CLEI Electronic Journal, Volume 27, Issue 3, paper 2, August 2024

The Fog Node Location Problem

Rodrigo A. C. da Silva
Center for Mathematics, Computing and Cognition,
Federal University of ABC (UFABC), Santo André, SP, Brazil
cardoso.rodrigo@ufabc.edu.br

and

Nelson L. S. da Fonseca
Institute of Computing, University of Campinas (Unicamp), Campinas, SP, Brazil
nfonseca@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.

Keywords: Fog computing, Cloud computing, Facility location, Unmanned Aerial Vehicle.

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 [1].

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 [2, 3] 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 [4]. 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" [5] 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 communicate with end-users via wireless interfaces. Other solutions 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. The thesis modeled variations of the fog node location problem using mixed-integer linear programming (MILP) and integer linear programming models (ILP), and it also proposed heuristic algorithms. The following list summarizes these contributions:

- A multi-objective MILP model for the fog node location problem with fixed fog nodes aimed at optimizing the acceptance ratio of requests, infrastructure cost, and fog node usage;
- A multi-objective MILP model and the Energy and Demand Trade-off algorithm for the fog node location problem with fixed fog nodes aimed at optimizing the acceptance ratio of requests, the energy spent by end-user devices, and the infrastructure cost;
- A multi-objective MILP model and the UAV Fog Node Location algorithm for the fog node location problem with fixed and mobile fog nodes aimed at optimizing the acceptance ratio of requests, and infrastructure cost;
- A multi-objective ILP model and the Sequential UAV Fog Node Location algorithm for the fog node location problem with mobile fog nodes mounted on rotary-wing UAVs aimed at optimizing the acceptance ratio of requests, infrastructure cost, and latency;
- A multi-objective ILP model and the Spatio-Temporal UAV Fog Node Location algorithm for the fog node location problem with mobile fog nodes mounted on fixed-wing UAVs aimed at optimizing the acceptance ratio of requests, infrastructure cost, and latency;
- The Gaussian Process Regression for Fog-Cloud Allocation mechanism for the allocation of tasks either in the fog or in the cloud, aimed at optimizing the acceptance ratio of requests and fog node usage.

The thesis expands the state-of-the-art in different manners. All algorithms and formulations were designed for fog computing and can be applied for different needs, such as in cellular networks, smart cities, and smart farming. The proposals considered variable demands in time and space, which addresses the need of mobile users. Moreover, a significant part of the thesis was devoted to the study of UAVs as fog nodes, showing that such vehicles can greatly impact future networks.

The remainder of this paper is organized as follows. Section 2 reviews related work. Section 3 introduces the main contributions of the thesis. Section 4 presents a mathematical formulation to the fog node location problem introduced by the thesis. Section 5 summarizes the main results of the thesis. Finally, Section 6 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, edge servers were considered in [6, 7]. Zhao et al. [6] proposed a solution for selecting the position of edge 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 until all users have been covered by an edge server. Lähderanta et al. [7] 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.

Oliveira and Viana [8] proposed a solution for the WiFi hotspot location that attempts to maximize the offloaded traffic by mobile users limited by the number of devices available for this deployment. The solution first identifies points of interest, i.e., locations commonly visited by users, to serve as candidate locations. It employed a graph that connects mobile users and points of interest. Results showed that the number of hotspots could be reduced by the use of the solution.

Larumbe and Sansò proposed solutions to the location of cloud data centers in a backbone network. Their work employed a MILP formulation [9] and a scalable tabu search algorithm [10] to decide on the location of data centers to minimize delay, energy consumption, costs, and the emission of greenhouse gases.

The placement of cloudlets has been explored in previous papers [11, 12]. Jia et al. [11] determined the location of cloudlets to reduce the delay of user tasks. Fan and Ansari [12] included the cloudlet cost in the decision. Using an optimization model, they showed that their solution can reduce deployment cost as long as additional delays are acceptable. All these papers on facility location [6, 7, 8, 9, 10, 11, 12] were designed for various types of infrastructures, whereas the work in the reviewed thesis considered the inherent characteristics of fog computing.

Some authors employed UAVs as fog nodes. For instance, the authors of [13] attempted to reduce the energy consumed by end-user devices that offload tasks to UAVs while optimizing the UAV trajectory. Their offload scheme significantly reduced the energy consumed by mobile devices in comparison with scenarios with processing only in the device. In [14], 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. [15] 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 Contributions

Some contributions of the thesis assumed an infrastructure with only terrestrial fog nodes. 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 as well as reduce capital expenditure (CAPEX). We formulated the optimization problem as a multicriterial MILP model with three objective functions. First, the model maximizes the processing of workload with strict latency requirements. Second, the total number of dedicated 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.

In a 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, the thesis also 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.

The third research question of the thesis 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 fog node location problem approaches aimed at evaluating infrastructures with only UAVs. The fourth research question is *"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?"*, while the fifth research question is *"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 to understand the role of different types of UAVs as fog nodes 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.

