Proactive Virtual Machine Migration in Fog Environments

Diogo Gonçalves⋆, Karima Velasquez†, Marilia Curado†, Luiz Bittencourt⋆ and Edmundo Madeira⋆

⋆Institute of Computing, University of Campinas, Brazil

†CISUC, Department of Informatics Engineering, University of Coimbra, Portugal

diogomg@lrc.ic.unicamp.br, {kcastro, marilia}@dei.uc.pt, {bit, edmundo}@ic.unicamp.br

***Abstract*—Fog computing provides a low latency access to resources at the edge of the network for resource-constrained devices. The high mobility of some of these devices, such as vehicles, brings great challenges related to resource allocation and management. In order to improve the management of computing resources utilized by mobile users connected to the Fog infrastructure, this paper proposes a virtual machine placement and migration decision model based on mobility prediction. Simulations have shown that moving the virtual machine to a Fog node ahead of the user's route using the proposed approach can decrease by almost 50% the number of migrations needed by the user. The Fog architecture provides an average latency of about 15 milliseconds for the users' applications and the proposed approach presents a lower latency compared to a greedy approach for the VM placement problem.**

***Index Terms*—Fog, Virtual Machine, Placement, Migration.**

I. Introduction

With the increasing focus on the Internet of Things (IoT) paradigm, more connected devices produce and consume data at the edge of the network. These resource-constrained devices frequently use Cloud resources to satisfy their computing and storage requirements. Some applications, however, have requirements that are not provided by Cloud environments (e.g.: strict delay) [1].

The Fog Computing paradigm has emerged to provide distributed resources geographically closer to the users at the edge of the network. The cloudlets are dedicated Fog nodes which provide computing and network resources with lower latency than the one experienced at Cloud. cloudlets are usually handled as a virtualized environment which shares its physical resources in the form of Virtual Machines (VM).

In the context of IoT in Smart Cities, user's devices with high mobility, such as vehicles, bikes, and trains, can bring new challenges to the Fog [2] and approaches to orchestrate this environment have been needed [3]. The improvement of systems for support and management of virtual machines, as VM migration, has been pointed as one of the research challenges related to the development of middleware services for the Future Internet for Smart Cities [4], specially because it "can substantially improve a mobility support solution both in terms of performance and applicability" [5].

Keeping the VM as close as possible to its user whilst still maintaining user mobility is a great challenge in this context [5]. Frequently migrating the application from one cloudlet to another may increase the application downtime; on the other hand, not migrating enough times may leave the application too far from its user, increasing the latency. Both cases can compromise the Quality of Service (QoS) for the mobile user.

Many users in the context of Smart Cities has a highly predictable mobility (e.g.: buses and trains). In face of that, some works have been presenting good results incorporating mobility prediction to improve, e.g. content-caching [6], for mobile users. However, current works which propose to solve the problem of VM migration in Fog, targeting especially lower latency and higher availability, does not include any mobility prediction mechanisms.

Given this scenario, this paper proposes a VM migration approach based on mobility prediction. The algorithm defines the set of candidate cloudlets to receive the user's VM according to the user's future position. Furthermore, an Integer Linear Programming (ILP) model is proposed to improves the placement of VMs within the candidate cloudlets set.

This paper is structured as follows. Section II presents the most relevant related work. Section III describes the decision migration model proposed for the proactive virtual machine migration approach. Section IV shows the simulation results and finally, Section V summarizes the work and presents some directions for future research.

II. Related Work

In face to the defined scenario, there is a need to design orchestration mechanisms, including service placement and migration approaches to improve the use of the Fog resources, which are different in nature than the ones in the Cloud [7].

Some works present approaches to solve the service application problem in Fog environments. Skarlat et al. [7] analyze the placement of IoT services in the Fog using an optimization ILP approach that takes into account QoS requirements. iFogSim [8] is used for the evaluation. Velasquez et al. [9] propose a service placement architecture including a simple ILP model for latency reduction in the Fog.

