A Truthful Online Mechanism for Allocating Fog Computing Resources

Extended Abstract

Fan Bi, Sebastian Stein, Enrico Gerding
University of Southampton
fb1n15,ss2,eg@ecs.soton.ac.uk

Nick Jennings
Imperial College London
n.jennings@imperial.ac.uk

Thomas La Porta
Penn State University
tlp@cse.psu.edu

KEYWORDS

Mechanism Design; Fog Computing; IoT; Resource Allocation

ACM Reference Format:
Fan Bi, Sebastian Stein, Enrico Gerding, Nick Jennings, and Thomas La Porta. 2019. A Truthful Online Mechanism for Allocating Fog Computing Resources. In *Proc. of the 18th International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2019), Montreal, Canada, May 13–17, 2019,* IFAAMAS, 3 pages.

1 INTRODUCTION

The Internet of Things (IoT) is developing rapidly, and it is estimated that by 2025 22 billion active devices will be in the IoT [13]. Since it is impossible to let the often low-powered IoT devices perform all computing tasks, some of which are highly computationally demanding, fog computing, which extends the cloud to be closer IoT devices, has been proposed as a solution [3].To make the most of the fog resources and maximise the efficiency, good fog computing resource allocation mechanisms are needed.

To address this challenge, researchers have proposed many resource allocation mechanisms for fog computing or similar computing paradigms [2, 4, 7, 8, 10, 19, 20]. However, most of these mechanisms were not specifically designed for settings where users act strategically to maximise their utility. Therefore, some researchers have proposed truthful mechanisms that incentivise users to truthfully reveal their private information [5, 12, 16–18, 21, 22]. However, these approaches cannot be applied directly to our model due to subtle but important differences. For example, [18] assumes single-minded users (i.e., users who do not get any value for a partially executed task). However, users in our model can get partial value for a partially executed task.

In this paper, we are the first to formulate the fog computing resource allocation problem as a constraint optimisation problem that considers bandwidth constraints (a key challenge in IoT settings) and allows flexible allocation of virtual machines (VMs) and of the bandwidth. Furthermore, we introduce a novel *dominant-strategy incentive compatible* (DSIC) and *individually rational* (IR) mechanism to maximise social welfare.[1] DSIC mechanisms guarantee that regardless of others' behaviours, users always maximise their utility by reporting truthfully. Furthermore, under an IR mechanism, no user will get a negative utility by participation.

[1]We define social welfare as the difference between the value and the operational costs of all tasks.

Proc. of the 18th International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2019), N. Agmon, M. E. Taylor, E. Elkind, M. Veloso (eds.), May 13–17, 2019, Montreal, Canada.

2 THE FOG RESOURCE MODEL

Next, we briefly describe our fog computing resource allocation model, which is shown in Figure 1. It contains a set P of geo-distributed micro data centres (MDCs) and a set L of locations, which are interconnected through a set $\mathbb{E}$ of data links. Furthermore, there is a set E_l of endpoints in each location l, and every MDC $p \in P$ has a set R of limited computational resources. Moreover, there are $A_{p,r}$ units of type $r \in R$ resources in MDC p, and the unit operational cost of resource r in MDC p is $o_{p,r}$. In addition, the bandwidth capacity and the unit operational cost of link $(j,k) \in \mathbb{E}$, which are assumed to be symmetrical for simplicity, are $b_{j,k}$ and $o_{j,k}$ respectively. Furthermore, the fog provider controls the resource allocation of the fog through a central control system.

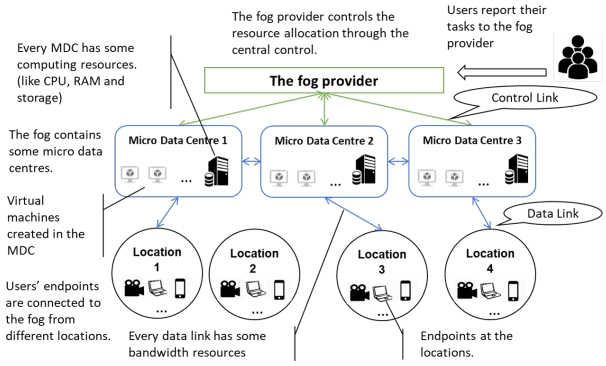

Figure 1: General view of a fog computing system.

