

# Cutting Throughput on the Edge: App-Aware Placement in Fog Computing

Francescomaria Faticanti, Francesco De Pellegrini, Domenico Siracusa, Daniele Santoro and Silvio Cretti<sup>◇</sup>

**Abstract**—Fog computing extends cloud computing technology to the edge of the infrastructure to let IoT applications access objects’ data with reduced latency, location awareness and dynamic computation. By displacing workloads from the central cloud to the edge devices, fog computing overcomes communication bottlenecks avoiding raw data transfer to the central cloud, thus paving the way for the next generation IoT-based applications.

In this paper we study scheduling and placement of applications in fog computing, which is key to ensure profitability for the involved stakeholders. We consider a scenario where the emerging microservice architecture allows for the design of applications as cascades of coupled microservice modules. It results into a mixed integer non linear problem involving constraints on both application data flows and computation placement. Due to the complexity of the original problem, we resort to a simplified version, which is further solved using a greedy algorithm. This algorithm is the core placement logic of the FogAtlas platform, a fog computing platform based on existing virtualization technologies.

Extensive numerical results validate the model and the scalability of the proposed solution, showing it attains performance close to the optimal solution and, in our real implementation, it scales well with respect to the number of served applications.

**Index Terms**—fog computing, microservice, resources allocation, placement

## I. INTRODUCTION

Fog computing adopts cloud technology to move computation to the edge. It promises to solve the core problem of data explosion in the IoT domain [1]. Instead of performing raw data transfer to the cloud, in fact, data flows generated from objects can be intercepted to extract information at the edge of network. This architectural choice prevents massive, diffused and continuous raw data injection which would ultimately create severe communication congestion [2]. Furthermore, compared to customary cloud-based IoT deployments, proximity to mobile or sensing devices lowers round-trip-time between objects and backends of processing applications [3].

Further incentive in the development of fog computing solutions include the standardization of IoT deployments, ease of management and maintenance of IoT services in industrial networks [4], and also overcoming privacy issues by confining raw data within specific geographical regions [5]. The fog system studied in this paper refers to FogAtlas, a platform designed to perform efficient deployment of fog computing applications according to the above guidelines.

The tradeoff in this context is represented by edge resource occupation: compared to standard cloud technologies – based

on overprovisioned datacenters – the business of edge infrastructure owners will not be able to rely on overprovisioning. Rather, they need to trade off premier service provision based on localized data processing and low round-trip time for storage, memory and processing capabilities of edge units [6], [7].

The paradigm of fog computing consists of a layered architecture, including a central cloud, a series of edge units, gateways to connect objects and, finally, objects which generate data and possibly actuate. Virtual machines or containers can run either in the central cloud, or over edge units, depending on the requirements of IoT-based applications.

To this respect, it is natural to assume that fog-native applications will adhere to the microservice paradigm [8]. Microservice applications, in fact, are made by the composition of multiple coupled modules, such as, e.g., a graphical user interface, a user repository, a web server, an image recognition module, a monitoring application, etc. Once interconnected using a specific communication and computing pattern, the microservice architecture delivers the intended functionality while preserving scalability, minimality and cohesiveness of the application. In fog computing, the modular structure is indeed appealing in order to simplify the dispatch of computing modules onto edge nodes.

Typically, the microservice components of an application can be deployed using independent containers. However, in this work, we make the baseline assumption that a fog application will be shipped using two modules. The rationale for such a minimal containerization is that all operations of monitoring and networking on the edge will be largely simplified. The first one – possibly a virtual machine hosting several containers in the cloud – will typically comprise microservice modules not involved in raw IoT data computation and can be hosted in the central cloud.

The second module, hosted on a single IoT container, comprises functionalities involving objects’ data processing. This container may reside either on the cloud or on edge nodes, depending on the scheduling operated by the fog orchestrator. We refer to the concrete example of application deployed on our FogAtlas platform, namely a plate recognition video application, able to be dispatched on an edge server close to a target video-camera. Stream mining is actually emerging as a core research field motivating fog-computing applications [9].

In such benchmark fog application, indeed, performing computing IoT operations directly on edge nodes provides a clear advantage in terms of bandwidth utilization. In fact, the raw video stream is filtered through an image detection

<sup>◇</sup>Fondazione Bruno Kessler, via Sommarive, 18 I-38123 Povo, Trento, Italy

Table I: Main notation used throughout the paper

Symbol	Meaning
$\mathcal{K}$	set of regions $ \mathcal{K}  = K$
$\mathcal{U}$	set of applications to be deployed $\mathcal{U} = \cup_{i=1}^K U_i$ , $ \mathcal{U}  = U$
$S_k$	set of server units in region $k$ , with $ S_i  = n_i, \forall i \in \mathcal{K}$ , $S_i = \{s_{i1}, \dots, s_{in_i}\}$
$S_0$	central cloud
$U_k$	set of applications requiring IoT data in region $k$
$\lambda_u^H/\lambda_u^L$	high/low throughput required by application $u$
$\Delta_u^H/\Delta_u^L$	large/small data unit of application $u$
$F_u$	output samples per second required by application $u$
$\mathbf{C}_{k_i}$	memory, storage and processing capacity of the $i$ -th server in region $k$ : $\mathbf{C}_{k_i} = (C_{k_i}^M, C_{k_i}^S, C_{k_i}^P)$
$\mathbf{c}_u$	memory, storage and processing requirements of application $u$ : $\mathbf{c}_u = (c_u^M, c_u^S, c_u^P)$
$x_{u,k,i} \in \{0, 1\}$	boolean variable indicating $u$ is placed on server unit $i$ of region $k$
$x_{u,k}$	$x_{u,k} = \sum_{i \in S_k} x_{u,k,i}$

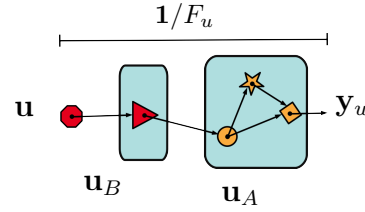
algorithm so that only tagged frames need to be forwarded to the cloud. As a result, only a small fraction of information is transferred toward the central cloud.