Once the placement is settled, considering mobile user, the conditions might change, rendering in a less-than-optimal location for the service; thus it is important to also consider how to alter the placement by performing a migration of the services. Mostly works which propose VM migration in Fog, as Bittencourt et al. [2], Yao et al. [10], and Ansari [11], does not use data about user's future position in the migration decisions (i.e., when and where).

Some current works with different proposes have been presenting interesting results by incorporating user mobility prediction on their algorithms. Gomes et al. [6] present an enhancing migration of content-caches at edge nodes. Mustafa et al. [12] tackle prediction for migration in vehicular networks with the purpose of improving the performance of the network without incurring in resource overuse.

To the best of our knowledge, most of the works focused on placement and migration in Fog environments are based on theoretical approaches using mathematical models (e.g.: ILP) and simulations, like this work, but does not use any data about users' future position in their approaches. This work is aimed at proposing a mobility prediction based service placement and migration model for mobile users, that could benefit from the Fog, targeted at reducing the latency. The model is described in the following section.

III. VM Placement and Migration decision model

An ILP model was drawn in order to optimize the placement of the VMs in a vehicular network. For this work, two objective functions are executed sequentially in order to achieve two goals: (1) maximize the accepted requests and (2) minimize the latency for the user. Table I lists and describes the variables, both input and decision, included in the model.

TABLE I: Variables for the ILP Model

INPUT VARIABLES	
Variable	Description
N	Set of nodes where the application can be executed
A	Set of applications, where an application is composed by one service
Ω_n	CPU available for node $n \in N$ (MIPS)
ϕ_n	RAM available for node $n \in N$ (MB)
β_n	Bandwidth available (in Mbps) for traffic for node $n \in N$
γ_n	Storage available (in MB) for node $n \in N$
ω_a	CPU requirement for the application $a \in A$ (MIPS)
φ_a	RAM requirement for the application $a \in A$ (MB)
ρ_a	Bandwidth (in Mbps) which an application receives as traffic
α_a	Storage requirement (in MB) for the application
$C_{a,n}$	Cost Matrix. Integer matrix. Latency between all pair a, n of nodes $n \in N$ and the user owner of the application $a \in A$
DECISION VARIABLES	
Variable	Description
$P_{a,n}$	Placement matrix. Binary matrix. 1 when application a is being executed in node n

Equation 1 maximizes the accepted requests. By maximizing the *Placement Matrix* the model aims at accepting the most requests. Equation 2 minimizes the latency between the user and the node where the application will be placed by minimizing the product between the *Placement Matrix* and the *Cost Matrix*, where this later one holds the latency between the user and the node where the application is being placed.

$$max \sum_{a \in A} \sum_{n \in N} P_{a,n} \tag{1}$$

$$min \sum_{a \in A} \sum_{n \in N} P_{a,n} \times C_{a,n} \tag{2}$$

The second objective function (Equation 2) is flexible enough to adapt the model to different objectives by simply varying the content of the *Cost Matrix*; for instance, minimizing the financial costs. Besides the objective functions, several constraints are included in the model.

Equation 3 guarantees that the application is executed on one server at most. Equation 4 enforces CPU constraints according to the resources on each node and the service requirements. The sum of CPU requested by the applications must not surpass the node resources.

$$\forall a \in A : \sum_{n \in N} P_{a,n} \leq 1 \tag{3}$$

$$\forall n \in N : \sum_{a \in A} P_{a,n} \times \omega_a \leq \Omega_n \tag{4}$$

Equation 5 enforces RAM constraints. Analog to Equation 4, the sum of RAM requested by the applications must not exceed the node's resources. Equation 6 limits the placement of applications according to the traffic requested by the applications and the link capacity of the executing node.

$$\forall n \in N : \sum_{a \in A} P_{a,n} \times \varphi_a \leq \phi_n \tag{5}$$

$$\forall n \in N : \sum_{a \in A} P_{a,n} \times \rho_a \leq \beta_n \tag{6}$$

Equation 7 enforces storage constraints. The sum of storage requested by the applications must not surpass the node resources.

$$\forall n \in N : \sum_{a \in A} P_{a,n} \times \alpha_a \leq \gamma_n \tag{7}$$

This model was implemented using the analytical decision support toolkit IBM ILOG CPLEX 12.7.1 [13]. Simulations were executed to evaluate its performance and are presented in the following section.