Finally, the thesis also analyzed how computational resources should be allocated to end-users' requests in fog and cloud computing. The sixth and last research question was *"Which tasks of an application should be processed by the fog and which ones should be processed by the cloud?"*. The answer to it was the Gaussian Process Regression for Fog-Cloud Allocation mechanism that considers the processing and networking requirement of applications to reduce blockage of user requests.

As part of the contributions, results of the thesis were published as papers in international journals and conferences [16, 17, 18, 19, 20, 21]. There were also collaborations during the Ph.D. that resulted in two journal publications [22, 23].

4 Mathematical formulation

The thesis proposed five linear programming formulations as well as five heuristic algorithms. This section reviews the investigation of the third research question in more details. The detailed investigations of the other research questions were not included in this paper due to space constraints; they are available in the thesis [5].

In the considered system, end-users request processing applications at different locations. Some applications have strict latency requirements and need to be processed in a nearby fog node. If there is no fog node available, the request is blocked since it cannot be migrated to more distant nodes due to latency requirements. Fog nodes serve users in their coverage area and can have more than one compute server.

UAVs can serve as fog nodes due to their capability of carrying a small processing server. In this investigation, UAVs can be in one of four different states: turned-off, stand-by, processing, and flying. In the turned-off state, UAVs do not consume energy, but they cannot be used. When they are needed, UAVs start their operation until their service is no longer required or their battery has been depleted. In the stand-by state, UAVs are powered on but they are not processing any workload. From stand-by, they can be quickly switched to another state. In the processing state, UAVs are processing end-users' workload and consume energy for processing and data transmission. In the flying state, the UAV is flying to a different location to augmented the computing capacity of the destination node. The flying state is the one that consumes the largest amount of energy, proportional to the distance traveled and the speed of traveling, both horizontally and vertically.

The fog node location problem consists of deciding the locations where fog nodes should be deployed, based on the candidate locations for fog nodes, the workload demand, and the available budget. The set of locations selected and the number of servers at each node is decided. Additionally, the number of UAVs and their flight plan is determined. The primary goal is to process the maximum possible amount of workload, and the secondary goal is to reduce the infrastructure cost. The problem was formulated as an optimization problem and modeled as a mixed-integer linear programming formulation. The model is bi-criteria and optimizes the served workload as well as the deployment cost. The notation in Table 1 is used in the formulation.

The model considers discrete time and $\mathcal{T}$ is the set of discretized intervals. The set of all candidate locations for fog nodes is given by $\mathcal{L}$. The workload demand is variable in each location over time, indicated by $\mathcal{W}_{lt}$. All fixed servers have the same capacity, their cost and capacity are $\mathcal{C}^S$ and $\mathcal{K}^S$, respectively, and UAVs cost and capacity are $\mathcal{C}^U$ and $\mathcal{K}^U$, respectively.

The output of the formulation is the workload served, the fixed servers used, number of UAVs employed, their trajectories as well as their states during the evaluation. The served workload is given by w_{lt} for each location l and time t. The number of fixed servers in each location l is indicated by n_l, and the number of used UAVs is n^u. $\mathcal{U}$ is set of all possible UAVs. The total battery capacity of any UAV is denoted by E, and energy cost for a trip between locations k and l is indicated by E_{kl}. The number of intervals required by a UAV to travel between a location k to a location l is given by D_{kl}.

The states of UAV are indicated by the variables a_{ult}, s_{ult}, f_{uklt}, p^f, and p^l; a_{ult} indicates that the UAV u is in processing state at location l during time t. Similarly, s_{ult} indicates that the UAV u is inactive at location l and time t. Being inactive means that the UAV is either in stand-by or in power-off state. A UAV plan starts in the turned-off state; then, after being turned on, it can be in the other three states. If turned off or the battery capacity has exhausted, it goes to the power-off state. The power-off state is denoted by p^f_{ut} and p^l_{ut}. If a UAV u remains in power-off state in the first T^f time intervals, $p^f_{ut} = 1$ for all $1 \leq t \leq T^f$. Similarly, if a UAV u goes to power-off state at time T^l, then $p^l_{ut} = 1$ for all $T^l \leq t \leq T$. The flying state is represented by the variable f_{uklt} that indicates whether the UAV u is traveling between locations k and l during time t. Finally, if a UAV u is not required, i.e., if it remains inactive during all time intervals in $\mathcal{T}$, then $i_u = 1$.

The formulation of the fog node location problem is given by Equations (1)–(18):

Table 1: Notation used in formulation for the third research question.

Input parameters	
Notation	**Description**
General parameters	
$\mathcal{T}$	Set of discrete time intervals: $\mathcal{T} = \{1, 2, \ldots, T\}$
$\mathcal{L}$	Set of candidate locations for fog nodes: $\mathcal{L} = \{1, 2, \ldots, L\}$
$\mathcal{W}_{lt}, l \in \mathcal{L}, t \in \mathcal{T}$	Workload at location l during time slot t
Cost	
$\mathcal{C}^S$	Cost of a single fixed server
$\mathcal{C}^U$	Cost of a UAV carring a mobile server
$\mathcal{C}$	Available budget
Capacity	
$\mathcal{K}^S$	Capacity of a single fixed server
$\mathcal{K}^U$	Capacity of a UAV mobile server
UAV parameters	
$\mathcal{U}$	Set of available UAVs, $\mathcal{U} = \{1, \ldots, U\}$
E	Total battery capacity of a UAV
E^P	Energy required by a UAV in processing state during one time interval
E^S	Energy required by a UAV in stand-by state during one time interval
$E_{kl}, k, l \in \mathcal{L}$	Energy consumption per time window required by a UAV to travel between locations l and k
$D_{kl}, k, l \in \mathcal{L}$	Number of discrete time intervals required by a UAV to travel between locations l and k

