Research article

UDC 004.75
PACS 02.50.Fz, 02.60.Pn,
DOI: 10.22363/2658-4670-2021-29-1-53-62

Optimization of mobile device energy consumption in a fog-based mobile computing offloading mechanism

Anastasia V. Daraseliya[1], **Eduard S. Sopin**[1, 2]

[1] *Peoples' Friendship University of Russia (RUDN University)*
6, Miklukho-Maklaya St., Moscow, 117198, Russian Federation
[2] *Federal Research Center "Computer Science and Control"*
of the Russian Academy of Sciences (FRC CSC RAS)
44-2, Vavilova St., Moscow 119333, Russian Federation

(received: March 5, 2021; accepted: March 12, 2021)

The offloading of computing tasks to the fog computing system is a promising approach to reduce the response time of resource-greedy real-time mobile applications. Besides the decreasing of the response time, the offloading mechanisms may reduce the energy consumption of mobile devices. In the paper, we focused on the analysis of the energy consumption of mobile devices that use fog computing infrastructure to increase the overall system performance and to improve the battery life. We consider a three-layer computing architecture, which consists of the mobile device itself, a fog node, and a remote cloud. The tasks are processed locally or offloaded according to the threshold-based offloading criterion. We have formulated an optimization problem that minimizes the energy consumption under the constraints on the average response time and the probability that the response time is lower than a certain threshold. We also provide the numerical solution to the optimization problem and discuss the numerical results.

Key words and phrases: queuing system, fog computing, cloud computing, queuing theory, optimization, Laplace–Stieltjes transform

1. Introduction

In recent years, fog computing has received attention from the scientific and industrial community. Many papers were related to opportunities and challenges of fog, focusing primarily on the networking context of the Internet of Things (IoT) [1]. Another one of the most popular topics and pressing research issue is the compromise between the energy-efficiency and the response time in offloading of mobile application tasks to fog computing infrastructure. The paper [2] presents the results of a study on energy consumption, execution delay and payment cost of offloading processes in a fog computing network in terms of queuing theory. Research in [3] focuses on energy-efficient task

offloading, whose main idea is taking into account both energy consumption and schedule delay under fog devices. Energy efficient offloading is also a vital task in the context of the Internet of Things concept [4].

In our previous paper [5], we developed an analytical framework for response time analysis that takes into account the variation of tasks in terms of processing volume. Then in the paper [6], we analyze the two-parameter offloading mechanism that takes into account both the computing complexity and the data size to be transferred in case of offloading. In [7] we derived the cumulative distribution function of the response time in terms of Laplace-Stieltjes Transform. In the current work, we solve the optimization problem by minimizing energy efficiency, subject to the average time constraint and taking into account the probability that the time exceeds a given threshold.

2. Mathematical model

We consider a distributed computing system that consist of mobile devices (MDs), a fog node and a remote cloud. MDs run real-time applications that require significant amount of computational resources. For each task, a MD makes a decision, whether it will be offloaded to the fog node or processed locally. The capacity of the fog node is limited, which means if there are too many tasks offloaded, then some of the offloaded tasks are redirected to the remote cloud to prevent the fog node congestion. In terms of queuing theory, the considered system can be represented as shown in the figure 1.

