Large-Scale IoT Network Offloading to Cloud and Fog Computing: a Fluid Limit Model

Gonzalo Belcredi*, Laura Aspirot†, Pablo Monzón*, Pablo Belzarena*
* Facultad de Ingeniería, Universidad de la República
{gbelcredi, monzon, belza}@fing.edu.uy
† Facultad de Ciencias Económicas y de Administración, Universidad de la República
laspirot@ccee.edu.uy

***Abstract*—This paper models a large-scale Internet of Things (IoT) network as a stochastic system that offloads computing towards Fog and Cloud via a shared access medium. The analysis of this large IoT system by stochastic methods is a challenging problem, if possible, to solve. This paper proposes the approximation of the dynamic of the IoT network via the fluid limit of the stochastic process. This method allows the analysis of the large-scale system and also allows finding the equilibrium point of the system. The results obtained with stochastic simulations show that the fluid model is an excellent approximation of the stochastic system.**

***Index Terms*—Edge Computing, Fog Computing, Cloud Computing, Computation Offloading, Markov Process, Fluid Limit, Switched Systems, Wireless Communications, Internet of Things**

I. Introduction

In recent years, the number of IoT devices has grown significantly, and the trend is expected to continue and accelerate with the arrival of 5G networks [1]. New applications and services will also require an increase in processing capabilities while lowering latency: Augmented Reality, Internet of Vehicles, Smart Home, Health Monitoring Devices. For this purpose, edge and fog computing emerge as key technological enablers.

While cloud computing is supposed to have a significant availability of resources, edge and fog processing could help in reducing latency and communication costs. However, these last devices are limited in their capabilities (energy constraints, buffer size, CPU, etc.). Therefore, one important problem is determining an optimal offloading factor at the different stages of the three-tier computing architecture (Local-Fog-Cloud) to maximize the system's overall performance.

The literature on this area has presented various stochastic offloading models, the articles [2], [3] are detailed surveys on this topic. Authors in [4]–[6] propose Markov Process models for offloading computation.

However, large-scale network offloading modeling remains an open challenge because as the network scales, typical models increases the decision delay and therefore the offloading delay. Although there are several works with scalable methods such as those based on game theory and machine learning, the challenge still exists [7].

In this context, in this article we are interested in knowing if the method of the fluid approximation of a Markov Process, which is inherently scalable, is appropriate for this type of scenario.

To the best of our knowledge, no previous work has studied the asymptotic behaviour of a large-scale IoT network that offloads processing towards Fog-Cloud via a shared access medium. The main contributions of this work are:

- Modeling of a large-scale population of IoT nodes that are processing, transmitting, or at idle state as a switched system via the fluid approximation of a Markov Process.
- Location of equilibrium points as a function of system parameters.

II. System Model

A set of IoT nodes (sensor nodes, mobiles, M2M devices) receives task requests. These tasks require processing that can be performed locally or offloaded to fog nodes or cloud servers, in a three-tier computing architecture (see Fig. 1). The arriving tasks follow a Poisson process with rate λ.

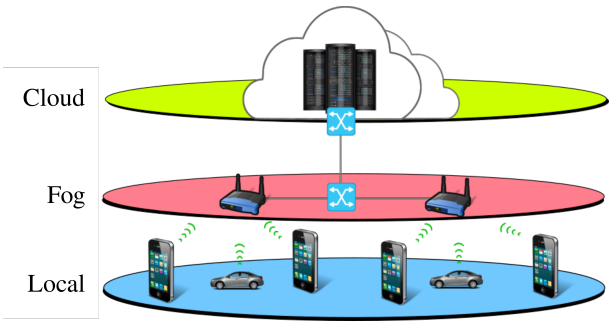


Fig. 1: Three-tier computing architecture

Suppose a new request arrives and the IoT device is available (neither processing nor transmitting). In that case, it is processed locally with probability α (the offloading factor) or transmitted wirelessly to the fog node with probability $1-\alpha$. The completion time for local processing and the transmission time are random variables with exponential distribution.

