The Fog Node Location Problem

Rodrigo A. C. da Silva[1,2], **Nelson L. S. da Fonseca**[1]

[1]Institute of Computing – University of Campinas (Unicamp)
Campinas - SP - Brazil

[2]Center for Mathematics, Computing and Cognition – Federal University of ABC (UFABC)
Santo André - SP - Brazil

rodrigo@lrc.ic.unicamp.br, nelson@ic.unicamp.br

Abstract. *This paper summarizes the thesis "The Fog Node Location Problem", which attempted to answer the question of how fog nodes should be located to process end-user demands that are variable in time and space. The problem was studied from different perspectives, optimizing the number of served users, deployment costs, energy consumption, and latency. This research considered both fixed servers and mobile fog nodes mounted on unmanned aerial vehicles (UAVs). The thesis has introduced several contributions, including linear programming models and novel algorithms. Results showed that the proposed solutions were quite efficient to design a fog computing infrastructure and that UAVs are suitable to be used as fog nodes.*

1. Introduction

Data generated by the Internet of Things (IoT) devices have commonly been processed in cloud data centers. Despite the high availability of resources, data centers are typically located in remote areas, preventing the deployment of several applications with strict latency requirements. Fog computing has been proposed to provide computing, storage, and networking capabilities along the continuum between cloud and end-users, reducing the latency between users and infrastructure.

A fog computing infrastructure relies on fog nodes, facilities that can host just a single device with processing capabilities or even a set of dedicated servers. Previous work [Kim and Chung 2018] has discussed the role of fog nodes in the network but has not addressed the impact of the geographical locations of fog nodes. The location problem consists in deciding where fog nodes should be placed given a set of potential locations and the devices available for deployment. Efficiently solving this problem is crucial for both users and fog provider since inappropriate decisions can increase the delay delivered to end-users and the deployment cost of the infrastructure.

Deciding the location of fog nodes is not a trivial task because of the variability of end-user demands in time and space. This variability is a result of end-users' mobility, which tends to concentrate a large number of computational demands in certain areas during short periods. This can cause overdimensioning of the fog infrastructure, making processing resources idle for long periods.

This paper summarizes the thesis "The Fog Node Location Problem" [da Silva and da Fonseca 2022b] that attempted to answer the question *"How should fog nodes be located to process end-user demands that are variable in time and* *space?"*. The thesis proposed solutions to select which locations should be used for the deployment of fog nodes and the computational capacity of each node. This is the fog node location problem, a variation of the facility location problem. The main goal is to process the maximum workload possible, but other criteria have also been adopted.

The work in the thesis considered a variety of scenarios. Some solutions considered a fog infrastructure with only terrestrial fog nodes, facilities that host dedicated servers and communicates with end-users via wireless interfaces. Other solutions in the thesis considered the employment of mobile fog nodes to process end-users' requests in different locations. These mobile nodes were mounted on unmanned aerial vehicles (UAVs) due to their small size, flexibility to access remote locations, autonomous operation, and onboard processing and networking devices. Moreover, a resource allocation algorithm was proposed to decide if a task should be instantiated in the fog or in the cloud.

The remainder of this paper is organized as follows. Section 2 reviews related work. Section 3 introduces the main contributions of the thesis considering only terrestrial nodes. Section 4 presents contributions that considered the employment of fog nodes mounted on UAVs. Section 5 summarizes the main results of the thesis. Section 6 lists the publications of the thesis. Finally, Section 7 draws the conclusions.

2. Related work

The facility location is a common problem in computer networks in which facilities range from single devices to large data centers. For instance, [Zhao et al. 2018] and [Lähderanta et al. 2021] proposed algorithms to place individual edge servers. Zhao et al. proposed a solution for selecting the position of servers to provide real-time data processing for IoT. They proposed a metric that quantifies the performance gain of candidate locations, and then used it to assign edge servers to locations. Lähderanta et al. proposed an algorithm customizable for different networks. Their evaluation showed that the number of hierarchical fog layers depends on the time and space variation of workloads.

