Resource Allocation with Service Availability and QoS Constraints in Mobile Fog Networks

Nader Daneshfar, Nikolaos Pappas, and Vangelis Angelakis
Department of Science and Technology, Linkoping University, Campus Norrkoping, 60 174, Sweden
Email: {nader.daneshfar, nikolaos.pappas, vangelis.angelakis}@liu.se

***Abstract*—The evolution of the Internet of Things (IoT) is bringing Cloud services closer to the networks' edge. Thus, fog networking presents itself as an approach aiming to utilize more and more resources in network edge devices to provide various networking tasks. This work presents an optimization formulation that minimizes the cost of executing a set of services, taking into account the availability of resources in mobile edge devices.**

I. Introduction

Driven primarily by requirements of the Internet of Things (IoT), Fog networking has been introduced as the offloading method of services to the edge of the network [1]. Fog can enhance distributed computing, management, control, storage and networking by providing such services at the edge of the network [2]. Comparing Fog against the Cloud we find key features that differ and can be categorized into three main parts as *Storage*, *Computation* and *Network Communication and Management*. A critical aspect of any operating system is highly related to data handling and processing. In this respect, application either has its own capability for storing data or utilizes a remote resource upon request. Fog can introduce storage and caching at the edge of network to further localize the file storage management. Fog Radio Access Network (F-RAN) architecture has been introduced to bridge the gap between existing technologies and combining the benefits of both edge and cloud processing.

Fog is still a relatively young term, with different definitions, architectures and scopes. Even the term Fog networking is often interchangeable with fog computing. Fog computing in [3] is defined as a very large number of interconnected heterogeneous and decentralized devices that have the ability to communicate and cooperate with each other and the existing network to facilitate performing tasks without the intervention of third parties. This definition may require a bit of tuning to address other type of usages for Fog, but it still conveys a comprehensive understanding. There are similar concepts to Fog computing such as Mobile Edge Computing (MEC) and Mobile Cloud Computing (MCC). The former is focused on bringing more computational power to the edge of the network where users demand higher level of computational power. The latter describes a method where both computational task and storage occur at a remote place with respect to the mobile node.

This project has received funding from the European Union's Horizon 2020 research and innovation programme under the Marie Sklodowska-Curie grant agreement No "642743" (WiVi-2020).

II. Problem Formulation and Mathematical Model

We consider that we are given a set $\mathcal{U}$ of users that demand services from a set $\mathcal{S}$ of servers. The amount of demand of user $u \in U$ is denoted by d_u. The demand is *unsplittable*, i.e., partial serving of a demand is prohibited. However, because we consider that the fog servers may not be available to execute the demand - especially if they are mobile, we allow multicasting of duplicates of the service to multiple servers. Duplicate flooding (i.e., sending of the demand to too many servers) is avoided by introducing a bound M on the number of servers in which a single user can post the service; in addition, sending a unit of the demand to server s has a *cost* w_s for each user and each user $u \in U$ has a limited *budget* B_u within which the total cost incurred by u's service assignments must remain. A server's resources may be shared between several users (i.e., more than one user may request service from a server); however, the total amount of demand served by server $s \in \mathcal{S}$ should not exceed the capacity D_s of the server.

We bring this into the domain of mobile fog computing, where the servers, being even the UEs, may not necessarily be there all the time. A server $s \in \mathcal{S}$ is available with probability $p_s \in (0, 1)$ and we assume that the availabilities of different servers are independent. Thus, successful completion of services cannot be guaranteed to the users. Each user $u \in U$ however has an expressed *minimum service level* requirement $l_u \in (0, 1)$: the user wants to have its demand served with probability greater than l_u.

We dub our problem *M-Fog Allocation* or *MFA* for short. Our goal is to decide, for each user, to which set of servers the user should multicast their demand. That is, we consider the problem for a centralized controller (middleware) between the users and the fog, that has *complete and correct* information about all the elements of the system. A natural objective function is the total cost incurred by the users when (multi) casting their demand. For simplicity we assume each service has a cost of w_s.

We formulate MFA as an Integer Program (IP) whose decision variables $x_{us}, u \in \mathcal{U}, s \in \mathcal{S}$ indicate whether user u includes server s into the set of servers to which it multicasts its demand:

$$x_{us} = \begin{cases} 1 & \text{if } u \text{ sends to } s \\ 0 & \text{otherwise} \end{cases}$$

Therefore, our mathematical model is formulated as

$$min. \qquad \sum_u \sum_s d_u w_s x_{us}, \tag{1}$$

$$s.t. \quad \sum_{s} x_{us} \leq M, \quad \forall u \in \mathcal{U}, \tag{2}$$

$$\sum_{s} d_u w_s x_{us} \leq B_u, \quad \forall u \in \mathcal{U}, \tag{3}$$

$$\sum_{u} d_u x_{us} \leq D_s, \quad \forall s \in \mathcal{S}, \tag{4}$$

$$\sum_{s} x_{us} \ln(1 - p_s) \leq \ln(1 - l_u), \quad \forall u \in \mathcal{U}. \tag{5}$$

The MFA objective in (1) is to minimize the total cost. By constraint (2) a limit for each user disabling them to excessively send request to servers is introduced. Each user is coupled with a respective cost budget that is constrained by (3). By (4) we control duplicate flooding limiting the total amount of service requests to each server in the Fog. The equations in (5) offers users their QoS by considering the probability of failure in service acquisition by any of the assigned servers.