Let N be the number of IoT nodes; $\tilde{N}^N_{tx}(t)$, $\tilde{N}^N_p(t)$ and $\tilde{N}^N_{idle}(t)$ be the stochastic processes corresponding to the number of IoT nodes that are transmitting, processing and idle at time t respectively with $\tilde{N}^N_{tx}(t)+\tilde{N}^N_p(t)+\tilde{N}^N_{idle}(t) = N$.

Shared access medium

The communication channel is shared among IoT nodes and has a throughput function $\tilde{C}^N(\tilde{N}_{tx})$ that is scalable with N. Let L be the number of Fog Gateways such that $L \propto N$, each Gateway can service up to M IoT devices with a constant throughput per transmitting device k, the maximum throughput per Gateway is $c_g = Mk$ and the maximum throughput per device $c = c_g L/N$.

A medium access control is implemented, i.e ALOHA, CSMA/CA, etc. Following the analysis obtained in [8] about the throughput of a CSMA wireless network varying the number of users, a piecewise linear function with parameters k and k_1 modeling the throughput is considered:

$$\tilde{C}^N(\tilde{N}_{tx}^N) = \begin{cases} k\tilde{N}_{tx}^N, & \text{if } 0 \leq \tilde{N}_{tx}^N \leq cN/k \\ cN(1+\frac{k_1}{k}) - k_1\tilde{N}_{tx}^N, & \text{if } cN/k < \tilde{N}_{tx}^N \leq N \end{cases}$$

Initially the channel throughput increases with the number of transmitting nodes, until the maximum capacity is reached and the performance of the link decreases (see Fig. 2).

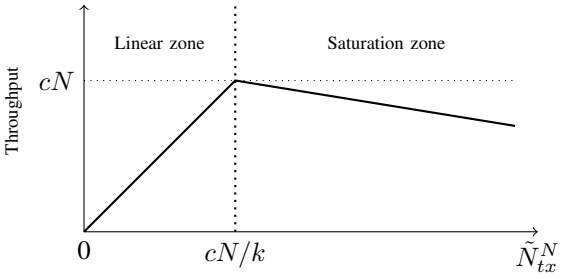


Fig. 2: Channel throughput $\tilde{C}^N(\tilde{N}_{tx}^N)$.

Number of transmitting and processing nodes as population processes

We can think of the system as a population process with a fixed number of similar "particles" of different classes. The particles, in this case, will be IoT nodes, each of them classified according to its state: transmitting, processing, or idle.

The rate at which each class evolves will depend on the intensity between the transitions of the different states. As an example, let us consider the intensity for the birth of a processing node. The birth will occur when a new task arrives at any of the idle nodes, and the system decides to process the task locally. Because the time between arrivals follows an exponential distribution with parameter λ, the time until a first task arrives to an idle node has an exponential distribution with parameter λN_{idle}. Considering the offloading factor, the intensity for the birth of a new processing node is $\alpha\lambda\tilde{N}_{idle}^N$. Similarly, the processing population will decrease by one, with the first task completed among all nodes processing locally. In this case, the death intensity of processing nodes is $\mu\tilde{N}_p^N$ where μ is the processing rate.

Analogously, the intensity for the birth of a new transmitting node is $(1-\alpha)\lambda\tilde{N}_{idle}^N$. The transmission rate for a single node (tasks transmitted per second) can be calculated as the available throughput per user $\tilde{C}^N(\tilde{N}_{tx}^N)/\tilde{N}_{tx}^N$ divided by the packet size $1/v$. Therefore, the time until a first transmission ends among all existing transmissions will be exponentially distributed with rate parameter $v\tilde{N}_{tx}^N\tilde{C}^N(\tilde{N}_{tx}^N)/\tilde{N}_{tx}^N = v\tilde{C}^N(\tilde{N}_{tx}^N)$.