The main objective of this work is to describe an efficient placement of fog applications' modules either on the edge or in the cloud. In order to determine such a placement, constraints on computational and bandwidth requirements have to be factored in. We shall introduce first the general problem of how to place a batch of applications with sufficient computational resources and yet efficient network usage. Then, we shall describe our algorithmic solution.

The rest of the paper is organized as follows. In Sec. II we describe the system model, including the abstractions we use for the applications' architecture, the network infrastructure and applications' deployment configurations. In Sec. III we present the problem formulation, introducing the most general problem setting. The placement problem is addressed in Sec. IV by reduction to a multi-dimensional knapsack problem, which can be solved using a greedy algorithm. The FogAtlas platform is described in V and numerical results are reported in Sec. VI. A concluding section ends the paper.

## II. SYSTEM MODEL

We consider a fog system deployed over a set of geographic regions  $\mathcal{K} = \{1, \dots, K\}$ . Region  $k$  hosts a set  $S_k$  of edge servers or units. We denote  $s_{k_i}$ , with  $i \in \{1, \dots, n_k\}$  a specific edge unit deployed within the  $k$ -th region; for the sake of notation we denote the central cloud as  $S_0$ . The resources of edge unit  $s_{k_i}$  are represented by capacity vector  $\mathbf{C}_{k_i} = (C_{k_i}^M, C_{k_i}^P, C_{k_i}^S)$ . The first component of the capacity vector is the memory capacity. The second component is the processing capacity, which determines the maximum load which can be sustained on the edge unit. Finally, the third component denotes the storage capacity, i.e., the data volume that can be accommodated on the storage of the edge unit. We assume that the storage of a containerized application is handled on the same unit where the container is deployed, with the aim to reduce the communication costs.


 Figure 1: The modules cascade outputs a result  $y_u$  every  $1/F_u$  sec.

In region  $k$ , IoT devices serve data required by a set of applications  $U_k$ . From here on out, we identify the application and the device from which data are requested with same symbol. The extension of the following optimization framework in the case of multiple requests for same IoT device is immediate, by considering virtual replicas of a tagged IoT device. We say that application  $u$  “belongs” to a given region because the IoT object is located there. Such region is denoted  $S_u$  for the sake of notation. We leave access of apps to IoT objects of different regions for future works.

**Network Architecture.** The fog system can be described by a weighted graph  $G = (V, E)$  where  $V = \{S_i \cup U_i\}_{i \in \mathcal{K}}$  and  $E \subseteq \binom{V}{2}$ . The weight of each edge  $(i, j) \in E$  consists of the delay,  $d_{ij}$ , of the link and the bandwidth of the link  $B_{ij}$ . Let  $\mathcal{N}(S_i) = \{S_j | (j, i) \in E\}$ .

**Application Architecture.** As depicted in Fig. 2, an application  $u \in \mathcal{U}$  consists of two containers:  $u_A$  and  $u_B$ . In order to account for computing and communication constraints in a practical case, we refer to a benchmark application for face recognition in a video stream. As introduced before, modules for processing IoT data streams – face detection processing over the sequence of video frames in our example – are containerized in  $u_B$ . They can be deployed in the central cloud  $S_0$  or on the edge, i.e., in regions  $S_i, i = 1, 2, 3$ . Conversely,  $u_A$  contains all remaining logic, including, e.g., alarm generation in case a positive match is returned. The application has to output every  $1/F_u$  seconds a result  $y_u$  – in this case a positive or negative face recognition match.  $u_A$  is installed in the central cloud  $S_0$ . We can hence consider the whole processing chain involved by the two-containers and the related data transmission delay. We should also include the processing delay  $d_u$  of application  $u$  (if deployed back to back to the IoT object), plus the communication delay  $d_{uj}$ , which is the additional delay to retrieve data from region where the sensor belongs to  $u_A$ , when  $u_B$  is installed in region  $j$ .

The IoT source – in the example a videocamera – generates information units – video frames – of size  $\Delta_u$ , which are served at rate  $B_u$  bit/s. We denote  $\Delta_u^H = \Delta_u$ . Conversely,  $u_B$  transfers smaller information unit  $\Delta_u^L$  to  $u_A$ .

Finally, we denote  $c_u^M, c_u^S, c_u^P$  the resource requirements of application  $u$ , in terms of memory, storage and processing capacity, respectively, of  $u_B$ ; with compact notation we denote  $\mathbf{c}_u = (c_u^M, c_u^S, c_u^P)$ .

In the placement problem we need to consider the processing and transferring time. Actually, the processing time for each information unit depends on the throughput be-

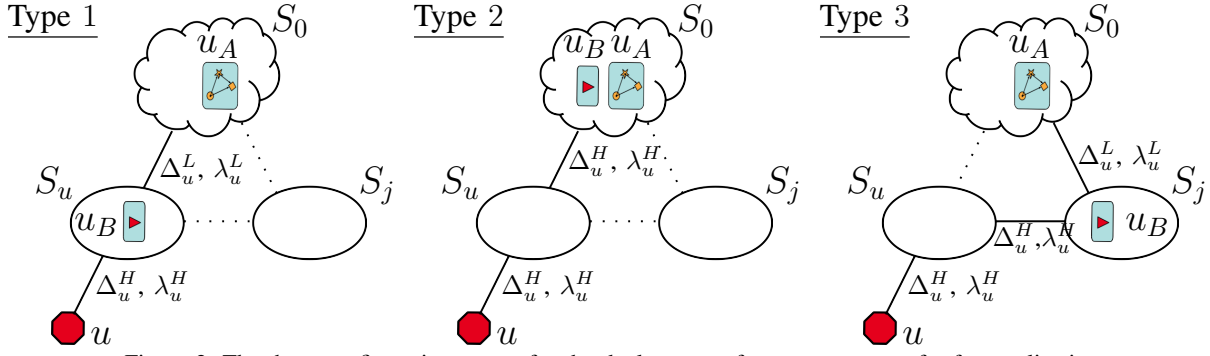


Figure 2: The three configurations types for the deployment of component  $u_B$  of a fog application.