Decision variables	
Notation	**Description**
$w_{lt}, l \in \mathcal{L}, t \in \mathcal{T}$	Workload originating at location $l \in \mathcal{L}$ at time $t \in \mathcal{T}$ and processed by the fog node
$n_l, l \in \mathcal{L}$	The number of fixed servers deployed at the fixed fog node located at l
n^u	Number of required UAVs
$i_u, u \in \mathcal{U}, l \in \mathcal{L}, t \in \mathcal{T}$	1 if UAV u was in inactive, i.e. it was in stand-by state over all time intervals; 0 otherwise
$a_{ult}, u \in \mathcal{U}, l \in \mathcal{L}, t \in \mathcal{T}$	1 if UAV u is at location l at time t in processing state; 0 otherwise
$s_{ult}, u \in \mathcal{U}, l \in \mathcal{L}, t \in \mathcal{T}$	1 if UAV u is at location l at time t in stand-by or power-off state; 0 otherwise
$f_{uklt}, u \in \mathcal{U}, k \in \mathcal{L}, l \in \mathcal{L}, t \in \{2, \ldots, T\}$	1 if UAV u is traveling between locations k and l at time t; 0 otherwise
$p^f_{ut}, u \in \mathcal{U}, t \in \mathcal{T}$	1 if UAV u is in power-off state at all time intervals $t*$ such that $t* \leq t$; 0 otherwise
$p^l_{ut}, u \in \mathcal{U}, t \in \mathcal{T}$	1 if UAV u is in power-off state at all time intervals $t*$ such that $t* \geq t$; 0 otherwise

$$maximize \sum_{l\in\mathcal{L}} \sum_{t\in\mathcal{T}} w_{lt} \tag{1}$$

$$minimize \quad C^S \sum_{l\in\mathcal{L}} n_l + C^U n^u \tag{2}$$

$$\mathcal{C}^S \sum_{l\in\mathcal{L}} n_l + \mathcal{C}^U n^u \leq \mathcal{C} \tag{3}$$

$$w_{lt} \leq \mathcal{K}^S n_l + \mathcal{K}^U \sum_{u\in\mathcal{U}} a_{ult}, l \in \mathcal{L}, t \in \mathcal{T} \tag{4}$$

$$w_{lt} \leq \mathcal{W}_{lt}, l \in \mathcal{L}, t \in \mathcal{T} \tag{5}$$

$$\sum_{k\in\mathcal{L}} \Big(a_{ukt} + s_{ukt} + \sum_{l\in\mathcal{L}\setminus\{k\}} f_{uklt} \Big) = 1, u \in \mathcal{U}, t \in \mathcal{T} \tag{6}$$

$$\begin{aligned} a_{ult} + s_{ult} \leq a_{ul(t-1)} + s_{ul(t-1)} + \sum_{k\in\mathcal{L}\setminus\{l\}} f_{ukl(t-1)}, \\ u \in \mathcal{U}, l \in \mathcal{L}, t \in \{2, \ldots, T\} \end{aligned} \tag{7}$$

$$\begin{aligned} f_{uklt} \leq a_{uk(t-1)} + s_{uk(t-1)} + f_{ukl(t-1)}, \\ u \in \mathcal{U}, k \in \mathcal{L}, l \in \mathcal{L} \setminus \{k\}, t \in \{2, \ldots, T\} \end{aligned} \tag{8}$$

$$\begin{aligned} a_{ukv} + s_{ukv} \leq 1 - a_{ult} - s_{ult}, \\ u \in \mathcal{U}, k \in \mathcal{L}, l \in \mathcal{L} \setminus \{k\}, t \in \mathcal{T}, \\ v \in \{t+1, \ldots, t + D_{lk}\} \end{aligned} \tag{9}$$

$$i_u \leq \frac{1}{T} \sum_{l\in\mathcal{L}} \sum_{t\in\mathcal{T}} s_{ult}, u \in \mathcal{U} \tag{10}$$

$$i_u > \Big(\sum_{l\in\mathcal{L}} \sum_{t\in\mathcal{T}} s_{ult} \Big) - T, u \in \mathcal{U} \tag{11}$$

$$p^f_{ut} \leq \sum_{l\in\mathcal{L}} s_{ult}, u \in \mathcal{U}, t \in \mathcal{T} \tag{12}$$

$$p^l_{ut} \leq \sum_{l\in\mathcal{L}} s_{ult}, u \in \mathcal{U}, t \in \mathcal{T} \tag{13}$$

$$p^f_{ut} \leq p^f_{u(t-1)}, u \in \mathcal{U}, t \in \{2, \ldots, T\} \tag{14}$$