The Markov Process $(\tilde{N}_p^N(t), \tilde{N}_{tx}^N(t))$ has transition rates $\tilde{q}((\tilde{N}_p^N, \tilde{N}_{tx}^N), (\tilde{N}_p'^N, \tilde{N}_{tx}'^N))$ from state $(\tilde{N}_p^N, \tilde{N}_{tx}^N)$ to state $(\tilde{N}_p'^N, \tilde{N}_{tx}'^N)$, defined by:

$$\tilde{q}((\tilde{N}_p^N, \tilde{N}_{tx}^N), (\tilde{N}_p^N + 1, \tilde{N}_{tx}^N)) = \alpha\lambda(N - \tilde{N}_p^N - \tilde{N}_{tx}^N) \quad (1)$$

$$\tilde{q}((\tilde{N}_p^N, \tilde{N}_{tx}^N), (\tilde{N}_p^N - 1, \tilde{N}_{tx}^N)) = \mu\tilde{N}_p^N \quad (2)$$

$$\tilde{q}((\tilde{N}_p^N, \tilde{N}_{tx}^N), (\tilde{N}_p^N, \tilde{N}_{tx}^N + 1)) = (1-\alpha)\lambda(N - \tilde{N}_p^N - \tilde{N}_{tx}^N) \quad (3)$$

$$\tilde{q}((\tilde{N}_p^N, \tilde{N}_{tx}^N), (\tilde{N}_p^N, \tilde{N}_{tx}^N - 1)) = v\tilde{C}^N(\tilde{N}_{tx}^N) \quad (4)$$

Fluid limit approximation

We are interested in evaluating the system's performance for a large number of IoT nodes as a function of the offloading factor. Although simulations can be run for different parameters, the asymptotic behavior can be more easily studied with an equivalent deterministic system.

For this purpose we introduce a fluid limit approximation for the Markov Process. More information about this method can be found in [9] [10] [11].

Following the notation used in [10], let $\tilde{X}^N(t)$ be a Markov process parametric in N and its Martingale decomposition:

$$\tilde{X}^N(t) = \tilde{X}^N(0) + \int_0^t \tilde{Q}^N(\tilde{X}^N(s))ds + \tilde{M}^N(t)$$

where $\tilde{M}^N(t)$ is a Martingale, $\tilde{Q}^N(l) = \sum_{m\in S}(m-l)q(l,m)$ is the process drift, $q(l,m)$ is the transition rate from state l to state m and the state space is denoted by S. Let $X^N(t) = \tilde{X}^N(t)/N$ be the scaled process,

$$X^N(t) = X^N(0) + \frac{1}{N}\int_0^t \tilde{Q}^N(\tilde{X}^N(s))ds + \frac{\tilde{M}^N(t)}{N}$$

If a Lipschitz function Q exists such that:

$$\lim_{N\to\infty} sup_{l\in S} \left\| \frac{\tilde{Q}^N(l)}{N} - Q(l/N) \right\| = 0$$

then $\frac{\tilde{M}^N(t)}{N}$ converges to zero in probability and $X^N(t)$ converges in probability to a deterministic process $x(t)$ described by the ordinary differential equation (ODE):

$$\dot{x} = Q(x(t))$$

We define $n_{tx}^N = \tilde{N}_{tx}^N/N$, $n_p^N = \tilde{N}_p^N/N$ and $n_{idle}^N = \tilde{N}_{idle}^N/N$ such that $0 \leq n_{tx}^N \leq 1$ and $0 \leq n_p^N \leq 1$. The average throughput function per IoT node, $C(n_{tx}^N) = \frac{\tilde{C}^N(\tilde{N}_{tx}^N)}{N}$ is given by:

$$C(n_{tx}^N) = \begin{cases} kn_{tx}^N, & \text{if } 0 \leq n_{tx}^N \leq c/k \\ c(1+\frac{k_1}{k}) - k_1 n_{tx}^N, & \text{if } c/k < n_{tx}^N \leq 1 \end{cases} \quad (5)$$