tween application modules. Any application placement has to guarantee that the application to process an information unit  $\Delta_u$  in  $\frac{1}{F_u}$  seconds. Thus, the allocation of such throughput depends on the app deployment configurations. Since  $u_A$  is always installed on the central cloud, the three basic fog configurations to deploy app  $u$  are as in Fig. 2:

*Type 1:*  $u_B$  deployed on  $S_u$ ; higher throughput  $\lambda_u^H$  flows between IoT object  $u$  and region  $S_u$ , with IoT data unit  $\Delta_u = \Delta_u^H$ .  $\Delta_u^L$  is served between  $S_u$  and  $S_0$  with low throughput  $\lambda_u^L$ ;

*Type 2:*  $u_B$  deployed on central cloud  $S_0$ ; the IoT data  $\Delta_u = \Delta_u^H$  is served between  $S_u$  and  $S_0$  with high throughput  $\lambda_u^H$ ;

*Type 3:*  $u_B$  deployed on a neighboring fog region  $S_j \neq S_u$ ; lower throughput required between  $S_j$  and central cloud  $S_0$ . However, the IoT data  $\Delta_u = \Delta_u^H$  is served between  $S_u$  and  $S_0$  with high throughput  $\lambda_u^H$ .

### III. PROBLEM FORMULATION

The resource allocation problem is tackled from the perspective of the edge-infrastructure owner. Her aim is to maximize the revenue obtained in the provision of her fog infrastructure to application tenants. In fact, she settles a cost in order to deploy an application using the traditional scheme of pay per use. A tenant owning application  $u$  pays  $f_{u,k} > 0$  euros per container installed in region  $k$ .

The objective is to schedule the containerized fog applications such in a way to maximize the owner revenue, while satisfying the applications' requirements. We can obtain the optimal reward for a given set of application requests. Hence, the following formulation provides an upper bound on the average reward that can be attained with perfect information.

Decision variables  $x_{u,k,i}$  are boolean variables indicating the placement of the application  $u$  on the  $i$ -th server of the region  $k$ . Further, decision variables  $\lambda_u^H, \lambda_u^L \in \mathbb{R}^+$  represent throughput in the large and small data unit transfer mode of application  $u$ , respectively. The optimal allocation policy using a mixed integer non linear program (MINLP) writes:

$$\text{maximize: } \sum_{(u,k) \in \mathcal{U} \times \mathcal{K} \setminus \{u\}} f_{u,k} x_{u,k} \quad (1)$$

subject to:

$$\sum_{u \in \mathcal{U}} \mathbf{c}_u x_{u,k,i} \leq \mathbf{C}_{k,i}, \quad \forall k \in \mathcal{K}, \forall i \in S_k \quad (2)$$

$$\sum_{u \in U_k} (x_{u,k} \lambda_u^L + x_{u,0} \lambda_u^H) + \sum_{j \in \mathcal{N}(S_k)} \sum_{v \in U_j} x_{v,j} \lambda_v^L \leq B_{k,0}, \quad \forall k \in \mathcal{K} \setminus \{0\} \quad (3)$$

$$\sum_{u \in U_k} x_{u,j} \lambda_u^H + \sum_{u \in U_j} x_{u,k} \lambda_u^H \leq B_{k,j}, \quad \forall jk \in E, j, k \neq 0 \quad (4)$$

$$d_u + \frac{\Delta_u^H}{B_u} + \left( d_{uj} + \frac{\Delta_u^H}{\lambda_u^H} + \frac{\Delta_u^L}{\lambda_u^L} \right) x_{u,j} + \left( d_{u0} + \frac{\Delta_u^H}{\lambda_u^H} \right) x_{u,0} + \left( d_{u0} + \frac{\Delta_u^L}{\lambda_u^L} \right) x_{u,u} \leq \frac{1}{F_u} \quad (5)$$

$$\forall u \in \mathcal{U}, \forall j \in \mathcal{N}(S_u) \quad (5)$$

$$\sum_{k \in \mathcal{K}} x_{u,k} \leq 1 \quad \forall u \in \mathcal{U} \quad (6)$$

$$\sum_{k \in \mathcal{K} \setminus \{\mathcal{N}(u) \cup \{u\}\}} x_{u,k} \leq 0 \quad \forall u \in \mathcal{U} \quad (7)$$

$$x_{u,k,i} \in \{0, 1\} \quad \forall (u, k) \in \mathcal{U} \times \mathcal{K} \quad \forall i \in S_k \quad (8)$$

$$\lambda_u^H, \lambda_u^L \in \mathbb{R}^+ \quad (9)$$

where we let  $x_{u,k} = \sum_{i \in S_k} x_{u,k,i} \quad \forall (u, k) \in \mathcal{U} \times \mathcal{K}$  for notation's sake. The objective function is the revenue gained by the infrastructure owner. The constraint (2) is meant component-wise: it bounds the resources utilization on fog servers in terms of memory, processing and storage capacity, respectively. Also, (3) and (4) bound the throughput generated by applications with respect to links' capacity. (3) accounts for all traffic from region  $k$  to the central cloud, whereas (4) accounts for the throughput across adjacent regions as in 2c. By constraint (5), the total transmission and computing time needs to be smaller than the service rate of the application. We assume that, according to (6), each application has at most one deployment region. In particular, (7) indicates that each application can be deployed only on neighbor regions or on its original region.

The decision variables are the binary variables for the

placement and the continuous variables for the throughput. The Prob. 1–9 is a combination of a placement problem and a multicommodity flow problem. For the sake of tractability, in the next section we offer a reduction to a pure placement problem, which is seen to correspond to a  $m$ -dimensional knapsack problem.

#### IV. PURE PLACEMENT PROBLEM

The reduction is attained by fixing the continuous decision variables of the MINLP, i.e.,  $\lambda_u^L$  and  $\lambda_u^H$ . To do so, we fix the minimum throughput required for each application  $u \in \mathcal{U}$  to deliver the output at target rate  $F_u$ , given the configuration type and the deployment region for  $u_B$ .