$$p^l_{ut} \leq p^l_{u(t+1)}, u \in \mathcal{U}, t \in \{1, \ldots, T-1\} \tag{15}$$

$$p^f_{ut} + p^l_{ut} \leq 1, u \in \mathcal{U}, t \in \mathcal{T} \tag{16}$$

$$n^u = U - \sum_{u\in\mathcal{U}} i_u \tag{17}$$

$$\begin{aligned} \sum_{t\in\mathcal{T}} \sum_{l\in\mathcal{L}} E^A \cdot a_{ult} + \\ \sum_{t\in\mathcal{T}} (\sum_{l\in\mathcal{L}} E^S \cdot s_{ult} - p^f_{ut} - p^l_{ut}) + \\ \sum_{t\in\{2,\ldots,T\}} \sum_{k\in\mathcal{L}} \sum_{l\in\mathcal{L}\setminus\{k\}} E_{kl} \cdot m_{uklt} \leq E, \\ u \in \mathcal{U} \end{aligned} \tag{18}$$

The objective function in Equation (1) is the maximization of the served workload. Equation (2) minimizes of the cost to build the infrastructure, considering the cost of both fixed servers and UAVs. These goals are subject to the constraints given by Equations (3)–(18). Equation (3) limits the cost of the infrastructure to the available budget. Equation (4) guarantees that the workload served is smaller than the capacity of the fog node, considering both fixed and UAV servers. Equation (5) limits the workload served to the demand at each location in each time interval. Equation (6) guarantees that every UAV is at exactly one single state at any moment. Equations (7)–(8) check whether the transitions between states are valid considering time and location limitations. In sequential time intervals, UAVs cannot change their location without going to travel state (Equation (7)), and the travel state can only be reached if the UAV was in travel state or in the source location in the previous time interval (Equation (8)). Equation (9) guarantees that the minimum trip time is respected when a UAV flies between different locations. Equations (10)–(11) check whether any UAV was inactive, i.e., in stand-by over all intervals. Equations (12)–(16) guarantee the right value of p^f_{ut} and p^l_{ut}, identifying the intervals of power-off state. Equation (17) accounts n^u, the number of required UAVs. Finally, (18) guarantees that the autonomy of each UAV is respected by summing the energy spent in each state and limiting it to the battery capacity.

The deployment planning given by this formulation is a network design problem, and, as such, is typically solved off-line. However, solving this problem optimally using existing solvers does not scale to large problem instances. The modeling of potential flight routes, the location of UAVs at every time interval, and the UAVs activity/inactivity periods lead to an exponential growth in the number of constraints. To circumvent these limitations, the UAV Fog Node Location (UFL) algorithm was proposed.

The UFL algorithm functions as follows: initially, it solves the formulation with only fixed nodes while adhering to the pre-defined budget that limits the number of deployed servers and nodes. From the derived solution, the algorithm identifies servers that might be substituted by UAVs; those servers are typically underutilized and primarily deployed to address peak demands. Subsequently, the UFL algorithm attempts to employ UAVs to cover multiple locations at different times, aiming to diminish deployment costs. The algorithm considers the ratio between the cost of UAVs and the cost of fixed servers. A formal definition and a more comprehensive discussion of the steps of the UFL algorithm can be found in the thesis [5].

5 Results

In this section, the main results obtained in the thesis are presented. Detailed analyses can be seen in the thesis [5]. 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 [24] 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. 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 (second research question), the results of the EDTA algorithm were compared to the optimal results for a small number of locations (Figure 2). Figure 2(a) displays the acceptance ratio for different application classes: fog-only applications can only be processed in the fog, fog-device applications can be processed either in the fog or by the mobile device,