IV. Simulation and results

This section describes the performance of the proactive migration approach proposed. The simulation scenario was designed by evaluating the resource management in the Fog to meet the QoS requirements from mobile users' applications. For the simulation of the resource management in the Fog, the MyIFogSim simulator [14] was used. MyIFogSim is an extension of IFogSim [8] which supports VM migrations by mobile users. For the simulation of mobility patterns of the users, a vehicular mobility scenario was built using Simulation of Urban MObility (SUMO) simulator [15].

A. *Simulation scenarios*

The simulation scenarios were built using realistic vehicular mobility patterns from Luxembourg SUMO Traffic (LuST) [16]. 2070 different bus traces were used to evaluate the scenario. The average speed of the buses is 22.3 km per hour in a route of, in average, 26.44 minutes. For the vehicular network settings, users connect to the cloudlets by access points uniformly distributed over the map. Each access point has a coverage radius of approximately 500 meters and it is connected to one cloudlet.

Simulations settings about cloudlets resources and VM requirements are consistent with related works [10], [14]. Each cloudlet was built with 2800 million instructions per second, 8 GB of RAM, 80 GB of storage, and one link to the access point with 100 Mbps of bandwidth and 4 milliseconds of latency. A uniformly distributed link from 1 and 10 Gbps connects each pair of cloudlets. The VM's size used was 200 MB.

In order to compare the proposed model, the migration strategy applied to determine the destination cloudlet used in this work was a greedy algorithm which selects the cloudlet with the lowest latency among a set of 10 candidate cloudlets. A live migration based technique was used in the implementation. The approaches were evaluated in a scenario with 30, 60 and 90 simultaneous users. Each approach was simulated using information of the users' future locations from a range up to 5 minutes. This information is used to optimize the set of candidate cloudlets to determine the destination of the VM. As an introductory study and aiming to avoid noise in the hypothesis, the mobility predictor model used has a 100% accuracy.

Average and worst case latencies of packets requested while the VM is not in the migration process, average number of VM migrations and unavailability time were the metrics used to evaluate the scenarios. The results are presented in the following section considering a 95% confidence interval. The plots present the numbers 30, 60 and 90 after the approaches' name to indicate the number of users in these simulations. The x axis is the range of prediction of the user's future location.

B. Results

Figure 1 shows the number of migrations made by each vehicle through its route. The number of migrations tends to decrease as the knowledge of the user's future route is improved. This additional information about the vehicle's route allows both the approaches to select a better set of candidate cloudlets to be used as the destination of the user's application. The knowledge of the following 5 minutes of user's position allows a decrease in the number of migrations by the greedy approach in almost 31% and a decrease of about 50% by the ILP approach.