*Type. 1:* processing each information unit and providing an output result should happen at rate  $\frac{1}{F_u}$ ; by accounting for all processing and communication delay we write

$$d_u + d_{u0} + \frac{\Delta_u^H}{B_u} + \frac{\Delta_u^L}{\lambda_u^L} \leq \frac{1}{F_u} \quad (10)$$

which can be solved for equality in  $\lambda_u^L$ ;

*Type. 2:* For each application  $u$ , we have

$$d_u + d_{u0} + \frac{\Delta_u^H}{B_u} + \frac{\Delta_u^H}{\lambda_u^H} \leq \frac{1}{F_u} \quad (11)$$

In this case we are solving for  $\lambda_u^H$ ; we observe that it must hold indeed  $\lambda_u^H \geq \lambda_u^L$ .

*Type. 3:* if  $u_B$  is deployed in a region neighbor of the original region of  $u$ , it holds

$$d_u + d_{uj} + d_{j0} + \frac{\Delta_u^H}{B_u} + \frac{\Delta_u^H}{\lambda_u^H} + \frac{\Delta_u^L}{\lambda_u^L} \leq \frac{1}{F_u} \quad (12)$$

In this case, in order to have a unique solution in the minimum throughout, we impose additional constraints, namely we restrict to the set of solutions such that

$$\frac{\lambda_u^H}{\lambda_u^L} = \frac{\Delta_u^H}{\Delta_u^L} \quad (13)$$

Once we performed the above identification, the original problem becomes:

$$\text{maximize: } \sum_{(u,k) \in \mathcal{U} \times \mathcal{K}} f_{u,k} x_{u,k} \quad (14)$$

subject to:

$$\sum_{u \in \mathcal{U}} \mathbf{c}_u x_{u,k,i} \leq \mathbf{C}_{k,i}, \quad \forall k \in \mathcal{K}, \forall i \in S_k \quad (15)$$

$$\begin{aligned} & \sum_{u \in U_k} (x_{u,k} \lambda_u^L + x_{u,0} \lambda_u^H) + \\ & + \sum_{j \in \mathcal{N}(S_k)} \sum_{v \in U_j} x_{v,j} \lambda_v^L \leq B_{k0}, \quad \forall k \in \mathcal{K} \setminus \{0\} \end{aligned} \quad (16)$$

$$\sum_{u \in U_k} x_{u,j} \lambda_u^H + \sum_{u \in U_j} x_{u,k} \lambda_u^H \leq B_{kj}, \quad \forall jk \in E, j, k \neq 0 \quad (17)$$

$$\sum_{k \in \mathcal{K}} x_{u,k} \leq 1 \quad \forall u \in \mathcal{U} \quad (18)$$

$$\sum_{k \in \mathcal{K} \setminus (\mathcal{N}(u) \cup \{u\})} x_{u,k} \leq 0 \quad \forall u \in \mathcal{U} \quad (19)$$

$$x_{u,k,i} \in \{0, 1\}, \quad \forall (u, k) \in \mathcal{U} \times \mathcal{K}, \quad \forall i \in S_k \quad (20)$$

**Proposition 1.** *Problem (14) is NP-hard.*

*Proof:* For every instance of a multidimensional knapsack with  $n$  decision variables and  $m$  constraints, we can reduce it to an instance of our problem. In fact, it is sufficient to consider an instance of (14)–(20) with  $n$  applications and a single  $m$  servers region, which proves NP-hardness. ■

We note that (14)–(20) appears as a  $m$ -knapsack instance, where  $m = K \sum_{k \in \mathcal{K}} n_k + |E| + 2U$ : in the decision form, the problem is hence NP-complete.

#### A. Placement algorithm

Hereafter, we describe FPA, a greedy solution for (14).

---

#### Algorithm 1: Fog Placement Algorithm (FPA)

---

**Input:**  $G = (V, E), \mathcal{U}$

**Output :** Container placement for each  $u \in \mathcal{U}$

```

1 while  $\mathcal{U} \neq \emptyset$  do
2   for  $i = 1, \dots, K$  do
3     for  $u \in U_i$  do
4        $\mathcal{A} \leftarrow \emptyset$ ;
5       if  $\text{verify}(S_i, u) = \text{TRUE}$  then
6          $\mathcal{A} \leftarrow \mathcal{A} \cup \{S_i\}$ ;
7       for  $S \in \mathcal{N}(S_i)$  do
8         if  $\text{verify}(S, u) = \text{TRUE}$  then
9            $\mathcal{A} \leftarrow \mathcal{A} \cup \{S\}$ ;
10      if  $|\mathcal{A}| \geq 2$  then
11         $(j^*, s_{j_h}^*) \leftarrow \text{select}(\mathcal{A}, u)$ ;
12        // where  $s_{j_h}^* \in S_{j^*}$ 
13      else if  $|\mathcal{A}| = 1$  then
14         $(j^*, s_{j_h}^*) \leftarrow S_{j^*}$  with  $S_{j^*} \in \mathcal{A}$ ;
15      // select the application to be deployed
16       $u^* \leftarrow \arg \min_{u \in \mathcal{U}} \|\bar{v}_{j^*}^u\|^2$ ;
17      deploy( $u^*, j^*$ );
18      updateServer( $S_{j^*}, s_{j_h}^*, u^*$ );
19      // Update  $G$ 
20      update( $G, S_{j^*}, S_{u^*}, u^*$ );
21       $\mathcal{U} \leftarrow \mathcal{U} \setminus \{u^*\}$ 

```

---

FPA operates an iterative application deployment. At each step, for each region and for each application  $u$  which belongs to that region, it selects the set  $\mathcal{A}$  of admissible regions for the deployment of module  $u_B$  container. Such set includes all the regions satisfying the computational and throughput requirements of a tagged application. Preliminarily, a feasibility check is performed through a *verify* procedure (pseudocode omitted for space's sake): given a region and application's requirement, it verifies whether exists some server in the region to host  $u_B$ . Further, throughput requirements are verified against each configuration type for each application, by ensuring that the residual bandwidth of involved links satisfies the minimum throughput requirement corresponding to the tagged configuration type.