Fog users with tasks arrive over time, and I denotes the set of all tasks. Note that we adopt a continuous time system, but the tasks can only start execution at discrete time steps, denoted by the set $T = \{1, 2, \ldots, |T|\}$. Each task $i \in I$ is owned by a user, which is also denoted as i for simplicity. In addition, the arrival time of task i is $T_i^a \in [0, |T|]$, which is the time when user i becomes aware of its task i, and the time interval that the task can run is from T_i^s to T_i^f. Here, we assume that no tasks arrive at the exact same time. The operational cost of task i is denoted as o_i, which is the total cost of task i. Furthermore, we also assume that every task only requires one VM to run but may require connections to several endpoints, and the endpoints of tasks do not change locations over time, VMs can migrate without costs, and all tasks are preemptive. Finally, we focus on time-oriented tasks, which are common in fog computing. Such a task i needs a certain capacity of resources for a time length t_i to get its full value, but can still

get part of the value if the processing time is less than t_i. Formally, the type of task i is a tuple $\theta_i = (T_i^a, T_i^s, T_i^f, \mathbf{v}_i, \{a_{i,r}\}_{r\in R}, \{\Gamma_l^i\}_{l\in L})$, where $a_{i,r}$ denotes the amount of resource $r \in R$ required, and Γ_l^i denotes the bandwidth demand between its VM and location $l \in L$. For simplicity, bandwidth demands are symmetrical. The valuation function is $\mathbf{v}_i = \{v_{i,0}, v_{i,1}, \ldots, v_{i,t_i}\}$, where $v_{i,t}$ is the value when task i gets usage time of t time steps. We make a mild assumption that the value monotonically increases with usage time.

Next, when receiving the type θ_i for task i, the fog provider will decide the resource allocation scheme λ_i to this task, and the payment $\tilde{p}_i$ right away. Formally, the fog provider solves a constraint optimisation problem, and the decision variables are: (1) $\{z_{p,t}^i \in \{0,1\}\}_{i\in I, p\in P, t\in T}$, indicating that the VM of task i is placed in MDC p ($z_{p,t}^i = 1$), or not ($z_{p,t}^i = 0$) at time step t. (2) $\{f_{l,p,j,k,t}^i \in \mathbb{R}^+\}_{i\in I, l\in L, p\in P, (j,k)\in\mathbb{E}, t\in T}$, indicating allocation of the bandwidth on each link for task i at time step t. (3) $\tilde{p}_i(\lambda_i, \theta^{\langle T_i^a\rangle}) \in \mathbb{R}^+$, denoting the payment of task i, which is a function of the allocation: λ_i and all tasks received by T_i^a: $\theta^{\langle T_i^a\rangle}$. So, for task i, its resource allocation scheme $\lambda_i = \{z_{p,t}^i\}_{i\in I, p\in P, t\in T} \cup \{f_{l,p,j,k,t}^i\}_{i\in I, l\in L, p\in P, (j,k)\in\mathbb{E}, t\in T}$ and its utility is $u_i = \mathbf{v}_i(\tilde{t}_i) - \tilde{p}_i(\lambda_i, \theta^{\langle T_i^a\rangle})$. The objective function maximises the total social welfare:

$$\underset{\lambda_i}{\text{maximise}} \sum_{i\in I} \mathbf{v}_i\Big(\sum_{p\in P, t\in T} z_{p,t}^i\Big) - o \tag{1}$$

where

$$o = \sum_{i\in I, r\in R, p\in P, t\in T} a_{i,r} z_{p,t}^i o_{p,r} + \sum_{i\in I, l\in L, p\in P, (j,k)\in\mathbb{E}, t\in T} 2o_{j,k} f_{l,p,j,k,t}^i$$

Then, the following equations are the constraints.

$$\text{Subject to:} \quad \sum_{p\in P} z_{p,t}^i \le 1 \quad \forall i \in I, t \in T \tag{2a}$$

$$\sum_{i\in I} z_{p,t}^i a_{i,r} \le A_{p,r} \quad \forall p \in P, r \in R, t \in T \tag{2b}$$

$$z_{p,t}^i = 0 \quad \forall i \in I, p \in P, t < T_i^s \text{ or } t > T_i^f \tag{2c}$$

$$\sum_{j:(j,p)\in\mathbb{E}} f_{l,p,j,p,t}^i = \Gamma_l^i z_{p,t}^i \quad \forall p \in P, i \in I, l \in L, t \in T \tag{2d}$$

$$\sum_{k:(l,k)\in\mathbb{E}} f_{l,p,l,k,t}^i = \Gamma_l^i z_{p,t}^i \quad \forall p \in P, i \in I, l \in L, t \in T \tag{2e}$$

$$\sum_{j:(j,k)\in\mathbb{E}} f_{l,p,j,k,t}^i = \sum_{j:(k,j)\in\mathbb{E}} f_{l,p,k,j,t}^i \forall p \in P, k \in P, i \in I, l \in L, t \in T \tag{2f}$$

$$\sum_{i\in I, l\in L, p\in P} f_{l,p,j,k,t}^i \le b_{j,k} \quad \forall (j,k) \in \mathbb{E}, t \in T \tag{2g}$$

$$f_{l,p,j,k,t}^i \ge 0 \quad \forall i \in I, l \in L, p \in P, (j,k) \in \mathbb{E}, t \in T \tag{2h}$$

The above optimisation problem is a mixed integer linear programming problem. Unfortunately, this problem is NP-hard.