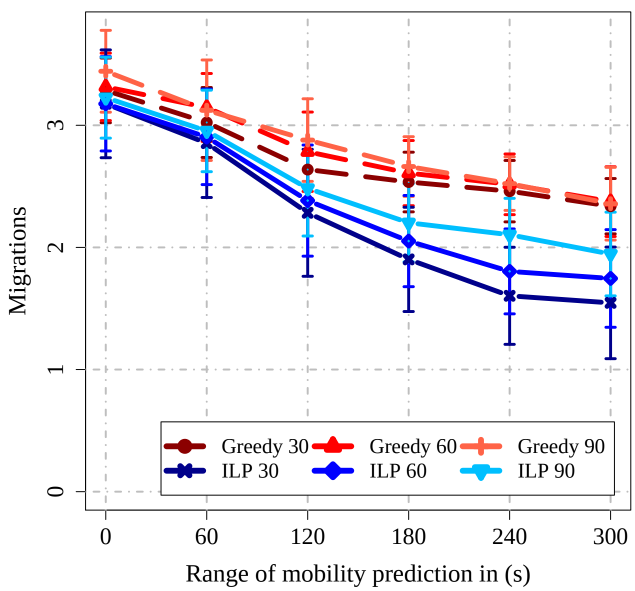

Fig. 1: Average number of migrations

Even in a live migration approach, the user is not able to access its application in the Fog during a period of time of his/her VM migration. A decrease in the number of migrations means a lower time in which the user has its application disconnected. Figure 2 presents the average time of unavailability that the user deals with his/her route.

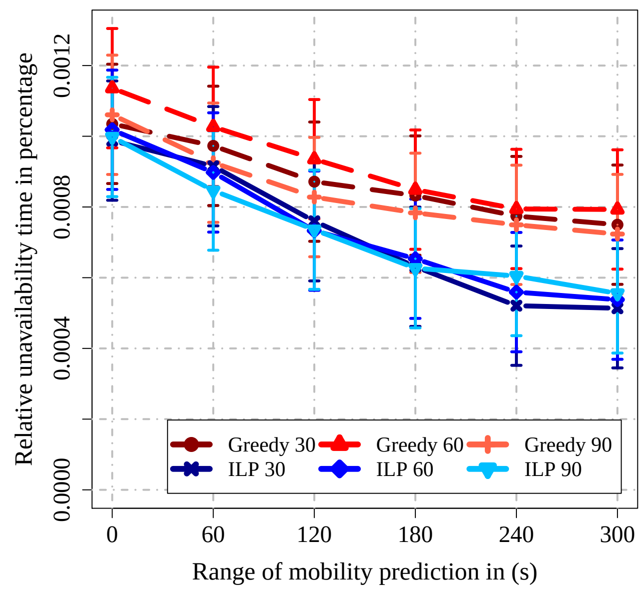

Fig. 2: Relative time of unavailability to access the virtual machine

The decrease in the number of migrations may suggest that the locations in which the VMs are being positioned are more appropriate for the user. However, latency is an important metric that also should be evaluated in applications which use Fog resources. An increase in the distance between the user and his/her application may compromise the levels of latency required by the user.

Figure 3 shows the average latency observed in the simulation scenarios. The Fog architecture has provided a 15 milliseconds latency for the users' applications. The results present, in the entire range of mobility prediction used in the evaluated scenarios, a low and stable level of the average latency. Furthermore, the simulations show that the ILP approach presents an even lower latency compared to the greedy approach in the three evaluated user densities.

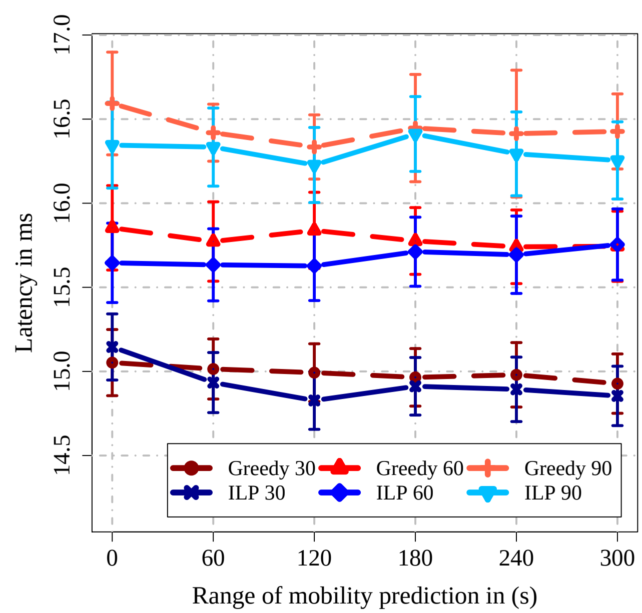

Fig. 3: Average latency

The results about worst case latency presented in Figure 4 indicate that, in general, there is not a significant increase in the application's latency using the mobility prediction approach. Like the results about average latency, also evaluating the worst case latency, the scenarios which use the ILP model present better results if compared to the greedy algorithm.