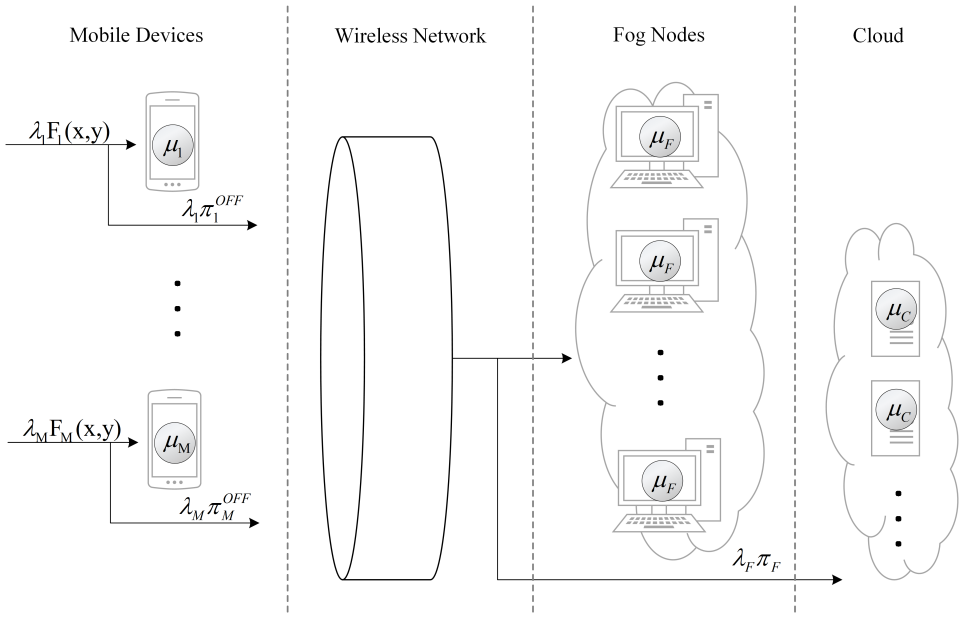

Figure 1. Mathematical model in terms of queuing network

Assume there are M MDs, each of them generating a flow of tasks with exponentially distributed interarrival times according to Poisson's law with intensity λ_i, $i = 1, \ldots, M$. Each task is characterized by the amount of processing volume required and the data size to be transferred in case of

offloading. We assume that the processing volume (measured in millions of instructions, MI) and the data size (measured in MB) are independent random variables with CDFs $W_i(x)$ and $S_i(x)$, probability density functions (PDF) $w_i(x)$ and $s_i(x)$ respectively. MDs process locally served tasks in the FCFS mode with constant serving rate μ_i, $i = 1, \dots, M$ (measured in MIPS).

We propose the offloading mechanism that implies offloading tasks that are "heavy" in terms of processing volume and "light" in terms of data size. Splitting to "heavy" and "light" tasks are done by the threshold O_w on the processing volume and the threshold O_s on the data size. Hence, the offloading probability $\pi_{i,O}$ on the i-th MD is evaluated [6], [7] according to the following formula

$$\pi_{i,O} = \int_{O_w}^{\infty} w_i(x)\,dx \int_0^{O_s} s_i(y)dy = (1 - W_i(O_w))S_i(O_s). \tag{1}$$

If a task is processed locally, then the response time consists of processing time on an MD only. If a task is offloaded to the fog node, then the total response time is the sum of task transmission time to the fog node through wireless network, the processing time on the fog node and the transmission time back to the MD. If the fog node is overloaded and an offloaded task is sent to the remote cloud, then the processing time at the fog is replaced by the transmission time between the fog node and the remote cloud, the processing time on the cloud and the transmission time back to the fog node.

We assume that the wireless network provides total bitrate R, which is used to transmit the data of tasks one-by-one in FCFS order, so the transmission time is obtained as the fraction of the data size of a task and total bitrate R. On the other side, the transmission time between the fog node and the cloud is assumed constant.

The fog node provide computational resources by means of virtual machines (VMs), each of them having the constant serving rate λ_F. The total number of VMs at the fog node is N. The constant serving rate μ_C of VMs at the cloud is greater than μ_F, and amount of computational resources (VMs) at the remote cloud is assumed to be large enough, so that it cannot be overloaded.

3. The response time analysis

3.1. CDFs of the response time components

The service process at MD i is modeled in terms of a queuing system $M/G/1$ with arrival intensity λ_i, $\sum\limits_{i=1}^{M} \lambda_i = \lambda$. The distribution function of the processing volume on a MD can be determined by conditional CDF $W_{MD,i}(x)$ as follows

$$W_{MD,i}(x) = \begin{cases} \dfrac{W_i(O_w) + (W_i(x) - W_i(O_w))(1 - S_i(O_s))}{1 - \pi_{i,O}}, & x > O_w, \\ \dfrac{W_i(x)}{1 - \pi_{i,O}}, & x \leqslant O_w. \end{cases} \tag{2}$$

Having obtained the distribution function of the processing volume $W_{MD,i}(\mu_i x)$, we can find the serving time at a MD. By virtue of the fact that the serving rate on the i-th MD is constant and being μ_i, its CDF is easily obtained as $T_{MD,i}(x) = W_{MD,i}(\mu_i x)$. The average serving time at MD i can be found through integration using the CDF $T_{MD,i}$.