Larumbe and Sansò proposed solutions to the location of cloud data centers in a backbone network. Their work employed a mixed-integer linear programming (MILP) formulation [Larumbe and Sansò 2012] and a scalable tabu search algorithm [Larumbe and Sansò 2013] to decide on the location of data centers to minimize delay, energy consumption, costs, and the emission of greenhouse gases.

Some authors employed UAVs as fog nodes. In [Wang et al. 2020], for instance, UAVs were dispatched to hover at specific areas to process user tasks within a given deadline. Employing UAVs increased the number of processed tasks compared with deployments with only terrestrial nodes, yet leading to latency fairness and avoiding the underutilization of resources. Zhou et al. [Zhou et al. 2021] considered UAVs with processing and caching capabilities to support virtual reality applications and proposed an algorithm to minimize latency by optimizing the 3D location of aerial nodes.

The work in the thesis differs from previous approaches in different ways. First, it takes into account the specific constraints of fog computing. Second, it was designed to deal with variable demands in time and space. Third, both the location and the capacity of fog nodes were jointly decided. Finally, the thesis introduced UAV type of node specificities in the location problem. Therefore, the presented thesis expands the state-of-the-art in different manners.

3. Terrestrial infrastructure

This part of the thesis considered a scenario with a cloud, various fog nodes, and end-user devices. Any device can access the cloud, but they can only access fog nodes within a limited range. Fog nodes host dedicated servers capable of processing end-user workload. Workloads have different latency requirements: the ones with strict-latency requirements can be processed on the fog, while others can be processed either on the fog or the cloud. End-users' devices can also process workloads if they have sufficient processing capacity.

The first research question is *"How should fog nodes be located to process end-user demands variable in time and space to reduce the cost of the fog infrastructure?"*. An answer to this question should improve the quality of the service delivered to end-users and reduce capital expenditure (CAPEX). We formulated the optimization problem as a multicriterial MILP model with three objective functions. First, it maximizes the processing of workload with strict latency requirements. Second, the total number of servers is minimized. Third, the model maximizes the total number of requests processed in the fog nodes, regardless of their latency requirements, to reduce the overall latency.

The second research question is *"How should fog nodes be located to process end-user demands variable in time and space to reduce the energy spent by end-user devices?"*. Answering this question is important since the batteries of mobile devices can be quickly drained, limiting the time end-users remain connected to the network. In the thesis, this problem was formulated as a MILP model with three objective functions. The first objective function maximizes the total number of requests with strict latency requirements processed by the fog. The second objective function minimizes the total energy consumption of mobile devices, accounting for the energy spent in data transmissions and processing. Finally, the total number of requests processed in the fog nodes is maximized. A heuristic algorithm denominated Energy and Demand Trade-off Algorithm (EDTA) was also proposed to obtain solutions for large fog deployments.

4. Aerial infrastructure

In the terrestrial infrastructure, once fog nodes are deployed, they cannot move to different locations. This represents a limitation since many servers remain underutilized while peaks of demands take place in other distant locations. To cope with that, proposals in this part of the thesis considered mobile fog nodes mounted on UAVs. Fixed servers are continuously connected to energy supplies and, therefore, can operate continuously. UAVs, on the other hand, are powered by onboard batteries, which limits considerably their operational time. The thesis considered these differences, showing how UAVs could potentially be used as complements to the terrestrial fog infrastructure.

This part attempted to answer three research questions. The first one is *"Are UAVs worth adopting to replace fixed nodes in a fog infrastructure?"*. To answer this question, a deployment with only fixed nodes was obtained, and then underloaded servers were replaced by UAVs as long as the CAPEX was not increased. The thesis proposed the UAV Fog Node Location algorithm to perform such a replacement. Moreover, a MILP model to design an infrastructure with both fixed and UAV fog nodes was proposed.