Proposition 1. *The scaled process* $(n_p^N(t), n_{tx}^N(t)) = \frac{1}{N}(\tilde{N}_p^N(t), \tilde{N}_{tx}^N(t))$ *converges in probability when* $N \to \infty$ *to* $(n_p(t), n_{tx}(t))$*, solution of the following deterministic ODE:*

$$(\dot{n}_p, \dot{n}_{tx}) = Q(n_p, n_{tx}) \tag{6}$$

where $Q(.)$ *is given by:*

$$Q(n_p, n_{tx}) = \begin{pmatrix} \alpha\lambda(1 - n_p - n_{tx}) - \mu n_p \\ (1-\alpha)\lambda(1 - n_p - n_{tx}) - vC(n_{tx}) \end{pmatrix}$$

The proofs on this paper are not included for lack of space.

III. Analysis of the deterministic ODE

The throughput defined in Eq. (5) generates an ODE with two right-hand sides, depending on channel congestion. For this reason, a polyhedral partition is considered: $\mathcal{X}_{lin} = \{(n_p, n_{tx}) | 0 \leq n_{tx} \leq c/k, 0 \leq n_p \leq 1, n_p + n_{tx} \leq 1\}$ and $\mathcal{X}_{sat} = \{(n_p, n_{tx}) | c/k \leq n_{tx} \leq 1, 0 \leq n_p \leq 1, n_p + n_{tx} \leq 1\}$ and $\mathcal{X} = \mathcal{X}_{lin} \cup \mathcal{X}_{sat}$ is the feasible triangle (see Fig. 3).

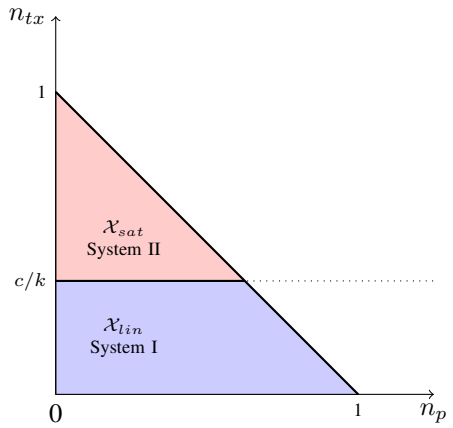


Fig. 3: Regions of the state-dependent switching system.

The system switches following a sequence $s = [(i_0, t_0), (i_1, t_1), ..., (i_N, t_N)]$ between two first-order linear systems of the form:

$$\dot{\mathbf{x}}(t - t_k) = \mathbf{A}_{\mathbf{i_k}} \mathbf{x}(t - t_k) + \mathbf{r}_{\mathbf{i_k}}, \tag{7}$$

for $t_k \leq t < t_{k+1}$ where $\mathbf{x} = (n_p, n_{tx})^T$, $\mathbf{A}_{\mathbf{i_k}}$ is the fundamental matrix, $\mathbf{r}_{\mathbf{i_k}}$ is a source input.

When $\mathbf{x} \in \mathcal{X}_{lin}$ then $i_k = 1$ and **System I** is active. If $\mathbf{x} \in \mathcal{X}_{sat}$ then $i_k = 2$ and **System II** is active. For the rest of this paper we denote $\mathbf{A}^{\text{lin}} = \mathbf{A}_{i_{k=1}}$, $\mathbf{A}^{\text{sat}} = \mathbf{A}_{i_{k=2}}$, $\mathbf{r}^{\text{lin}} = \mathbf{r}_{i_k=1}$ and $\mathbf{r}^{\text{sat}} = \mathbf{r}_{i_{k=2}}$.

We face a planar switched system that swings between two linear dynamics. We want to find the equilibria of the system and to study their stability properties. This can be done for each subsystem using the eigenvalues of the fundamental matrices. However, we must be careful, since switched systems can be unstable even when both subsystems are stable [12]. We first look at every subsystem.