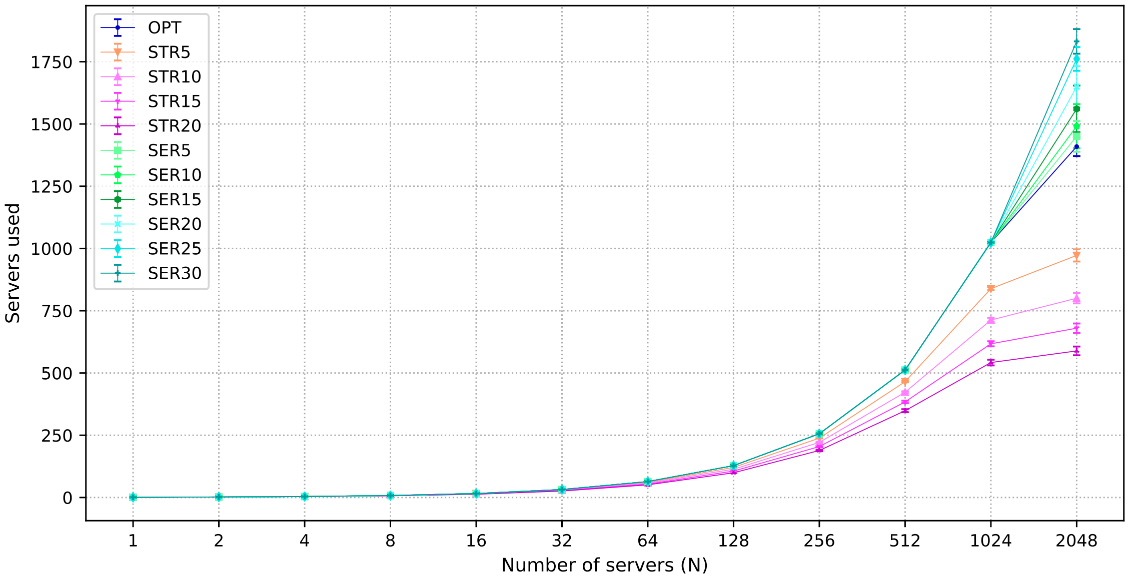
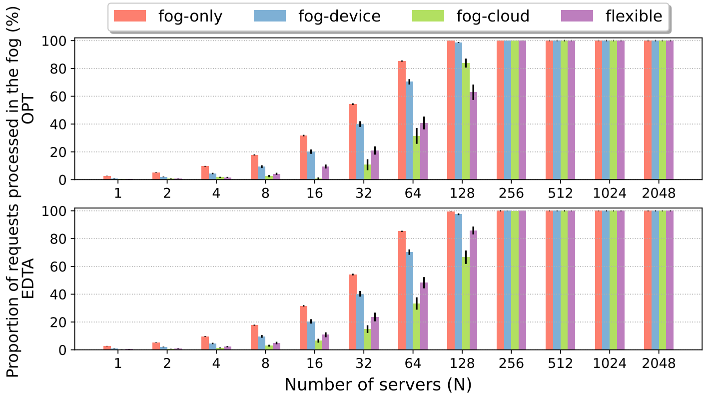


Figure 1: Employed terrestrial servers for different policies. Figure taken from [17].

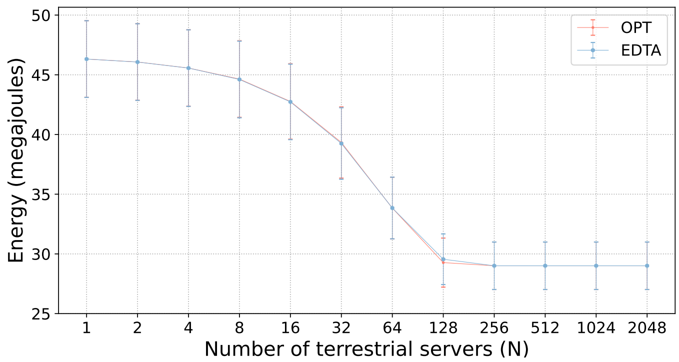
fog-cloud applications can be processed in the fog or in the cloud, and flexible applications can be processed anywhere. The number of applications processed in the fog increases as a function of the number of servers (N) up to $N = 128$. The proportion of the fog-only class processed in the fog is greater than that of the other classes because this is the main goal of the solution. For all values of N, the EDTA produces results very close to those given by the OPT, which shows that the proposed algorithm manages to deploy servers in a efficient manner. Figure 2(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 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.

For the investigation of the third research question, Figure 3(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 3(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 in the thesis. In relation to rotary-wing UAVs (fourth research question), different deployment scenarios were considered: under `ground` scenario, UAVs flies to a location and land, operating as a temporary fog node 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 4 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, in this case, 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

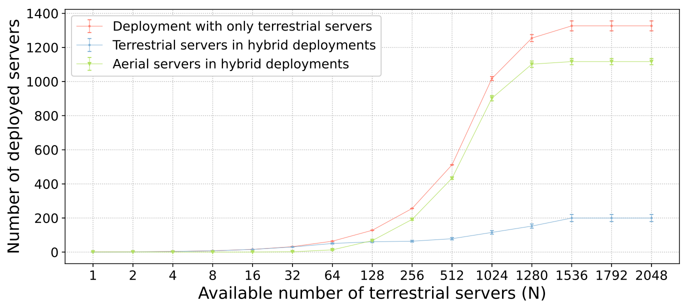


(a) Proportion of requests processed in the fog for the EDTA and the optimal solution.

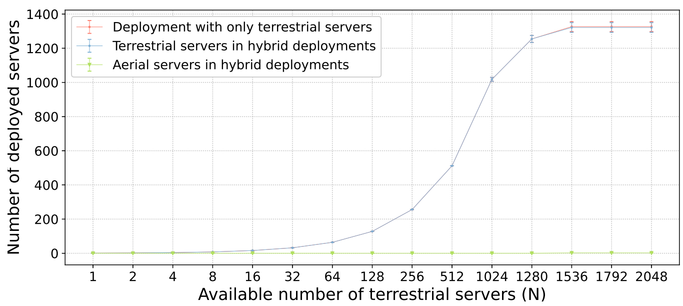


(b) Energy consumption for the EDTA and the optimal solution.

Figure 2: Results for reducing the energy consumption of end-user devices.



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



(b) UAVs four times more expensive than a terrestrial server.

Figure 3: Number of employed devices (terrestrial servers and UAVs) for different deployments.