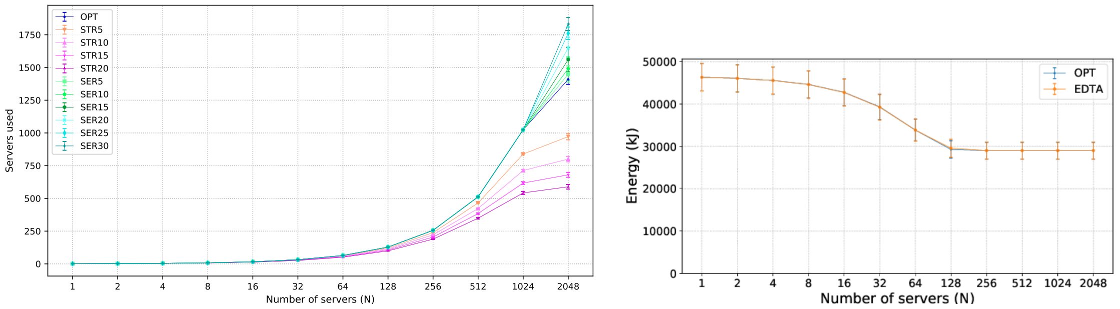
The remaining research questions aimed at evaluating infrastructures with only UAVs: i) *"What should be the location and operation period of fog nodes mounted on* *rotary-wing UAVs to maximize the number of end-users served while reducing the delay between ground nodes and UAVs?"* and ii) *"How should fixed-wing UAVs be positioned to provide a fog computing infrastructure to deal efficiently with variable demands in time and space, as well as maximize the number of processed requests?"*. Both questions were investigated since rotary-wing and fixed-wing UAVs present quite different energy and mobility models. The thesis modeled the problems as integer linear programming (ILP) models and two algorithms were proposed, the Sequential UAV Fog Node Location algorithm and the Spatio-Temporal UAV Fog Node Location algorithm.

5. Results

In this section, the main results obtained in the thesis are presented. Detailed analyses can be seen in the thesis [da Silva and da Fonseca 2022b]. Replicating experiments using real deployments with thousands of users, several candidate locations for fog node deployment, and many servers and UAVs is a costly task. Therefore, results in the thesis were obtained using simulation. Codes were mainly written in Python, and linear programming models were implemented in the Gurobi Optimizer solver. The data sets by Telecom Italia [Barlacchi et al. 2015] and the OpenCellId project were combined to model end-user requests in time and space. These data sets represent data from cellular networks in the metropolitan area of Milan, Italy. This allowed the simulation of hundreds of locations and thousands of users. UAVs were simulated based on real aircraft as well as realistic energy consumption and wireless channel models.

In the numerical evaluations, we varied the number of available devices for deploying fog nodes. The exact number of employed devices was lower than the number of available devices whenever there were sufficient resources to process the requests of end-users. For the investigation of the first research question, the number of employed servers as a function of the number of available servers is displayed in Figure 1(a). The curve identified by OPT presents the results considering hierarchical objective functions, which means that the number of servers was minimized only after the number of strict-latency requests was maximized, and the total number of requests processed in the fog was only optimized at the end. Curves identified by `SERXX` represent solutions that allow reducing the processed workload in order to reduce the number of employed servers in relation to OPT. Similarly, curves identified by `SERXX` represent the solutions that allow a degradation of `XX` % in the number of servers in relation to OPT, using more servers in the solution. A notable result is the one obtained by allowing 5 % of degradation in the processed workload ($STR5$) with 2048 available servers: less than 400 servers are used compared to OPT, which accounts to about 30 % of savings in server costs. This happens since the removal of one or two servers from each fog node does not lead to great blockage. To fully process all workloads, many fog nodes employ servers that process only a small number of strict workloads, remaining idle for long periods. Thus, even if the acceptable degradation is small, high infrastructure costs could be avoided.

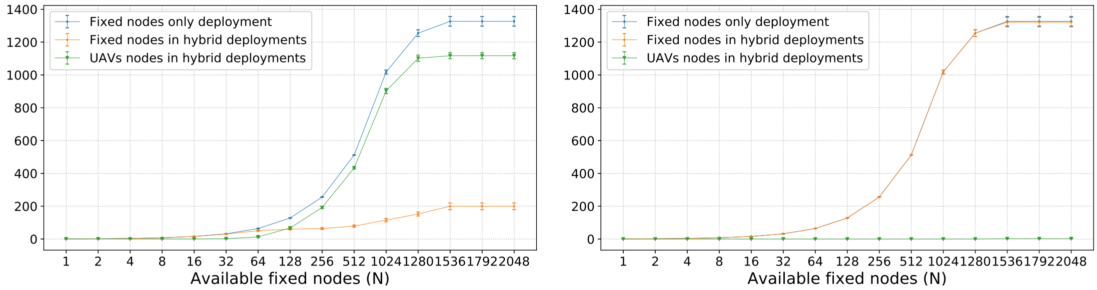
For the investigation of the energy consumption of end-user devices, the results of the EDTA algorithm were compared to the optimal results for a small number of locations. Figure 1(b) shows the energy consumed by all end-user devices as a function of the number of available servers for both EDTA and the optimal solution (OPT). The availability of fog nodes at all locations ($N \geq 128$) reduced the energy spent since users do not need to transmit data over very long links, reducing the energy consumption by