If a task is offloaded to the distributed computing infrastructure, it is first transferred through the wireless network to the fog node. The delays in wireless networks are obtained analogously, by employing $M/G/1$ queue. The arrival intensity λ_F is the sum of the offloading intensities from all MDs $\lambda_F = \sum_{i=1}^{M} \lambda_i \pi_{i,O}$. The CDF $S_{tr,i}(x)$ of the file size to be transmitted is

$$S_{tr,i}(x) = \begin{cases} \dfrac{1}{\pi_{i,O}(1 - W_i(O_w))S_i(x)}, & x \leqslant O_s, \\ 1, \quad x > O_s, \end{cases} \tag{3}$$

and the service time distribution in the wireless network is $T_{tr,i}(x) = S_{tr,i}(Rx)$.

At the fog node, there are N VMs to serve offloaded tasks, so the service process may be modeled by $M/G/N/0$ queue, where the blocked customers are redirected to the next layer — remote cloud. The arrival intensity is the same as for wireless network — λ_F. The service time is determined by

$$W_{F,i}(x) = \begin{cases} \dfrac{1}{\pi_{i,O}}(W_i(x) - W_i(O_w))S_i(O_s), & x > O_w, \\ 0, \quad x \leqslant O_w. \end{cases} \tag{4}$$

The service time is simply processing volume divided by the service rate μ_F, so the CDF of the service time at the fog node is

$$T_{F,i}(x) = W_{F,i}(\mu_F x), \tag{5}$$

$$T_F(x) = \frac{\lambda_i \pi_{i,O}}{\lambda_F} T_{F,i}(x). \tag{6}$$

The probability that a task is redirected to the remote cloud π_F is obtained from Erlang formula for $M/G/N/0$ queues as

$$\pi_F = \frac{(\lambda_F \tau_F)^N}{N!} \left(\sum_{k=0}^{N} \frac{(\lambda_F \tau_F)^k}{k!} \right), \tag{7}$$

where τ_F is the average serving time at the fog node, which can be easily evaluated from the CDF $TF(x)$. The service time distribution $T_C(x)$ at the cloud is

$$T_{C,i}(x) = W_{C,i}(\mu_C x). \tag{8}$$

3.2. The average response time

The total response time is the conditional sum of processing and transmission delays. A task from i-th MD is processed locally with probability $1-\pi_{i,O}$ on the fog node with probability $\pi_{i,O}(1-\pi_F)$ and on the cloud with probability $\pi_{i,O}\pi_F$. In [7], we derived the Laplace-Stieltjes Transforms (LST) for all delay components for the case of Gamma distribution of both processing volume and data size, and obtain the LST of the total response time.

First, we derived the LST $\tilde{T}_{MD,i}(s)$ of the service time at the MD i as

$$\tilde{T}_{MD,i}(s) = \int_0^\infty e^{-sx}\, d(T_{MD,i}(x)) = \frac{1}{1-\pi_{i,O}}\left[\frac{\mu_i^2}{(s\delta_W+\mu_i)^2} - \right.$$
$$\left. - \left(1 - e^{-\frac{O_s}{\delta_s}}\left(1+\frac{O_s}{\delta_s}\right)\right) e^{-\left(\frac{sO_w}{\mu_i}+\frac{O_w}{\delta_w}\right)} \frac{\mu_i O_w(s\delta_w+\mu_i)+\mu_i^2\delta_w}{\delta_w(s\delta_w+\mu_i)^2}\right] \quad (9)$$

with the the LST of $\phi_{MD,i}$ the sojourn time distribution on mobile device i

$$\omega_{MD,i}(s) = \frac{s(1-\rho_i)}{s-\lambda_i+\lambda_i\tilde{T}_{MD,i}(s)}, \quad (10)$$

$$\phi_{MD,i}(s) = \tilde{T}_{MD,i}(s)\omega_{MD,i}(s). \quad (11)$$

Then we obtained the LST $\tilde{T}_{F,i}(s)$ and $\tilde{T}_{C,i}(s)$ of the service time distribution at the fog node and cloud, respectively. LST $\tilde{T}_{F,i}(s)$ of the service time distribution at the fog node is derived from CDF

$$\tilde{T}_{F,i}(s) = \int_0^\infty e^{-sx}\, d(T_{F,i}(x)) = e^{-\left(\frac{sO_w}{\mu_F}\right)} \frac{\mu_F O_w(s\delta_w+\mu_F)+\mu_i^2\delta_w}{\delta_w(s\delta_w+\mu_i)^2}. \quad (12)$$

LST of the service time distribution in the cloud is obtained by analogy with $\tilde{T}_{F,i}(s)$.