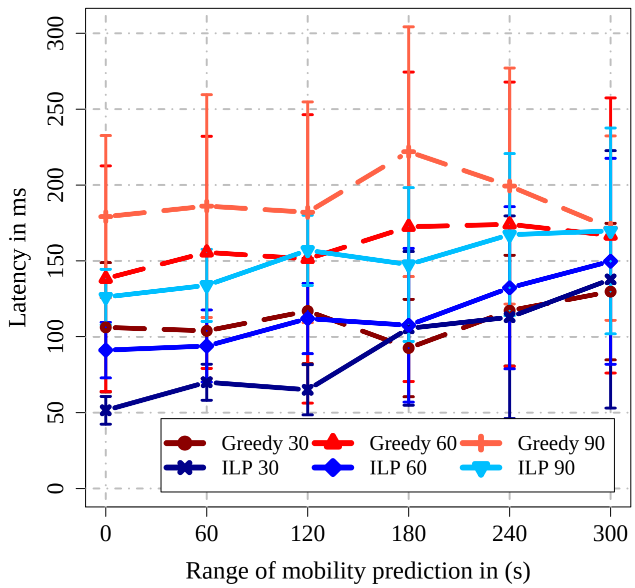

Fig. 4: Average latency in worst case

V. Conclusion

This work proposes a proactive VM migration approach to determine the set of candidate cloudlets to receive the user's VM. The approach uses users' future position to improve the set of candidate cloudlets. In order to optimize the placement of the VMs, an ILP model was proposed. Although the solution will not necessarily match with the optimal that would be achieved by using a single objective function, it is sufficient since it gets improved results with less computational costs.

Simulations suggest that the presented policy reduce the total of migration along the user's path without affecting the latency of VMs allocated to the Fog. This decrease in the number of migrations results in a lower time of unavailability to access the VM in all the evaluated scenarios. The Fog architecture has shown a low average latency of about 15 milliseconds for the users' applications. In all the scenarios which use the mobility prediction approach, the average latency of the application was kept at a low level, and the worst case latency did not show a significant increase. The ILP model proposed has shown a lower latency in average and worst case scenarios compared to a greedy algorithm to select the candidate cloudlets. The ILP model also presents a lower number of migrations.

As future works, we intend to improve the model incorporating historical data about previous migrations. Furthermore, an evaluation of the approaches using more realistic vehicular mobility predictors will be carried out.

Acknowledgment

This work is part of the INCT of the Future Internet for Smart Cities (CNPq 465446/2014-0, CAPES 88887.136422/2017-00 and FAPESP 2014/50937-1) and CNPq grant 420907/2016-5.

The work presented in this paper was partially carried out in the scope of the projects SORTS, financed by the CAPES - Coordenção de Aperfeiçoamento de Pessoal de Nível Superior ≪CAPES-FCT/8572/14-3≫ and by the FCT - Foundation for Science and Technology ≪FCT/13263/4/8/2015/S≫; and DenseNet financed by national funds via FCT - Foundation for Science and Technology within the scope of project PTDC/EEI-SCR/6453/2014 - DenseNet.

Karima Velasquez wishes to acknowledge the Portuguese funding institution FCT - Foundation for Science and Technology for supporting her research under the Ph.D. grant SFRH/BD/119392/2016. Diogo Gonçalves whishes to acknowledge the CAPES for the financial support.

References

[1] A. V. Dastjerdi and R. Buyya, "Fog computing: Helping the internet of things realize its potential," *Computer*, vol. 49, no. 8, pp. 112–116, Aug 2016.

[2] L. F. Bittencourt, M. M. Lopes, I. Petri, and O. F. Rana, "Towards virtual machine migration in fog computing," in *2015 10th International Conference on P2P, Parallel, Grid, Cloud and Internet Computing (3PGCIC)*, Nov 2015, pp. 1–8.