Analysis of ***System I*** *(*$\mathbf{x} \in \mathcal{X}_{lin}$*)*

The dynamic in this zone is given by:

$$\begin{cases} \dot{n}_p = \alpha\lambda(1 - n_{tx} - n_p) - n_p\mu \\ \dot{n}_{tx} = (1-\alpha)\lambda(1 - n_{tx} - n_p) - kvn_{tx} \end{cases}$$

with matrices:

$$\mathbf{A}^{\text{lin}} = \begin{bmatrix} -\alpha\lambda - \mu & -\alpha\lambda \\ -(1-\alpha)\lambda & -kv - (1-\alpha)\lambda \end{bmatrix}, \ \mathbf{r}^{\text{lin}} = \begin{bmatrix} \alpha\lambda \\ (1-\alpha)\lambda \end{bmatrix}$$

Since $\mathbf{A}^{\text{lin}}$ is non-singular, there is a unique equilibrium point $\hat{n}^{lin} = (\hat{n}_p^{lin}, \hat{n}_{tx}^{lin})$ which is always asymptotically stable and it always lies inside the feasible triangle. Whether it is inside $\mathcal{X}_{lin}$ or $\mathcal{X}_{sat}$ depends on parameter values.

Analysis of ***System II*** *(*$\mathbf{x} \in \mathcal{X}_{sat}$*)*

The dynamic in this zone is given by:

$$\begin{cases} \dot{n}_p = \alpha\lambda(1 - n_{tx} - n_p) - n_p\mu \\ \dot{n}_{tx} = (1-\alpha)\lambda(1 - n_{tx} - n_p) + k_1 vn_{tx} - c(1 + k_1/k)v \end{cases}$$

with matrices:

$$\mathbf{A}^{\text{sat}} = \begin{bmatrix} -\alpha\lambda - \mu & -\alpha\lambda \\ -(1-\alpha)\lambda & k_1 v - (1-\alpha)\lambda \end{bmatrix}, \ \mathbf{r}^{\text{sat}} = \begin{bmatrix} \alpha\lambda \\ (1-\alpha)\lambda - c(1+k_1/k)v \end{bmatrix}$$

Once again, there is a unique equilibrium point $\hat{n}^{sat} = (\hat{n}_p^{sat}, \hat{n}_{tx}^{sat})$. Both its stability property and its location depends on parameter values. It can be stable or unstable and it can be inside or outside the feasible triangle.

A careful analysis, not included here, reveals a total of six different scenarios regarding the stability and location of equilibria. A thorough analytical examination indicates that all the trajectories of the switched system converge to a unique equilibrium point within the feasible triangle, either $\hat{n}^{lin}$ or $\hat{n}^{sat}$. At this stage, given the system parameters, we can analytically determine the system's attractor, that is, the steady-state population density of transmitting and processing nodes.

IV. Results

For the purpose of this study we carried out a discrete-event simulation in Python to recreate the stochastic processes presented in this paper. In this section we present the results obtained for a particular set of parameters to show that the fluid limit model is an excellent approximation of the dynamics of the simulated system.

One interesting set of parameters to simulate are those of the scenario that depending on the offloading factor we could have an equilibrium point in where the channel is saturated or not.

The simulation parameters are listed in Table I and the results are shown in Fig. 4. We can observe that for each population density (transmitting, processing and idle), the solution of the fluid model ODE is a good approximation of the simulated densities, not only regarding the equilibrium point but also the transient response.