The LST $\omega_{tr,i}(s)$ of the waiting time distribution and the LST $\phi_{tr,i}(s)$ of sojourn time in the wireless network are:

$$\omega_{tr,i}(s) = \frac{s(1-\rho_{tr})}{s-\lambda_F+\lambda_F\tilde{T}_{tr,i}(s)}, \quad \phi_{tr,i}(s) = \tilde{T}_{tr,i}(s)\omega_{tr,i}(s). \quad (13)$$

Having obtained the LST of all these delay component distributions, we made use of the convolution formula and obtain the LST $\tilde{\tau}(s)$ of the response time distribution of a task from MD i:

$$\tilde{\tau}_i(s) = (1-\pi_{i,0})\phi_{MD,i}(s) + \pi_{i,0}(1-\pi_F)\tilde{T}_{F,i}(s)\phi_{tr,i}^2(s) +$$
$$+ \pi_{i,0}\pi_F\tilde{T}_{C,i}(s)\phi_{tr,i}^2(s)\tilde{T}_{FC}^2(s). \quad (14)$$

After this we used numerical Reverse LST $\tilde{\tau}(s)$ to evaluate the CDF $\tau(s)$ of the response time.

Actually, the average response time can be calculated as

$$\tau = \sum_{i=0}^{M} \frac{\lambda_i}{\left(\sum\limits_{j=1}^{M} \lambda_j\right)} \tau_i. \tag{15}$$

The resulting expressions allow to get the probability $\Pi(T)$ that the response time is lower than a threshold T

$$\Pi(T) = \tau(T). \tag{16}$$

4. The energy consumption analysis

In this section, we present the formulas for the average power consumption of MDs obtained at an earlier stage of research [6].

The energy consumption for tasks processed on MD is proportional to the processing volumes of tasks, therefore the average energy consumption $E_{pr,i}$ during locally executing on i-th MD can be evaluated as follows:

$$E_{pr,i} = P_{pr,i} t_{MD,i}, \tag{17}$$

where $P_{pr,i}$ is the power consumption (W) during the processing of the i-th MD, which is considered constant for simplicity of calculations.

The average file size transmitted by i-th MD, can be calculated through integration using CDF $S_{tr,i}(x)$ from the previous section.

The energy consumption during transmitting is also proportional to the transmission time, so the average energy consumption $E_{tr,i}$ of the i-th VD during task transmission is

$$E_{tr,i} = P_{tr,i} \frac{\theta_i}{R}. \tag{18}$$

Then the average energy consumption for any i-th MD is the weighted sum of processing and transmission energies:

$$E_i = (1 - \pi_{i,0}) E_{pr,i} + \pi_{i,0} E_{tr,i}. \tag{19}$$

At the end, we can evaluate the average energy consumption for a task from an arbitrary MD as follows

$$E = \sum_{i=0}^{M} \frac{\lambda_i}{\left(\sum\limits_{j=1}^{M} \lambda_j\right)} E_i. \tag{20}$$

5. Optimization problem

In order to find the minimum energy consumption E under constraints on the average response time and the probability $\Pi(T)$ that the response time is lower than a threshold T, we formulate the optimization problem as follows:

$$\begin{cases} E = \sum\limits_{i=0}^{M} \dfrac{\lambda_i}{\left(\sum\limits_{j=1}^{M} \lambda_j\right)} E_i \to \min, \\ \tau = \sum\limits_{i=0}^{M} \dfrac{\lambda_i}{\left(\sum\limits_{j=1}^{M} \lambda_j\right)} \tau_i \leqslant T, \\ \Pi(T) \leqslant \Pi^*. \end{cases} \tag{21}$$

6. Numerical results

In this section, we presented the numerical results of our study. The main metric of interest here is the minimum power consumption E for a task from an arbitrary MD under the constraints from the optimization problem.