The *select* procedure is reported in Algo. 2: *select* first calculates, for all eligible applications to be still deployed, a gradient  $\bar{v}_S$  for each feasible region. Its components are

calculated at lines 1, 2, 3, 7-8, 11, and 14, respectively, by estimating the normalized decrease of each resource type in case of deployment with tagged configuration. The output is the application minimizing the gradient (line 16). Once

---

**Algorithm 2:** *Select* procedure

---

**Input:**  $\mathcal{A}$ , set of admissible regions for the deployment of the module  $u_B$   
**Output :** A region for the deployment  
 // Build a gradient vector for each region in  $\mathcal{A}$

```

1 for  $S \in \mathcal{A}$  do
2    $v_m \leftarrow \frac{c_u^M}{\text{residual\_mem}(S)}$ ;
3    $v_p \leftarrow \frac{c_u^P}{\text{residual\_proc}(S)}$ ;
4    $v_s \leftarrow \frac{c_u^S}{\text{residual\_stor}(S)}$ ;
5   if  $S \neq S_u$  then
6     if  $S \in \mathcal{N}(S_u)$  then
7       // Case 3
8        $b_1 \leftarrow \frac{\lambda_u^H}{\text{residual\_band}(\{u, S\})}$ ;
9        $b_2 \leftarrow \frac{\lambda_u^L}{\text{residual\_band}(\{S, 0\})}$ ;
10       $\bar{v}_S \leftarrow (v_m, v_p, v_s, b_1, b_2)$ ;
11    else
12      //  $S = S_0$ , case 2
13       $b_1 \leftarrow \frac{\lambda_u^H}{\text{residual\_band}(\{0, u\})}$ ;
14       $\bar{v}_S \leftarrow (v_m, v_p, v_s, b_1, 0)$ ;
15  else
16    // Case 1
17     $b_1 \leftarrow \frac{\lambda_u^L}{\text{residual\_band}(\{0, u\})}$ ;
18     $\bar{v}_S \leftarrow (v_m, v_p, v_s, b_1, 0)$ ;
19 return  $\arg \min_{S \in \mathcal{A}} \{\|\bar{v}_S\|^2\}$ 

```

---

the algorithm has selected the application to be deployed, it updates the computational capacities of the server hosting the module of that application. Afterwards, the algorithm updates the graph structure decreasing the bandwidth of the links that connected the regions selected for the deployment (line 17). It iterates until all applications have been considered.

**Complexity.** Now we look at the complexity of FPA. The procedures *verify*, *updateServer* and *update* have constant time complexity. The procedure *select* computes a vector for each eligible region in the set  $\mathcal{A}$ . In the worst case, the cardinality of  $\mathcal{A}$  is at most  $K - 1$ . Hence, the complexity of the *select* procedure is  $O(K)$ . The cardinality of  $\mathcal{U}$  is  $O(U)$ , and the maximum cardinality of a neighborhood of a certain region is  $O(K)$  in the worst case. Finally, the complexity of FPA is  $O(U^2 \cdot K^3)$ .

## V. REAL IMPLEMENTATION: FOGATLAS

FPA is the fog scheduler of FogAtlas, a fog platform derived from several extensions of the early platform described in [10]. It handles microservice deployment and workload placement by managing a distributed fog infrastructure split into one cloud region and one or more fog regions. Actually, FogAtlas has a region-oriented architecture. In fact, existing OpenSource technologies such as OpenStack and Kubernetes handle well

resources orchestration in traditional data centers where the cloud is centralized (optionally also spread across few large regions). However, they do not handle natively distributed and/or decentralized fog systems, where heterogeneous computing devices lay in diverse IoT regions and must be internetworked with a central cloud, often with bandwidth-limited and/or partially reliable connections. Ultimately, FogAtlas handles the orchestration among regions, while delegating intra-region orchestration to standard OpenStack or Kubernetes controllers.

The platform instantiates fog applications accounting for a set of optional deployment *requirements*. The application owner can specify requirements as constraints imposed to the deployment/execution of microservices in terms of requested resources and/or specific application needs. She is allowed to declare connections of IoT objects with a certain Microservice, see Fig 3. She can also require a specific target region for dispatching.

In this context *Microservice* is a unit of software which plays a specific role as part of a larger fog application. But it can be deployed, upgraded or replaced independently from other microservices of same application. In FogAtlas it is distributed via Docker container images, which are stored in an *Application Repository*, in fact a Docker registry.

FogAtlas adds above OpenStack and Kubernetes an *Orchestrator*, an *Inventory*, a *Monitor* and a set of RESTful API together with some other components needed to operate the whole platform.

**FogAtlas Inventory.** The *Inventory* maintains an annotated topology of the distributed infrastructure and the applications deployed with up-to-date information on the state of resources. The *Inventory* maps infrastructural objects (i.e., regions, nodes, things) and application objects (i.e. applications, microservices) keeping track of their location and deployment status. As far as the infrastructural objects are concerned, the *Inventory* is populated with information from external systems like SDN network orchestrators and/or IaaS managers (e.g., ONOS, OpenStack). On the other hand, application related information is taken from PaaS managers (i.e., Kubernetes). Information is maintained based on a distributed and highly available key value store [11].

**FogAtlas Orchestrator.** The *Orchestrator* (see Fig 3a) receives *Deployment* requests referred to an *Application* and try to place related *Microservice* in a way that best satisfies the imposed requirements. An *Application* is modeled as a graph of *Vertices* (*Devices* or *Things* used by the *Application* and *Microservices*) and *Dataflows*. Both *Vertices* and *Dataflows* can specify requirements in terms of usage of resources and geographical location. We use *Inversion of Control* design principle in order to inject into the *Orchestrator* the specific implementation of the placement algorithm and of the PaaS manager in use (i.e. Kubernetes).