(a) Employed servers for different policies.

(b) Energy consumption for EDTA and the optimal solution.

Figure 1. Results for the terrestrial infrastructure.



(a) Fixed servers and UAVs with the same cost.

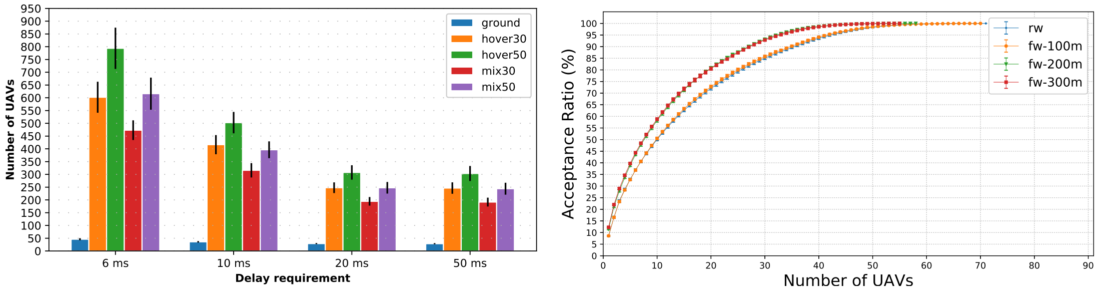
(b) Fixed servers four times cheaper than a UAV.

Figure 2. Number of fixed nodes and UAVs for different deployments.

about 40 % in comparison with a scenario with a single fog node ($N = 1$). The employment of the proposed algorithm promotes excellent energy savings compared to the exact solution, minimizing battery drain and allowing end-users to remain connected longer.

In the remainder of this section, results for aerial infrastructures are discussed. We first evaluated whether terrestrial fog nodes could be replaced by fog nodes mounted on UAVs. Figure 2(a) shows the number of employed devices as a function of the number of available servers (N). Two scenarios are considered: one with only fixed nodes, and another one with fixed and aerial fog nodes. For $N < 128$, fixed servers were heavily used, which prevents using UAVs due to their limited battery capacity to operate for long periods. When more devices are available for deployment, a large number of fixed servers is replaced by UAVs, with only about 200 fixed servers not replaced by UAVs. This implies that only 20 % of the infrastructure could employ terrestrial nodes. However, these results were obtained assuming that fixed servers and UAVs have the same cost. A UAV was only employed if its cost was less than or equal the cost of the replaced servers. Figure 2(b) displays results that assume a UAV costs four times the price a fixed server, which is more realistic nowadays. In this case, employing several UAVs is not advantageous: the average number of employed UAVs is very close to zero. This happens because using a UAV to replace servers in different locations is not always possible since the flights between the locations lead to a quick drain of the battery. Results of this investigation have revealed that a fog computing deployment with a large number of UAVs depends on the prices of aircraft being close to that of traditional servers.

The performance of rotary-wing and fixed-wing UAVs as fog nodes was assessed



(a) Average number of rotary-wing UAVs to process all requests for different deployment scenarios and delay requirements.