We consider a system with $M = 20$ homogeneous MDs that run the same applications, so the distributions of processing volume and data size of tasks are also the same. The fog nodes can run maximum $N = 8$ VMs. All values of parameters used in the section are gathered in table 1.

Table 1

Parameter values for the numerical analysis

Parameter	Value
M	20
N	8
R	150 Mbps
λ_i	2 tasks/s
μ_i	4 MIPS
μ_F	6 MIPS
μ_C	10 MIPS
δ_s	0.25
δ_ω	0.75
t_{FC}	0.5 s
P_{pr}	16 W
P_{tr}	0.2 W

Figure 2 shows the probability $\Pi(T)$ as a function of processing volume threshold with the response time threshold $T = 0.5$. The power consumption graph begins to descend only at values of $\Pi^* = 0.94$ and above. This shows that only with a very high threshold value of the probability $\Pi(T)$ that the response time is lower than a threshold T, there will be a gain in terms of energy costs.

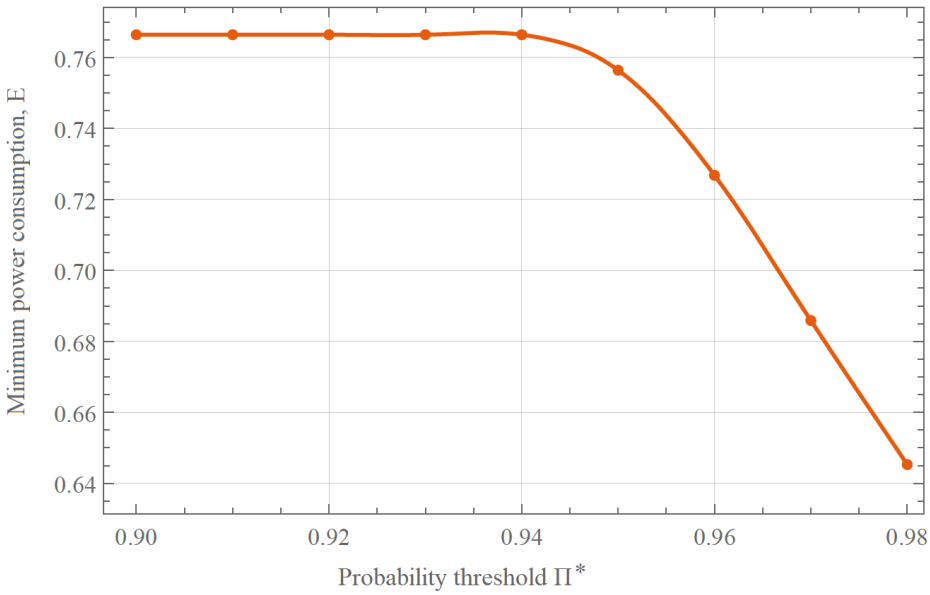

Figure 2. Dependence of the minimum power consumption $E \to \min$ on the threshold value Π^*, $T = 0.5$

7. Conclusions

In the paper, we focused on the analysis of the reducing energy consumption of MD's that use fog computing infrastructure to increase the performance and to improve the battery life of mobile devices. We have formulated and solved the problem of energy consumption optimization using constraints on the average response time and the probability that the response time is lower than a certain threshold, on the basis of which we offer some recommendations for offloading the system.

Acknowledgments

This paper has been supported by the RUDN University Strategic Academic Leadership Program (recipient Sopin E., mathematical model development). The reported study was funded by RFBR, project number 20-07-01052 (recipient Daraseliya A., optimization problem). The reported study was funded by RFBR, project number 19-07-00933 (recipient Sopin E., numerical analysis).

References

[1] 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, 2016. DOI: 10.1109/JIOT.2016.2584538.

[2] Z. Chang, Z. Zhou, T. Ristaniemi, and Z. Niu, "Energy efficient optimization for computation offloading in fog computing system," in *GLOBECOM 2017 — 2017 IEEE Global Communications Conference*, 2017, pp. 1–6. DOI: 10.1109/GLOCOM.2017.8254207.