We remark that in FogAtlas we support geographical constraints (regions) and bandwidth constraints which are not standard features of traditional cloud schedulers. The

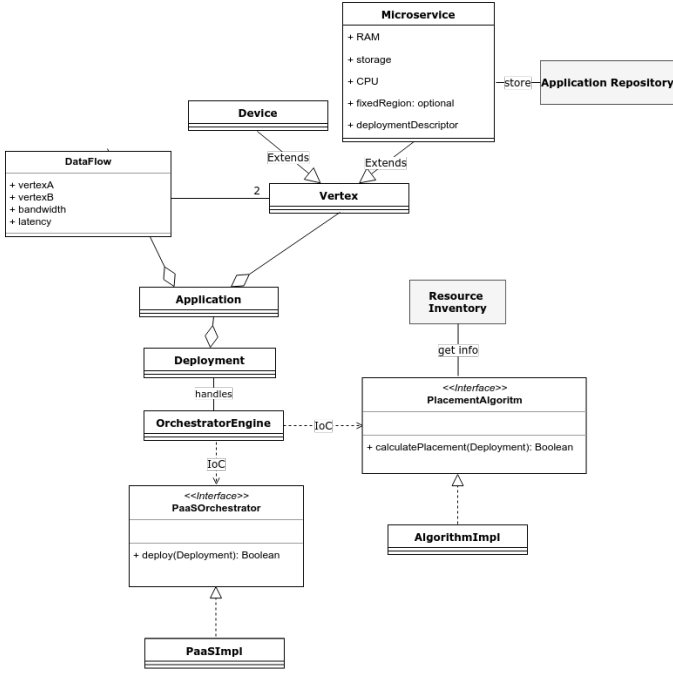


Figure 3: The FogAtlas Orchestrator and its implementation

Application Repository is a Docker registry, typically deployed on the cloud tier, and contains the application images, i.e., Microservice components.

The application deployment is performed as follows. A deployment request is submitted using the FogAtlas RESTful API. Requests can be processed in batches or sequentially (unitary batch). The first step is performed by the Orchestrator: it queries the Inventory, applies filtering and ranking rules as defined by the PlacementAlgorithm to identify the best regions to host the Microservice of the requested Application. Regions satisfying the requirements specified in the deployment request are identified: hence, the Orchestrator operates according to the results of the PlacementAlgorithm. The PaaSOrchestrator finally deploys on the target region the container image of the Microservice. The actual deployment of the Microservice on a node of the selected region is left to the PaaS manager (in this case Kubernetes). At the end of the process, the FogAtlas monitor component updates the Inventory to reflect the global status of infrastructure resources.

#### FogAtlas Implementation

We provide hereafter a few technological details on FogAtlas. In order to combine IaaS availability with flexible management of edge nodes, in FogAtlas the IaaS layer is provided by OpenStack while Kubernetes performs container orchestration. In particular, the OpenStack deployment adheres to the architecture proposed by the Edge Computing Group [12]. Namely, the OpenStack controller lies in the cloud tier while compute nodes cover the edge devices. They are interconnected via "WANWide" links. A Kubernetes cluster

is distributed on top of OpenStack virtual machines, covering both cloud and edge nodes. In case of small edge devices (with respect to available resources) OpenStack is not installed and Kubernetes workers are deployed directly on bare metal.

*Physical testbed and measurements:* the FBK data center holds the cloud tier and the edge cloudlet tier. Server nodes mount an Intel i7 CPU, 16GB RAM, and 480GB SSD. Furthermore, dedicated edge gateways can connect small and low power consumption devices (Raspberry Pi version 3), to perform hardware abstraction layer and to connect for non-IP IoT devices. TP-Link TL-WR740N access points and Tervis JPT3815W-HD cameras are finally connected to our plate recognition application [10].

In order to provide realistic scenarios for our numerical evaluation, we have measured resources demands of such benchmark application (see Tab. II). In the same way, placement constraints due to server characteristics (memory, CPU and storage) do mimic current expected consumer electronics specifications, FogAtlas servers (see Tab. III). The objective is to test the scalability of our fog placement mechanism with the applications batch size, as described in the next section.

## VI. NUMERICAL RESULTS

First, we describe the setup of the tested scenarios. Where not otherwise specified, we intend the infrastructure owner to maximize the number of deployed applications, i.e.,  $f_{u,k} \equiv 1$ .

*Network topology:* we consider a reference undirected network graph with a fixed number of regions  $K = 10$ , where the central cloud and regions form a star topology of cloud-to-fog connections, namely cloud-links. For every topology realization, crosslinks among regions are added according to an Erdős Renyi random graph model, where a link exists between two regions with probability  $q$ . Finally, we assign to each link in the resulting network a bandwidth of 15 Mbps, both for the cloud-links and crosslinks.

*Application Batch Generation:* a batch of fog applications is generated for each experiment; we considered  $U = \{10, 50, 100, 150, 250\}$ . The demands of each application of the batch for CPU, storage, memory and throughput are uniform independent random variables. The mean value of such variables is dictated by the nominal value we measured for our benchmark application. That application, as recalled in the previous section, is a plate-recognition application packaged as a two-modules microservice. The second microservice module can process the video stream either in the cloud or on a fog node. The resulting distribution values for the application batches are enlisted in Tab. II; symbol  $u_0$  refers to the nominal values we measured on FogAtlas for the plate recognition app.

Finally, the probability that an application belongs to region  $k \in \{1, \dots, K\}$  follows a truncated Pareto distribution of parameter  $\alpha$ , i.e.,  $\mathbb{P}\{R_u > k\} = k^{-\alpha}/\gamma$ , where  $R_u$  is the random variable representing the index of the region assigned to the application  $u$  and normalization constant  $\gamma = \sum_{h=1}^K h^{-\alpha}$ .

*Fog Server Classes:* the servers available within each region belong to three classes, depending on the resources they are equipped with, namely *low*, *medium* and *high* class.

Table II: Distribution of the application requirements of CPU, memory, storage and throughput.

Requirement	Mean Value ( $u_0$ )	Range ( $u \in \mathcal{U}$ )
CPU ( $c_u^P$ )	1250 MIPS	[500, 2000] MIPS
Memory ( $c_u^M$ )	1.2 Gbytes	[0.5, 2] Gbytes
Storage ( $c_u^S$ )	3.5 Gbytes	[1, 8] Gbytes
Low throughput ( $\Delta_u^L$ )	1.5 Mbps	[1, 2] Mbps
High throughput ( $\Delta_u^H$ )	4.25 Mbps	[3.5, 5] Mbps

Table III: Characteristics of the three classes of fog servers: low, medium and high.