III. Numerical Results

In this section we present the results for sets of simulations that has been done using Matlab. Here we aim to asses the optimal total cost of serving all service requests of the system. This is conducted with different configurations of cost per unit of service request as well as the probability of availability for servers and the minimum QoS requirement by users.

The former test condition is aimed to study the effect of different distribution in the cost variable while the latter is investigating the trade-off between QoS request and resource availability. It is important to note that each user's available budget is ensured to be sufficiently adjusted, ensuring that they can send as many number of requests as required to achieve their desired quality of service. Additionally, all scenarios are performed with fixed number of servers having constant accumulated available resources.

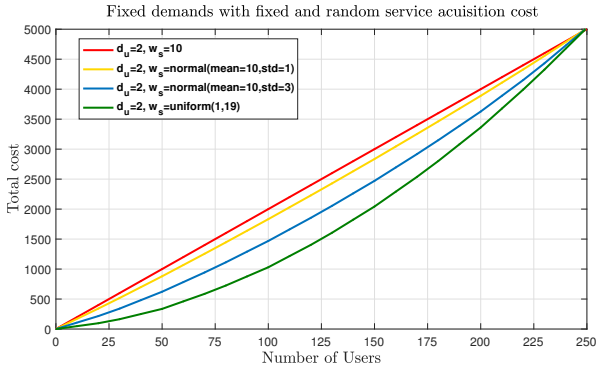

Fig. 1. Total cost vs. Number of users in various server distribution.

The total cost is at its highest optimal value when the cost per unit of service (w_s) is fixed at a constant value and it increases linearly. It is caused by the fact that all servers provide services with identical cost thus eliminating users to choose which server to send their service request. The optimal cost for a number of users decreases as the servers get more diverse with their available cost per unit of request. This phenomenon is clearly visible in Fig. 1 especially when the system is not saturated. This is because as more diversity is implemented into server set, the users can send their request to the server with the most inexpensive resources.

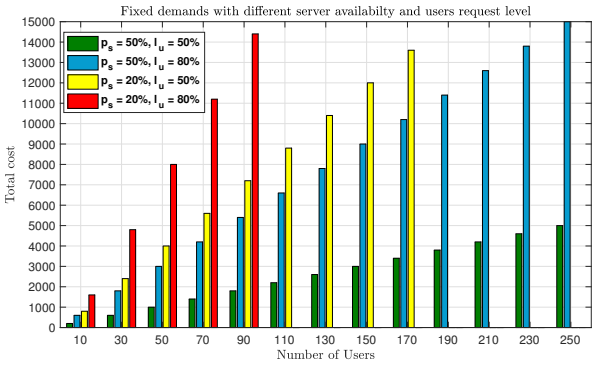

Fig. 2. Total cost vs. Number of users served under different QoS requirements

In addition, as shown in Fig. 2, when w_s is fixed, the total cost of serving all users' demands is in linear relation with the number of replications each user must have. This is to ensure that all users' required QoS are met. The bars reflect the fact that for higher QoS request, with a fixed server availability, more replications of service request are needed. It is remarkable that the multiplication factor for request replications is not constant with QoS increase in different server availabilities. This is due to the logarithmic behavior of the model. Moreover, optimum values for some scenarios (namely red and yellow) are discontinued because of the available resources and the fact that service replications extend beyond server set capacity to provided services.

The numerical results driven from scenarios outline the ability of our model to minimize total cost. They also successfully indicates the effect of various system variables such as probability of availability of each server in respect with the minimum service required by each user and diversity in the server set regarding their service cost. Overall, users can achieve their requested level of QoS by making duplicates of their requests to multiple servers but it comes with a price and also saturates the whole network. On the other hand, our model indicates the direct relation of diversity in cost of services provided by servers and their quantity. As a Fog network tends to have more diverse servers, the further a system can reach its lowest optimized value in relation to its total cost of providing services.

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: ACM, 2012, pp. 13–16.

[2] V. B. C. Souza, W. Ramírez, X. Masip-Bruin, E. Marín-Tordera, G. Ren, and G. Tashakor, "Handling service allocation in combined fog-cloud scenarios," in *2016 IEEE International Conference on Communications (ICC)*, May 2016, pp. 1–5.

[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.