(b) Acceptance ratio for deployments with rotary-wing and fixed-wing UAVs

Figure 3. Results for infrastructures with only rotary-wing or fixed-wing UAVs.

in our investigation. For rotary-wing UAVs, different deployment scenarios were considered: under `ground` scenario, UAVs flies to a location and land, temporarily operating on the ground; under `hoverXX` scenarios, UAVs hover all the time at a constant height of `XX` (30 or 50) meters; finally, under `mixXX` scenarios, UAVs hover all constant height (`XX`, 30 or 50 meters) while they are processing end-users' requests, but they land when there are no requests being submitted. Moreover, different delay requirements were considered, given by the number of milliseconds to transmit the workload to the infrastructure using the wireless channel. Figure 3(a) shows the number of required rotary-wing UAVs to process all requests for different deployment scenarios and delay requirements. The `ground` deployment required the smallest number of UAVs since the energy consumption of UAVs is minimal, allowing a long operation. When the height is fixed, about 20 % of the budget can be reduced if UAVs can land during standby (`mixXX` instead of `hoverXX`). This reveals that UAVs benefit from sharing landing spots at strategic locations to reduce the CAPEX. Moreover, for stringent delay requirements (6 ms), more than 600 UAVs were needed with `hover30` deployment, and more than 800 for the `hover50`. However, when delays are flexible, these values are less than 250 and 300 UAVs, respectively. This happens because, when limited delays are required, UAVs cannot process requests from distant users, which leads to solutions with UAVs at more locations.

Finally, employing fixed-wing UAVs as fog nodes was also investigated. Fixed-wing UAVs cannot hover and cannot easily land on limited spaces. Therefore, circular trajectories with different radii were assumed. For the sake of comparison, these results were compared to those of rotary-wing UAVs. Figure 3(b) shows the acceptance ratio as a function of the number of available UAVs; the acceptance ratio is calculated as the ratio between the number of requests processed by the infrastructure and the total number of requests submitted. For the `fw-XXXm`, `XXX` is the radius in meters and `rw` results are the ones obtained by rotary-wing UAVs. Results of `fw-200m` and `fw-300m` deployments resulted in greater acceptance since a large radius requires less energy from the UAV battery, which allows a longer operational time, leading to more requests processed. The energy consumption for a small radius is so high that it is close to the one of a rotary-wing UAV, making results of `fw-100m` and `rw` similar. In this case, the rotary-wing UAV can be more advantageous since it will provide a stable wireless channel due to its ability to hover. Nevertheless, fixed-wing UAVs are generally the best alternative as long as the radius of circular trajectory is the maximum possible to extend the battery operation.

6. Publications

Results of the thesis were published as papers in international journals and conferences: [da Silva and Fonseca 2018] in the IEEE International Conference on Communications, [da Silva and da Fonseca 2019] in the MDPI Sensors (Impact Factor 3.847), [da Silva and da Fonseca 2020] in the IEEE Transactions on Green Communications and Networking (Impact Factor 3.525), [da Silva et al. 2021] in the IEEE Wireless Communications (Impact Factor 12.777), [da Silva and da Fonseca 2022a] in the IEEE Global Communications Conference, and [da Silva and da Fonseca 2023] in the Elsevier Vehicular Communications (Impact Factor 8.373).

Collaborations during the Ph.D. resulted in journal publications: [Lago et al. 2021] in the Elsevier Simulation Modelling Practice and Theory (Impact Factor 4.199), and [Montoya-Munoz et al. 2022] in the Computers and Electronics in Agriculture (Impact Factor 6.757).

7. Conclusions

The fog node location is a crucial decision in the deployment of a fog computing infrastructure, requiring good solutions to provide efficient service to fog users. The thesis reviewed in this paper introduced novel algorithms and mathematical formulations to the fog node location problem, always considering users' demands variable in time and space. Results obtained showed that the proposed solutions were quite efficient to design a fog computing infrastructure based on terrestrial or aerial fog nodes. The work about the terrestrial infrastructure showed how to design an efficient fog infrastructure with the available resources and showed that the energy of mobile devices can be reduced with a proper fog node location. The investigation of the aerial infrastructure showed that a large part of the terrestrial infrastructure could be replaced by UAVs. It also showed that both rotary-wing and fixed-wing UAVs can be pretty efficient as fog nodes to complement the terrestrial network as long as their limitations are accounted for. This work can be extended in different manners, such as considering more complex user applications, evaluating real testbeds, employing infrastructure for recharging UAVs in the middle of the operation, and integrating other types of aircraft in the infrastructure.