Type	CPU (MIPS)	Memory (GB)	Storage (GB)
Low	5000	2	60
Medium	15000	8	80
High	44000	16	120

The computational characteristics are listed in Table III. The number of servers per region is determined per realization as follows. Each region is meant to satisfy same fraction of the expected aggregated demand. More precisely, each region is equipped with aggregated resource vector  $(1 + \beta) \frac{U}{K} \mathbf{c}_{u_0}$ . The parameter  $\beta$  is a slack parameter tuning the probability that fog resources are underprovisioned/overprovisioned compared to the aggregated demand. Finally, the servers' population of the tagged region is determined by allocating servers of random type until the region resource budget is exhausted.

#### A. Experimental Results

In Fig. 4a we have depicted the number of deployed applications for increasing batch size. The upper graph reports the results averaged on 10 instances of a scenario with parameter  $\beta = 1.5$  (top) and  $\beta = 2.5$  (bottom), respectively. The red line is the optimal solution obtained by the Gurobi ILP solver [13], the blue line is FPA, whereas the green one is the variant of FPA implemented in FogAtlas, namely FPA-R. It considers region-wise aggregated resources and delegates the intra-region, per-server deployment to Kubernetes schedulers using a randomized placement policy<sup>1</sup>. As seen in Fig. 4a, up to  $U = 50$ , the deployment of the batch of applications is complete. In the last part of the curve, communication constraints dominate, saturating around 100 deployed applications in the optimal case. Increasing from  $\beta = 1.5$  to  $\beta = 2.5$  provides moderate improvement, confined around  $U = 100$ , where the communication constraint is not dominating yet.

Fig. 4b repeats the same experiment in the case of different crosslink density among regions. The figure on top represents the case of denser topologies ( $q = 0.5$ ) and the bottom one the case of sparser ones ( $q = 0.3$ ). We observe first that using  $\beta = 0.3$ , and  $q = 0.5$  (top graph), this scenario has close performance to the ones seen in Fig. 4a, but for much lesser computational resources assigned to fog regions. However, when the network is sparser (bottom graph), the demand peaks for regions of lower indexes – according to the Pareto distribution – are not offloaded to neighboring regions.

<sup>1</sup>Basically, the algorithm runs FPA as if there exists a unique server having aggregated capacity of the entire region.

This causes the bottleneck visible even for smaller batch sizes, i.e.,  $U = 10, 50$ .

From Fig. 4a and b we observe that for the chosen settings, FPA has performance close to the optimal solution, whereas FPA-R pays some performance loss which is traded off for implementation's simplicity.

Fig. 4c reports on the tests performed on the orchestration delay on the FogAtlas platform, defined as the time needed from the instant when the batch of application is offered to the scheduler until the placement is calculated. As we can see, the expected time complexity is moderately super-linear, confirming scalability to larger batch sizes.

We tested again the sparser deployment ( $q = 0.3$ ) already described in Fig. 4b, for  $U = 100$ . In Fig. 4d, we have generated a typical instance and described the configurations of the deployments produced by FPA and by the optimal solution. The latter prefers type 3 configurations over type 1 configurations, whereas the opposite occurs for FPA. The impact onto the link utilization is different: we tested the link utilization in Fig. 4e and f. Actually, crosslinks are fully utilized in both cases, see Fig. 4e. But, offloading using Type 3 configurations is less frequent with the greedy algorithm: in turn cloud-links are underutilized (Fig. 4f). The different behaviour is due to the fact that, in a throughput-dominated scenario, optimal solutions prioritize communication constraints more efficiently than FPA's ones.

Finally, Fig. 4g and h characterize deployed applications for different weights. We depicted there the Cumulative Distribution Function (CDF) for the memory, storage and CPU required by the selected applications. The distribution is uniform in the case of equal weights, indicating that both optimal and FPA solutions sample applications to deploy uniformly at random with respect to computing requirements. This is what desired in a throughput-dominated scenario, proving that FPA behaves correctly by prioritizing the communication constraints. In the second scenario, half applications are generated with the maximum CPU value and the others uniform. We have assigned to each application  $u$  the weight  $\frac{c_u^P}{\max_{\mathcal{U}} c_u^P}$ , i.e., according to their probability mass distribution. Doing so, both the optimal and the FPA solutions have deployed applications according to the weight distribution, prioritizing higher CPU consumption.

## VII. RELATED WORK

Efficient service deployment is a core topic in cloud computing [14], [15]. In fog computing, the presence of remote, heterogeneous devices on edge nodes motivated novel schemes to match QoS requirements and maximize network usage. Authors of [16] focus on the provision of QoS constrained, eligible deployments for applications. The problem is showed NP-hard with a reduction from the subgraph isomorphism problem. Preprocessing plus backtracking determines the final eligible deployment restricting the search space. But, no performance target is optimized.

In [6], application provisioning is studied from the perspective of the network infrastructure. A fully polynomial-