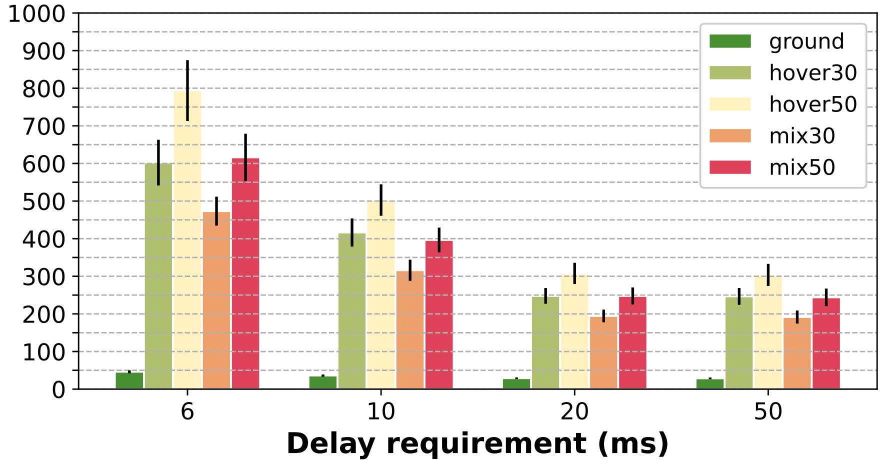


Figure 4: Average number of rotary-wing UAVs to process all requests for different deployment scenarios and delay requirements.

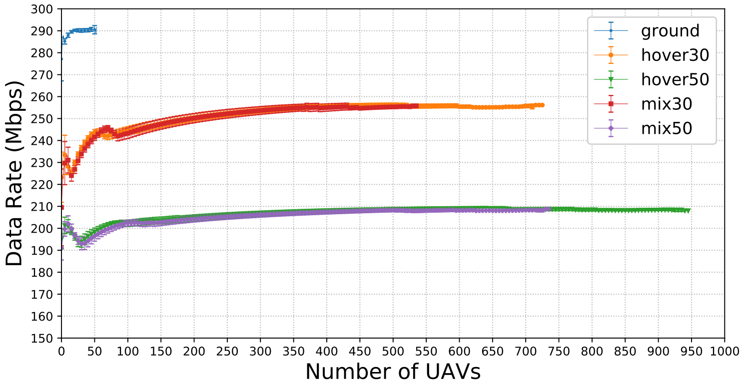
with `hover30` deployment, and more than 800 for the `hover50`. However, when delays are flexible, these values are less than 250 and 300 UAVs, respectively. This happens because, when limited delays are required, UAVs cannot process requests from distant users, which leads to solutions with UAVs at more locations.

To further explore the impact of the deployment scenario for rotary-wing UAVs, Figure 5 presents the data rates for the applications with a delay requiment of 6 ms (Figure 5(a)) and 50 ms (Figure 5(b)). The `ground` scenario produces the highest data rates, which is explained by the proximity with ground nodes that reduces the path loss. Nonetheless, even the smaller data rates are sufficient to provide fast communications between aerial and terrestrial nodes. The data rate is also affected by stricter delay requirements. For the strict requirement, in which 100 Kb should be transmitted in under 6 ms, data rates above 200 Mbps are needed, which requires that the UAVs should be deployed very close to the BSs serving the users. On the other hand, data rate less than 100 Mbps are needed under when the delay requirement is 50 ms. These results show that the more restricted delays, the closer to the users the aerial fog nodes should be.

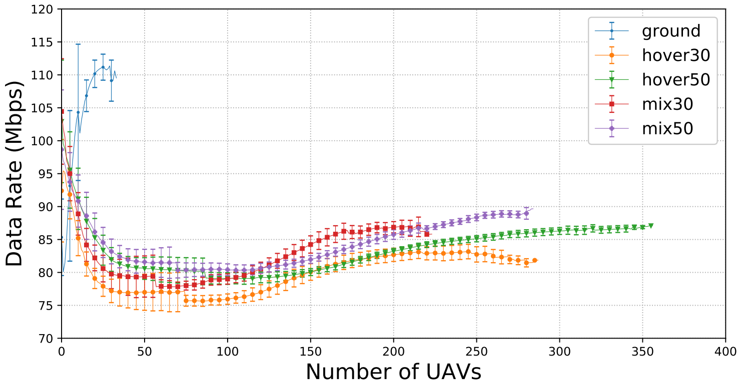
Finally, employing fixed-wing UAVs as fog nodes was also investigated (fifth research question). Fixed-wing UAVs cannot hover and cannot easily land on limited spaces. Therefore, to operate as a fog node, this work assumed they performed circular trajectories with different radii. For the sake of comparison, these results were compared to those of rotary-wing UAVs. Figure 6(a) 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.

The average data rate (Figure 6(b)) was in the range 65–75 Mbps, higher than the minimum requirement (50 Mbps). The highest data rate was that produced by the rotary-wing UAV deployment since such UAVs maintain the same position in the air, guaranteeing better wireless links. Differently, the data rate for the fixed-wing UAVs was calculated considering the circular trajectory that may increase the distance between UAVs and ground nodes. Nevertheless, trajectories with greater radii lead to only small reductions in data rate. In the experiments, fixed-wing UAVs with 300 m radii represented the most advantageous deployment, with small energy consumption and good wireless links. However, if rotary-wing UAVs are cheaper than fixed-wing ones, rotary-wing employment can be quite advantageous, since they 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.