TABLE I: Simulation parameters

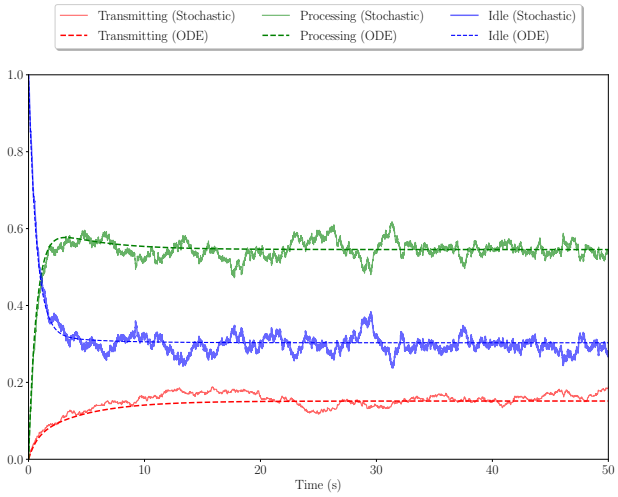
Parameter	Description	Value
N	Number of IoT devices	500
M	Number of Fog gateways	25
λ	Task arrival rate	1 task/s
μ	Task processing rate	0.5 task/s
c_g	Gateway Maximum Throughput	100 Mbps
k	Throughput in linear zone	10 Mbps/user
k_1	Throughput saturation parameter	3
v	Task transmission bit ratio	2×10^{-7} task/bit
$N_{tx}(t_0)$	Initial transmitting nodes	0
$N_p(t_0)$	Initial processing nodes	0
$N_{idle}(t_0)$	Initial idle nodes	500
T	Simulation time	50 s

In the first case with $\alpha = 0.9$ the equilibrium point is in $\mathcal{X}_{lin}$ and the trajectory never enters $\mathcal{X}_{sat}$, so there is no switching between systems. However with $\alpha = 0.1$, the system is initially in $\mathcal{X}_{lin}$ but the equilibrium point is in $\mathcal{X}_{sat}$ and a switching occurs when the transmission density reaches $c/k = 0.5$.

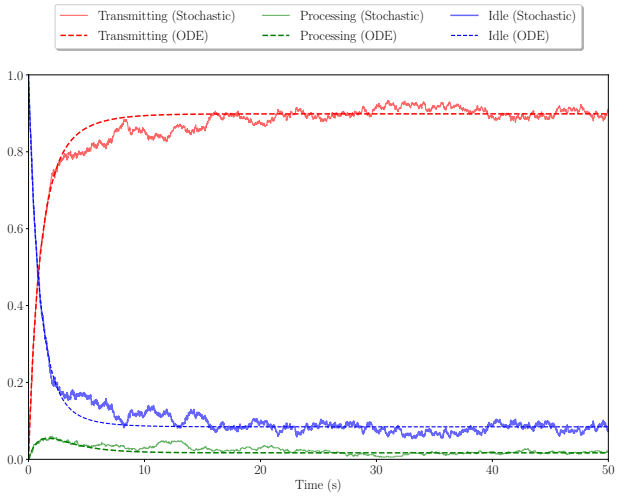
In order to numerically compare the expected equilibrium point with the mean of the stochastic processes and to reduce the effect of the transient response, we conducted a new simulation changing the initial condition to match the expected equilibrium point. Therefore in the case with $\alpha = 0.9$ we set $(N_{tx}(0), N_p(0)) = (N\hat{n}_p^{lin}, N\hat{n}_{tx}^{lin})$ and for the case with $\alpha = 0.1$ we set $(N_{tx}(0), N_p(0)) = (N\hat{n}_p^{sat}, N\hat{n}_{tx}^{sat})$. The results are shown in Table II, for each class and offloading factor we compare the mean value of the simulated system with the equilibrium coordinate of the fluid model. Although the number of IoT devices is finite we can conclude that the fluid model is a good approximation to obtain the expected value of the stochastic processes.

TABLE II: Mean value of the stochastic processes vs Fluid Model Equilibrium

	$\alpha = 0.1$		$\alpha = 0.9$	
Param.	**Mean Sim.**	**Fluid Model**	**Mean Sim.**	**Fluid Model**
n_{tx}	0.8947	0.8985	0.1503	0.1515
n_p	0.0172	0.0169	0.5512	0.5454



(a) Densities as a function of time with $\alpha = 0.9$



(b) Densities as a function of time with $\alpha = 0.1$

Fig. 4: Evolution of population densities. The stochastic trajectory of each class is compared with the corresponding fluid limit approximation. Two simulations are represented varying the offloading factor.