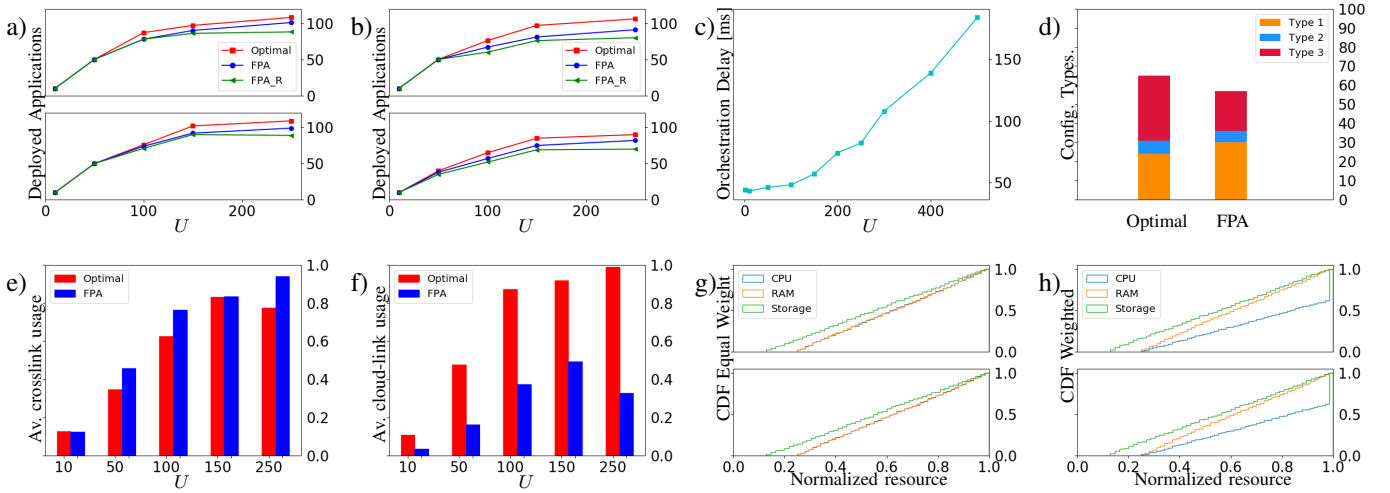


Figure 4: a/b) Number of deployed applications: a)  $q = 0.4$   $\beta = 1.5$  (top) and  $\beta = 2.5$  and (bottom); b)  $\beta = 0.3$  and  $q = 0.5$  (top) and  $q = 0.3$  (bottom); c) Orchestration delay,  $\beta = 0.3$  and  $q = 0.5$  (top) and  $q = 0.3$ ; d) Configuration types distribution for a typical solution instance with  $U = 100$ ,  $q = 0.3$  and  $\beta = 0.3$ ; e/f) Average link usage (settings as in d); e) cloud-links and f) crosslinks; g/h) CDF of the demands for the deployed applications g) Equal weight, optimal and FPA solutions,  $q = 0.5$  and  $\beta = 1.5$  and h) Weighted, optimal and FPA solutions,  $q = 0.5$  and  $\beta = 0.5$ ;

time approximation scheme is derived for single and multiple application deployment, showing large QoS performance improvement with respect to applications' bandwidth and delay figures; computational requirements are not accounted for.

Taneja et al. [7] define a placement algorithm by mapping the directed acyclic graph of the modules of an IoT-based application into fog and cloud nodes. Numerical results show performance gains in terms of latency, energy and bandwidth constraints compared to edge-agnostic placement schemes. In our work, conversely, we provide an optimization framework to account for the coupling of traffic and computing demands of a batch of applications to be deployed over multiple regions.

## VIII. CONCLUSIONS

In this paper, we have introduced an optimization framework for microservice scheduling over fog infrastructures, where different configurations are used to orchestrate fog computation modules to the edge or in cloud. The problem combines a multi-commodity flow and a placement problem, but can be reduced to a  $m$ -dimensional knapsack problem by introducing throughput proportionality. We proposed a greedy algorithm, namely FPA, which performs efficiently with respect to the optimal solution by performing placement using a gradient approach. We have tested numerically our framework under realistic dimensioning, leveraging our platform FogAtlas. Extensive numerical experiments have confirmed the scalability properties of the proposed fog orchestration technique.

## REFERENCES

- [1] D. Miorandi, S. Sicari, F. De Pellegrini, and I. Chlamtac, "Internet of things: Vision, applications and research challenges," *Ad Hoc Networks*, vol. 10, no. 7, pp. 1497 – 1516, 2012. [Online]. Available: <http://www.sciencedirect.com/science/article/pii/S1570870512000674>
- [2] 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*. ACM, 2012, pp. 13–16.
- [3] M. Chiang and T. Zhang, "Fog and IoT: An overview of research opportunities," *IEEE Internet of Things Journal*, vol. 3, no. 6, pp. 854–864, Dec 2016.
- [4] G. Li, J. Wu, J. Li, K. Wang, and T. Ye, "Service popularity-based smart resources partitioning for fog computing-enabled industrial Internet of Things," *IEEE Transactions on Industrial Informatics*, pp. 1–1, 2018.
- [5] Y. Guan, J. Shao, G. Wei, and M. Xie, "Data security and privacy in fog computing," *IEEE Network*, pp. 1–6, 2018.
- [6] R. Yu, G. Xue, and X. Zhang, "Application provisioning in Fog Computing-enabled Internet-of-Things: a network perspective," in *Proc. of INFOCOM*, 2018.
- [7] M. Taneja and A. Davy, "Resource aware placement of iot application modules in fog-cloud computing paradigm," in *Proc. of IFIP/IEEE IM*, 2017, pp. 1222–1228.
- [8] Y. Gan and C. Delimitrou, "The architectural implications of cloud microservices," *IEEE Computer Architecture Letters*, vol. 17, no. 2, pp. 155–158, Dec 2018.
- [9] L. Canzian and M. V. D. Schaar, "Real-time stream mining: online knowledge extraction using classifier networks," *IEEE Network*, vol. 29, no. 5, pp. 10–16, Sept. 2015.
- [10] D. Santoro, D. Zozin, D. Pizzolli, F. De Pellegrini, and S. Cretti, "Foggy: A platform for workload orchestration in a fog computing environment," in *Proc. of IEEE CloudCom*, Dec 2017, pp. 231–234.
- [11] "Etcd." [Online]. Available: <https://coreos.com/etcd>
- [12] "Openstack Edge Computing Group," Available Online, [https://wiki.openstack.org/wiki/Edge\\_Computing\\_Group](https://wiki.openstack.org/wiki/Edge_Computing_Group).
- [13] Gurobi Optimization, LLC, "Gurobi optimizer reference manual," 2018. [Online]. Available: <http://www.gurobi.com>
- [14] R. Cohen, L. Lewin-Eytan, J. S. Naor, and D. Raz, "Almost optimal virtual machine placement for traffic intense data centers," in *Proc. of IEEE INFOCOM*, 2013, pp. 355–359.
- [15] J. W. Jiang, T. Lan, S. Ha, M. Chen, and M. Chiang, "Joint VM placement and routing for data center traffic engineering," in *Proc. of IEEE INFOCOM*, vol. 12, 2012, pp. 2876–2880.
- [16] A. Brogi, S. Forti, and A. Ibrahim, "How to best deploy your Fog applications, probably," in *Proc. of IEEE ICPEC*, 2017, pp. 105–114.