The main results of the thesis can be summarised as follows. Dealing with variable demands in time and space is challenging when solving the fog node location problem. Results of the investigations of the first research questions showed that reducing the number of fog servers does not have a proportional impact on the acceptance ratio due to the distribution of demands. Therefore, even limited budgets could still be used to provide efficient infrastructures for end-users. Moreover, fog nodes can be located to minimize the battery drain of end-users' devices, allowing users to remain connected longer. Investigations that considered fog nodes mounted on UAVs showed that a large part of the terrestrial infrastructure could be replaced by



(a) Results for the 6 ms delay requirement.



(b) Results for the 50 ms delay requirement.

Figure 5: Average data rate as a function of the number of available rotary-wing UAVs.

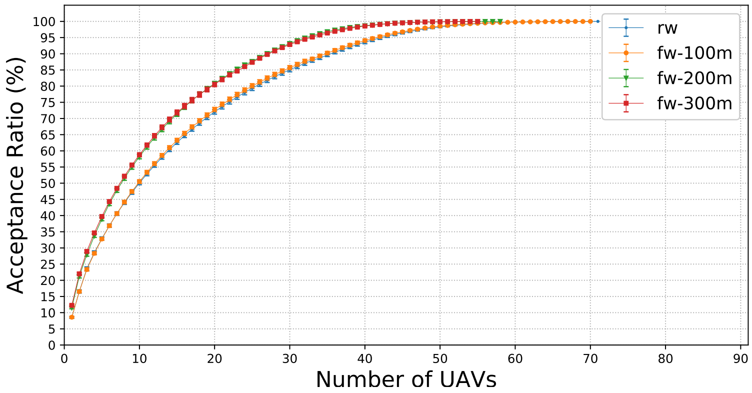
UAVs as long as the aircraft prices reduce in the future. The thesis also showed that different types of UAVs (rotary-wing and fixed-wing) can be employed as fog nodes, offering low-latency processing due to the physical proximity between them and the ground. UAVs present better flexibility than do other vehicles, but their energy consumption must be accounted for to guarantee an efficient operation. Both rotary-wing and fixed-wing UAVs can be used as fog nodes; the former is useful to provide stable and low latency processing, while the latter can endure a longer operation due to reduced energy consumption.

6 Conclusions

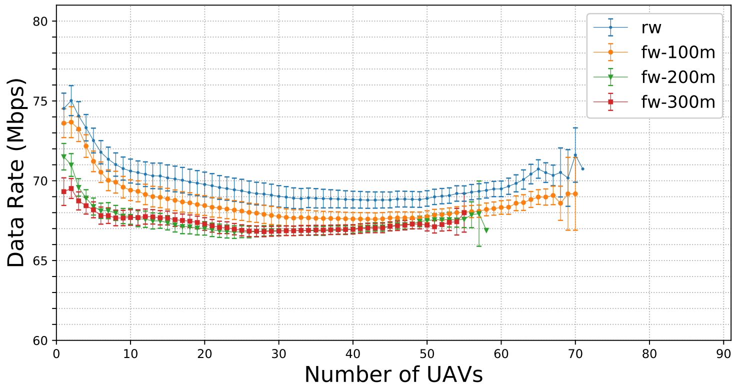
The fog node location is a crucial decision in the deployment of a fog computing infrastructure, requiring good solutions to provide efficient service to fog users. The thesis reviewed in this paper introduced novel algorithms and mathematical formulations to the fog node location problem, always considering 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 the Government of Canada through the Emerging Leaders in the Americas Program



(a) Acceptance ratio as a function of the number of available UAVs.



(b) Average data rate as a function of the number of available UAVs.

Figure 6: Results for the fixed-wing UAVs.

(ELAP). This study was financed in part by the National Council for Scientific and Technological Development (CNPq), grant 140464/20182. This study was financed in part by the São Paulo Research Foundation (FAPESP), grant 2015/24494-8.

References

[1] F. Bonomi, R. Milito, J. Zhu, and S. Addepalli, "Fog computing and its role in the Internet of Things," in *Proceedings of the First Edition of the MCC Workshop on Mobile Cloud Computing*, ser. MCC '12. New York, NY, USA: Association for Computing Machinery, 2012, p. 13â16.

[2] W. Kim and S. Chung, "User-participatory fog computing architecture and its management schemes for improving feasibility," *IEEE Access*, vol. 6, pp. 20 262–20 278, 2018.

[3] R. Vilalta, V. Lopez, A. Giorgetti, S. Peng, V. Orsini, L. Velasco, R. Serral-Gracia, D. Morris, S. De Fina, F. Cugini, P. Castoldi, A. Mayoral, R. Casellas, R. Martinez, C. Verikoukis, and R. Munoz, "Telcofog: a unified flexible fog and cloud computing architecture for 5G networks," *IEEE Communications Magazine*, vol. 55, no. 8, pp. 36–43, 2017.