V. Conclusions and Future Work

This work proposes a novel approach to analyze a large-scale IoT stochastic system using a fluid limit model. The results show that the fluid model is an excellent approximation to the stochastic system. They also show that the fluid model is a suitable tool to determine the equilibrium point of a large-scale IoT network.

Future studies will have to investigate how to incorporate QoS metrics for determining an optimal offloading factor considering cost functions with QoS restrictions. Additionally, energy models can be studied to investigate the effect of energy scarcity in the dynamics of the population of transmitting and processing nodes.

Acknowledgment

This research was partially supported by ANII grant POS-NAC-2018-1-151203.

References

[1] K. Shafique, B. A. Khawaja, F. Sabir, S. Qazi, and M. Mustaqim, "Internet of things (iot) for next-generation smart systems: A review of current challenges, future trends and prospects for emerging 5g-iot scenarios," *IEEE Access*, vol. 8, pp. 23022–23040, 2020.

[2] A. Shakarami, M. Ghobaei-Arani, M. Masdari, and M. Hosseinzadeh, "A Survey on the Computation Offloading Approaches in Mobile Edge/Cloud Computing Environment: A Stochastic-based Perspective," *Journal of Grid Computing*, vol. 18, pp. 639–671, Dec. 2020.

[3] M. Masdari and H. Khezri, "Efficient offloading schemes using Markovian models: a literature review," *Computing*, vol. 102, pp. 1673–1716, July 2020.

[4] B. Liu, Q. Zhu, W. Tan, and H. Zhu, "Congestion-Optimal WiFi Offloading with User Mobility Management in Smart Communications," Aug. 2018. ISSN: 1530-8669 Pages: e9297536 Publisher: Hindawi Volume: 2018.

[5] L. Chen, S. Zhou, and J. Xu, "Computation Peer Offloading for Energy-Constrained Mobile Edge Computing in Small-Cell Networks," *IEEE/ACM Transactions on Networking*, vol. 26, pp. 1619–1632, Aug. 2018.

[6] T. Q. Dinh, Q. D. La, T. Q. S. Quek, and H. Shin, "Learning for Computation Offloading in Mobile Edge Computing," *IEEE Transactions on Communications*, vol. 66, pp. 6353–6367, Dec. 2018. Conference Name: IEEE Transactions on Communications.

[7] H. Lin, S. Zeadally, Z. Chen, H. Labiod, and L. Wang, "A survey on computation offloading modeling for edge computing," *Journal of Network and Computer Applications*, vol. 169, p. 102781, Nov. 2020.

[8] K. Voulgaris, A. Gkelias, I. Ashraf, M. Dohler, and A. H. Aghvami, "Throughput analysis of wireless CSMA/CD for a finite user population," in *IEEE Vehicular Technology Conference*, pp. 1–5, Sep. 2006.

[9] S. Ethier and T. Kurtz, *Markov Processes: Characterization and Convergence*. Wiley Series in Probability and Statistics, Wiley, 2009.

[10] C. Rattaro, L. Aspirot, E. Mordecki, and P. Belzarena, "QoS provision in a dynamic channel allocation based on admission control decisions," *ACM Trans. Model. Perform. Eval. Comput. Syst.*, vol. 5, Feb. 2020.

[11] P. Robert, *Stochastic Networks and Queues*. Stochastic Modelling and Applied Probability, Berlin Heidelberg: Springer-Verlag, 2003.

[12] M. Branicky, "Multiple Lyapunov functions and other analysis tools for switched and hybrid systems," *IEEE Transactions on Automatic Control*, vol. 43, pp. 475–482, April 1998.