[3] K. Velasquez, D. Perez Abreu, D. Goncalves, L. Bittencourt, M. Curado, E. Monteiro, and E. Madeira, "Service orchestration in fog environments," in *2017 IEEE 5th International Conference on Future Internet of Things and Cloud (FiCloud)*. IEEE, 2017, pp. 329–336.

[4] D. M. Batista, A. Goldman, R. Hirata, F. Kon, F. M. Costa, and M. Endler, "Interscity: Addressing future internet research challenges for smart cities," in *Network of the Future (NOF), 2016 7th International Conference on the*. IEEE, 2016, pp. 1–6.

[5] C. Puliafito, E. Mingozzi, and G. Anastasi, "Fog computing for the internet of mobile things: issues and challenges," in *Smart Computing (SMARTCOMP), 2017 IEEE International Conference on*. IEEE, 2017, pp. 1–6.

[6] A. S. Gomes, B. Sousa, D. Palma, V. Fonseca, Z. Zhao, E. Monteiro, T. Braun, P. Simoes, and L. Cordeiro, "Edge caching with mobility prediction in virtualized lte mobile networks," *Future Generation Computer Systems*, vol. 70, pp. 148–162, 2017.

[7] O. Skarlat, M. Nardelli, S. Schulte, and S. Dustdar, "Towards qos-aware fog service placement," in *2017 IEEE 1st International Conference on Fog and Edge Computing (ICFEC)*, May 2017, pp. 89–96.

[8] H. Gupta, A. Vahid Dastjerdi, S. K. Ghosh, and R. Buyya, "ifogsim: A toolkit for modeling and simulation of resource management techniques in the internet of things, edge and fog computing environments," *Software: Practice and Experience*, vol. 47, no. 9, pp. 1275–1296, 2017.

[9] K. Velasquez, D. Perez Abreu, M. Curado, and E. Monteiro, "Service placement for latency reduction in the internet of things," *Annals of Telecommunications*, vol. 72, no. 1, pp. 105–115, Feb 2017.

[10] H. Yao, C. Bai, D. Zeng, Q. Liang, and Y. Fan, "Migrate or not? exploring virtual machine migration in roadside cloudlet-based vehicular cloud," *Concurrency and Computation: Practice and Experience*, vol. 27, no. 18, pp. 5780–5792, 2015.

[11] X. Sun and N. Ansari, "Edgeiot: Mobile edge computing for the internet of things," *IEEE Communications Magazine*, vol. 54, no. 12, pp. 22–29, December 2016.

[12] A. M. Mustafa, O. M. Abubakr, O. Ahmadien, A. Ahmedin, and B. Mokhtar, "Mobility prediction for efficient resources management in vehicular cloud computing," in *Mobile Cloud Computing, Services, and Engineering, 5th IEEE International Conference on*. IEEE, 2017, pp. 53–59.

[13] I. I. CPLEX, "12.7, users manual for cplex, 2016."

[14] M. M. Lopes, W. A. Higashino, M. A. Capretz, and L. F. Bittencourt, "Myifogsim: A simulator for virtual machine migration in fog computing," in *Companion Proceedings of the10th International Conference on Utility and Cloud Computing*, ser. UCC '17 Companion. New York, NY, USA: ACM, 2017, pp. 47–52.

[15] M. Behrisch, L. Bieker, J. Erdmann, and D. Krajzewicz, "Sumo – simulation of urban mobility: An overview," in *SIMUL 2011*, S. . U. of Oslo Aida Omerovic, R. I. R. T. P. D. A. Simoni, and R. I. R. T. P. G. Bobashev, Eds. ThinkMind, October 2011.

[16] L. Codeca, R. Frank, and T. Engel, "Luxembourg sumo traffic (lust) scenario: 24 hours of mobility for vehicular networking research," in *Vehicular Networking Conference (VNC), 2015 IEEE*. IEEE, 2015, pp. 1–8.