[4] M. C. Gonzalez, C. A. Hidalgo, and A.-L. Barabasi, "Understanding individual human mobility patterns," *Nature*, vol. 453, no. 7196, pp. 779–782, June 2008.

[5] R. A. C. da Silva and N. L. S. da Fonseca, "The Fog Node Location Problem," Ph.D. dissertation, Universidade Estadual de Campinas, 6 2022.

[6] Z. Zhao, G. Min, W. Gao, Y. Wu, H. Duan, and Q. Ni, "Deploying edge computing nodes for large-scale IoT: a diversity aware approach," *IEEE Internet of Things Journal*, vol. 5, no. 5, pp. 3606–3614, Oct 2018.

[7] T. Lähderanta, T. Leppänen, L. Ruha, L. Lovén, E. Harjula, M. Ylianttila, J. Riekki, and M. J. Sillanpää, "Edge computing server placement with capacitated location allocation," *Journal of Parallel and Distributed Computing*, vol. 153, pp. 130–149, 2021.

[8] E. M. R. Oliveira and A. C. Viana, "From routine to network deployment for data offloading in metropolitan areas," in *2014 Eleventh Annual IEEE International Conference on Sensing, Communication, and Networking (SECON)*, June 2014, pp. 126–134.

[9] F. Larumbe and B. Sanso, "Cloptimus: A multi-objective cloud data center and software component location framework," in *2012 IEEE 1st International Conference on Cloud Networking (CLOUDNET)*, Nov 2012, pp. 23–28.

[10] ——, "A tabu search algorithm for the location of data centers and software components in green cloud computing networks," *IEEE Transactions on Cloud Computing*, vol. 1, no. 1, pp. 22–35, Jan 2013.

[11] M. Jia, J. Cao, and W. Liang, "Optimal cloudlet placement and user to cloudlet allocation in wireless metropolitan area networks," *IEEE Transactions on Cloud Computing*, vol. 5, no. 4, pp. 725–737, Oct 2017.

[12] Q. Fan and N. Ansari, "Cost aware cloudlet placement for big data processing at the edge," in *2017 IEEE International Conference on Communications (ICC)*, May 2017, pp. 1–6.

[13] S. Jeong, O. Simeone, and J. Kang, "Mobile edge computing via a UAV-mounted cloudlet: Optimization of bit allocation and path planning," *IEEE Trans. Veh. Technol.*, vol. 67, no. 3, pp. 2049–2063, 2018.

[14] J. Wang, K. Liu, and J. Pan, "Online UAV-mounted edge server dispatching for mobile-to-mobile edge computing," *IEEE Internet Things J.*, vol. 7, no. 2, pp. 1375–1386, Feb 2020.

[15] Y. Zhou, C. Pan, P. L. Yeoh, K. Wang, M. Elkashlan, B. Vucetic, and Y. Li, "Communication-and-computing latency minimization for UAV-enabled virtual reality delivery systems," *IEEE Transactions on Communications*, vol. 69, no. 3, pp. 1723–1735, 2021.

[16] R. A. C. da Silva and N. L. S. da Fonseca, "Resource allocation mechanism for a fog-cloud infrastructure," in *2018 IEEE International Conference on Communications (ICC)*, 2018, pp. 1–6.

[17] ——, "On the location of fog nodes in fog-cloud infrastructures," *Sensors*, vol. 19, no. 11, 2019.

[18] ——, "Location of fog nodes for reduction of energy consumption of end-user devices," *IEEE Transactions on Green Communications and Networking*, vol. 4, no. 2, pp. 593–605, 2020.

[19] R. A. C. da Silva, N. L. S. da Fonseca, and R. Boutaba, "Evaluation of the employment of UAVs as fog nodes," *IEEE Wireless Communications*, vol. 28, no. 5, pp. 20–27, 2021.

[20] R. A. C. da Silva and N. L. S. da Fonseca, "Design of fog computing infrastructures with rotary-wing UAVs," in *GLOBECOM 2022 - 2022 IEEE Global Communications Conference*, 2022, pp. 5789–5794.

[21] ——, "Location of fog nodes mounted on fixed-wing UAVs," *Vehicular Communications*, 2023.

[22] D. G. Lago, R. A. da Silva, E. R. Madeira, N. L. da Fonseca, and D. Medhi, "SinergyCloud: A simulator for evaluation of energy consumption in data centers and hybrid clouds," *Simulation Modelling Practice and Theory*, vol. 110, p. 102329, 2021.

[23] A. I. Montoya-Munoz, R. A. da Silva, O. M. C. Rendon, and N. L. da Fonseca, "Reliability provisioning for fog nodes in smart farming IoT-fog-cloud continuum," *Computers and Electronics in Agriculture*, vol. 200, p. 107252, 2022.

[24] G. Barlacchi, G. Barlacchi, M. De Nadai, R. Larcher, A. Casella, C. Chitic, G. Torrisi, F. Antonelli, A. Vespignani, A. Pentland, and B. Lepri, "A multi-source dataset of urban life in the city of Milan and the Province of Trentin," *Scientific Data*, vol. 2, no. 1, p. 150055, Oct 2015.