3 FLEXIBLE ONLINE GREEDY MECHANISM

The key idea of our mechanism is that it only commits the usage time $\tilde{t}_i$ to task i but keeps its allocation scheme flexible, so that it can achieve higher social welfare. We briefly describe our mechanism flexible online greedy (FlexOG) as follows. After receiving a report of task i, FlexOG finds the allocation that maximises the social welfare of all unfinished tasks given the constraints of their committed usage time. Then, FlexOG computes the usage time $\tilde{t}_i$ for task i from its corresponding allocation scheme, and commits it to task i, which means that task i is guaranteed to get $\tilde{t}_i$ usage time before its finish time T_i^f. Afterwards, FlexOG requires payment $\tilde{p}_i$ for task i as the marginal total operational cost.

Theorem 3.1. *The FlexOG mechanism is DSIC and IR.*

4 SIMULATION RESULTS

We have tested the robustness of our mechanism by running simulations with different parameters. To this end, Figure 2 shows a representative result below due to space limitation.[2] Here, we compare the total social welfare achieved by FlexOG with other benchmarks under different resource coefficients k, which indicates the abundance of resources (i.e., a higher k means that there is less competition for resources). We use synthetic data in this simulation, and the three benchmarks are offline optimal (optimally allocate resources knowing all tasks' information beforehead), online greedy (OG; it greedily allocate resource for each newly arrived task and commit that allocation scheme), social welfare maximisation online auction 2 (SWMOA2; a variant of SWMOA [16], which greedily allocate resource for each new arrived task based on a virtual cost and commit that allocation scheme).

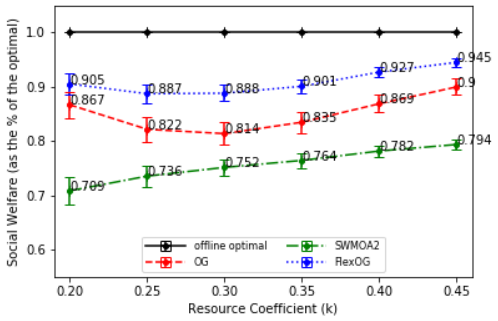

Figure 2: The social welfare achieved by four mechanisms.

We can see from the figure that FlexOG consistently achieves better social welfare than two truthful benchmarks (OG and SWMOA2), and achieves around 90% social welfare of the upper bound (offline optimal). FlexOG performs better than OG because the way in which committed time steps are allocated to tasks is flexible. Due to this, it can reschedule unfinished tasks to allocate more time steps for newly arrived tasks.

5 CONCLUSION

This paper formulates the fog computing resource allocation problem as a constrained optimisation problem and proposes a novel truthful online mechanism for solving it. In the future, we plan to design online mechanisms that combine machine learning and online mechanism design to further improve social welfare.

Acknowledgements: This research was sponsored by the U.S. Army Research Laboratory and the U.K. Ministry of Defence under Agreement Number W911NF-16-3-0001. The views and conclusions contained in this document are those of the authors and should not be interpreted as representing the official policies, either expressed or implied, of the U.S. Army Research Laboratory, the U.S. Government, the U.K. Ministry of Defence or the U.K. Government. The U.S. and U.K. Governments are authorized to reproduce and distribute reprints for Government purposes notwithstanding any copyright notation hereon.

[2]This figure is with 95% confidence intervals based on 200 trials, and the relative tolerance of the CPLEX optimizer is set to 1% for offline optimal, and 5 % for others. (A 1% tolerance means that the optimizer stops when a solution is within 1% of optimality)

REFERENCES

[1] 2015. *White paper: Fog Computing and the Internet of Things: Extend the Cloud to Where the Things Are.* Technical Report. CISCO. https://www.cisco.com/c/dam/en_us/solutions/trends/iot/docs/computing-overview.pdf

[2] Mohammad Aazam and Eui-Nam Huh. 2015. Fog computing micro datacenter based dynamic resource estimation and pricing model for IoT. In *Proc. of 29th International Conference on AINA*. IEEE, 687–694.

[3] Flavio Bonomi, Rodolfo Milito, Jiang Zhu, and Sateesh Addepalli. 2012. Fog computing and its role in the internet of things. In *Proc. of the first MCC workshop*. ACM, 13–16.

[4] Valeria Cardellini, Vincenzo Grassi, Francesco Lo Presti, and Matteo Nardelli. 2015. On QoS-aware scheduling of data stream applications over fog computing infrastructures. In *ISCC*. IEEE, 271–276.