Acknowledgment

This study was financed in part by Google through the Latin America Research Awards. This study was financed in part by CNPq, grant 140464/20182. This study was financed in part by the São Paulo Research Foundation (FAPESP), grant 2015/24494-8.

References

Barlacchi, G., Barlacchi, G., De Nadai, M., Larcher, R., Casella, A., Chitic, C., Torrisi, G., Antonelli, F., Vespignani, A., Pentland, A., and Lepri, B. (2015). A multi-source dataset of urban life in the city of Milan and the Province of Trentin. *Scientific Data*, 2(1):150055.

da Silva, R. A. C. and da Fonseca, N. L. S. (2019). On the location of fog nodes in fog-cloud infrastructures. *Sensors*, 19(11).

da Silva, R. A. C. and da Fonseca, N. L. S. (2020). Location of fog nodes for reduction of energy consumption of end-user devices. *IEEE Transactions on Green Communications and Networking*, 4(2):593–605.

da Silva, R. A. C. and da Fonseca, N. L. S. (2022a). Design of fog computing infrastructures with rotary-wing uavs. In *GLOBECOM 2022 - 2022 IEEE Global Communications Conference*, pages 5789–5794.

da Silva, R. A. C. and da Fonseca, N. L. S. (2022b). *The Fog Node Location Problem*. PhD thesis, Universidade Estadual de Campinas.

da Silva, R. A. C. and da Fonseca, N. L. S. (2023). Location of fog nodes mounted on fixed-wing UAVs. *Vehicular Communications*, 41:100600.

da Silva, R. A. C., da Fonseca, N. L. S., and Boutaba, R. (2021). Evaluation of the employment of uavs as fog nodes. *IEEE Wireless Communications*, 28(5):20–27.

da Silva, R. A. C. and Fonseca, N. L. S. d. (2018). Resource allocation mechanism for a fog-cloud infrastructure. In *2018 IEEE International Conference on Communications (ICC)*, pages 1–6.

Kim, W. and Chung, S. (2018). User-participatory fog computing architecture and its management schemes for improving feasibility. *IEEE Access*, 6:20262–20278.

Lago, D. G., da Silva, R. A., Madeira, E. R., da Fonseca, N. L., and Medhi, D. (2021). Sinergycloud: A simulator for evaluation of energy consumption in data centers and hybrid clouds. *Simulation Modelling Practice and Theory*, 110:102329.

Larumbe, F. and Sansò, B. (2012). Cloptimus: A multi-objective cloud data center and software component location framework. In *2012 IEEE 1st International Conference on Cloud Networking (CLOUDNET)*, pages 23–28.

Larumbe, F. and Sansò, B. (2013). A tabu search algorithm for the location of data centers and software components in green cloud computing networks. *IEEE Transactions on Cloud Computing*, 1(1):22–35.

Lähderanta, T., Leppänen, T., Ruha, L., Lovén, L., Harjula, E., Ylianttila, M., Riekki, J., and Sillanpää, M. J. (2021). Edge computing server placement with capacitated location allocation. *Journal of Parallel and Distributed Computing*, 153:130–149.

Montoya-Munoz, A. I., da Silva, R. A., Rendon, O. M. C., and da Fonseca, N. L. (2022). Reliability provisioning for fog nodes in smart farming iot-fog-cloud continuum. *Computers and Electronics in Agriculture*, 200:107252.

Wang, J., Liu, K., and Pan, J. (2020). Online UAV-mounted edge server dispatching for mobile-to-mobile edge computing. *IEEE Internet Things J.*, 7(2):1375–1386.

Zhao, Z., Min, G., Gao, W., Wu, Y., Duan, H., and Ni, Q. (2018). Deploying edge computing nodes for large-scale IoT: A diversity aware approach. *IEEE Internet of Things Journal*, 5(5):3606–3614.

Zhou, Y., Pan, C., Yeoh, P. L., Wang, K., Elkashlan, M., Vucetic, B., and Li, Y. (2021). Communication-and-computing latency minimization for UAV-enabled virtual reality delivery systems. *IEEE Transactions on Communications*, 69(3):1723–1735.