[3] Y. Jiang, Y. Chen, S. Yang, and C. Wu, "Energy-efficient task offloading for time-sensitive applications in fog computing," *IEEE Systems Journal*, vol. 13, no. 3, pp. 2930–2941, 2019. DOI: 10.1109/JSYST.2018.2877850.

[4] Q. Li, J. Zhao, Y. Gong, and Q. Zhang, "Energy-efficient computation offloading and resource allocation in fog computing for Internet of Everything," *China Communications*, vol. 16, no. 3, pp. 32–41, 2019. DOI: 10.12676/j.cc.2019.03.004.

[5] E. S. Sopin, A. V. Daraseliya, and L. M. Correia, "Performance analysis of the offloading scheme in a fog computing system," in *2018 10th International Congress on Ultra Modern Telecommunications and Control Systems and Workshops (ICUMT)*, 2018, pp. 1–5. DOI: 10.1109/ICUMT.2018.8631245.

[6] E. Sopin, K. Samouylov, and S. Shorgin, "The analysis of the computation offloading scheme with two-parameter offloading criterion in fog computing," pp. 11–20, 2019. DOI: 10.1007/978-3-030-34914-1_2.

[7] E. Sopin, N. Zolotous, K. Ageev, and S. Shorgin, "Analysis of the response time characteristics of the fog computing enabled real-time mobile applications," *Lecture Notes in Computer Science*, vol. 12525, pp. 764–779, 2020. DOI: 10.1007/978-3-030-65726-0_9.

For citation:

A. V. Daraseliya, E. S. Sopin, Optimization of mobile device energy consumption in a fog-based mobile computing offloading mechanism, Discrete and Continuous Models and Applied Computational Science 29 (1) (2021) 53–62. DOI: 10.22363/2658-4670-2021-29-1-53-62.

Information about the authors:

Daraseliya, Anastasia V. — PhD student of Department of Applied Probability and Informatics of Peoples' Friendship University of Russia (RUDN University) (e-mail: avdaraseliya@sci.pfu.edu.ru, phone: +7(495)9550927, ORCID: https://orcid.org/0000-0002-6603-2596)

Sopin, Eduard S. — Candidate of Physical and Mathematical Sciences, Assistant professor of Department of Applied Probability and Informatics of Peoples' Friendship University of Russia (RUDN University); Senior Researcher of Institute of Informatics Problems of Federal Research Center "Computer Science and Control" Russian Academy of Sciences (e-mail: sopin-es@rudn.ru, phone: +7(495)9550927, ORCID: https://orcid.org/0000-0001-9082-2152)

УДК 004.75
PACS 02.50.Fz, 02.60.Pn,
DOI: 10.22363/2658-4670-2021-29-1-53-62

Оптимизация энергопотребления мобильных устройств в системе туманных вычислений

А. В. Дараселия[1], **Э. С. Сопин**[1, 2]

[1] *Российский университет дружбы народов*
ул. Миклухо-Маклая, д. 6, Москва, 117198, Россия
[2] *Федеральный исследовательский центр «Информатика и управление» РАН*
ул. Вавилова, д. 44, кор. 2, Москва, 119333, Россия

Выгрузка задач мобильных вычислений в систему туманных вычислений представляется многообещающим подходом для снижения времени отклика ресурсоёмких мобильных приложений, функционирующих в режиме реального времени. Помимо снижения времени отклика, механизмы выгрузки вычислений помогут также снизить энергопотребление мобильных устройств. В этой статье мы проводим анализ энергопотребления мобильных устройств, которые используют инфраструктуру туманных вычислений для повышения производительности и увеличения времени их автономной работы. Рассматривается трёхуровневая вычислительная система, состоящая из непосредственно мобильного устройства, узла системы туманных вычислений и удалённого облака. Задачи мобильных вычислений могут быть обработаны локально на устройстве или быть выгружены в соответствии с пороговым критерием выгрузки. Сформулирована и решена задача оптимизации энергопотребления при наличии ограничений на среднее время отклика и на вероятность того, что время отклика ниже определённого порога.

Ключевые слова: система массового обслуживания, туманные вычисления, облачные вычисления, оптимизация, преобразование Лапласа–Стилтьеса