[5] Shuchi Chawla, Nikhil R Devanur, Alexander E Holroyd, Anna R Karlin, James B Martin, and Balasubramanian Sivan. 2017. Stability of service under time-of-use pricing. In *Proc. of the 49th Annual ACM SIGACT Symposium on Theory of Computing*. ACM, 184–197.

[6] Shanzhi Chen, Hui Xu, Dake Liu, Bo Hu, and Hucheng Wang. 2014. A vision of IoT: Applications, challenges, and opportunities with china perspective. *IEEE Internet of Things journal* 1, 4 (2014), 349–359.

[7] Wuhui Chen, Incheon Paik, and Zhenni Li. 2017. Cost-aware streaming workflow allocation on geo-distributed data centers. *IEEE Trans. Comput.* 66, 2 (2017), 256–271.

[8] Cuong T Do, Nguyen H Tran, Chuan Pham, Md Golam Rabiul Alam, Jae Hyeok Son, and Choong Seon Hong. 2015. A proximal algorithm for joint resource allocation and minimizing carbon footprint in geo-distributed fog computing. In *ICOIN*. IEEE, 324–329.

[9] Charalampos Doukas and Ilias Maglogiannis. 2012. Bringing IoT and cloud computing towards pervasive healthcare. In *Proc. of Sixth International Conference on IMIS*. IEEE, 922–926.

[10] Yunan Gu, Zheng Chang, Miao Pan, Lingyang Song, and Zhu Han. 2018. Joint radio and computational resource allocation in IoT fog computing. *IEEE Transactions on Vehicular Technology* 67, 8 (2018), 7475–7484.

[11] Keiichiro Hayakawa, Enrico H Gerding, Sebastian Stein, and Takahiro Shiga. 2018. Price-based online mechanisms for settings with uncertain future procurement costs and multi-unit demand. In *Proc. of the 17th International Conference on AAMAS*. 309–317.

[12] Brendan Lucier, Ishai Menache, Joseph Seffi Naor, and Jonathan Yaniv. 2013. Efficient online scheduling for deadline-sensitive jobs. In *Proc. of the twenty-fifth annual ACM symposium on Parallelism in algorithms and architectures*. ACM, 305–314.

[13] Knud Lasse Lueth. 2018. State of the IoT 2018: Number of IoT devices now at 7B – Market accelerating. https://iot-analytics.com/state-of-the-iot-update-q1-q2-2018-number-of-iot-devices-now-7b/. (2018).

[14] Noam Nisan, Tim Roughgarden, Eva Tardos, and Vijay V. Vazirani. 2007. *Algorithmic Game Theory.* Cambridge University Press, New York, NY, USA.

[15] Anam Sajid, Haider Abbas, and Kashif Saleem. 2016. Cloud-assisted IoT-based SCADA systems security: A review of the state of the art and future challenges. *IEEE Access* 4 (2016), 1375–1384.

[16] Weijie Shi, Chuan Wu, and Zongpeng Li. 2017. An online auction mechanism for dynamic virtual cluster provisioning in geo-distributed clouds. *Proc. of IEEE Transactions on Parallel and Distributed Systems* 28, 3 (2017), 677–688.

[17] Changjun Wang, Weidong Ma, Tao Qin, Xujin Chen, Xiaodong Hu, and Tie-Yan Liu. 2015. Selling Reserved Instances in Cloud Computing Changjun. In *IJCAI*, Vol. 17. 265–278. https://doi.org/10.1109/TFUZZ.2008.924315

[18] Qian Wang, Kui Ren, and Xiaoqiao Meng. 2012. When cloud meets ebay: Towards effective pricing for cloud computing. In *Proc. of INFOCOM*. IEEE, 936–944.

[19] Jie Xu and Shaolei Ren. 2016. Online learning for offloading and autoscaling in renewable-powered mobile edge computing. In *Proc. of GLOBECOM*. IEEE, 1–6.

[20] Huaqing Zhang, Yong Xiao, Shengrong Bu, Dusit Niyato, F Richard Yu, and Zhu Han. 2017. Computing resource allocation in three-tier IoT fog networks: A joint optimization approach combining Stackelberg game and matching. *IEEE Internet of Things Journal* 4, 5 (2017), 1204–1215.

[21] Xiaoxi Zhang, Chuan Wu, Zongpeng Li, and Francis C M Lau. 2015. A truthful (1-ϵ)-optimal mechanism for on-demand cloud resource provisioning. In *Proc. of INFOCOM*. IEEE, 1053–1061.

[22] Yifei Zhu, Silvery D. Fu, Jiangchuan Liu, and Yong Cui. 2018. Truthful Online Auction Toward Maximized Instance Utilization in the Cloud. *IEEE/ACM Trans. Netw.* 26, 5 (Oct. 2018), 2132–2145. https://doi.org/10.1109/TNET.2018.2864726