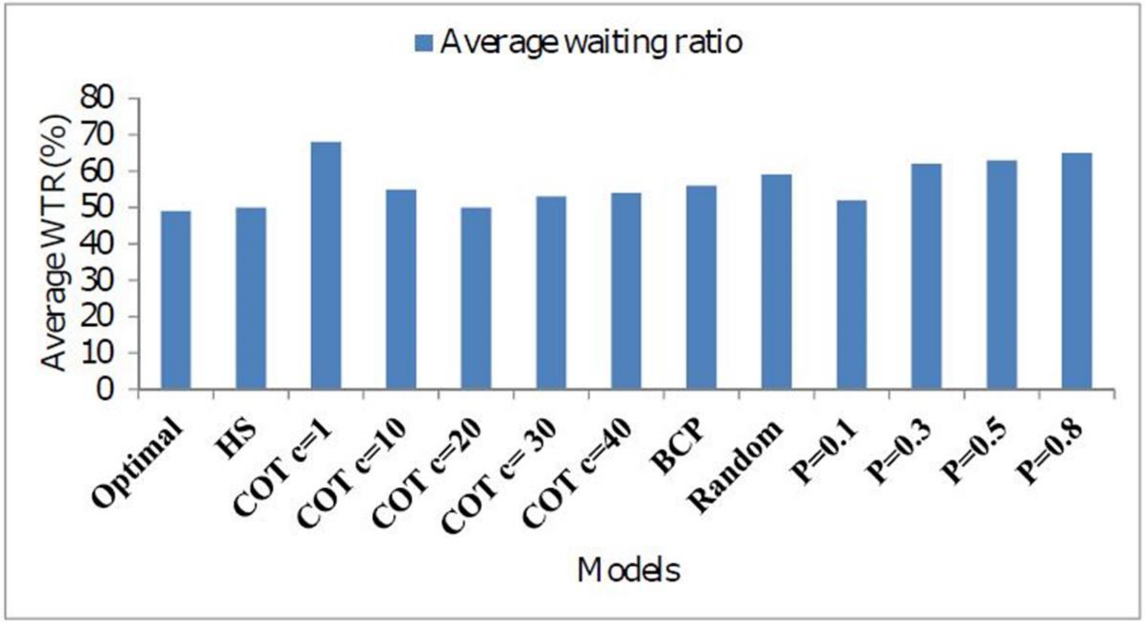


Figure 6: Average WTR for all models in a competitive setting.

Ibrahim Alghamdi, Christos Anagnostopoulos, Dimitrios P Pezaros [*University of Glasgow*]

Future Work and Conclusion

- **What have we learned?**
  - *It is not beneficial to offload at the very first server; the mobile node should, at least, pass a couple of servers to obtain a lower total delay and lower WTR when competing with other nodes.*
- **What comes next?**
  - *Define the cost based on a use case.*
  - *Try different OST models with different use cases.*

Ibrahim Alghamdi, Christos Anagnostopoulos, Dimitrios P Pezaros [*University of Glasgow*]

- Thank you
- Questions
- i.alghamdi.1@research.gla.ac.uk