System Model for Multi-level Cloud Based Tactile Internet System

Abdelhamied A. Ateya, Anastasia Vybornova, Konstantin Samouylov, Andrey Koucheryavy

System Model for Multi-level Cloud Based Tactile Internet System

Abdelhamied A. Ateya[1], Anastasia Vybornova[1], Konstantin Samouylov[2], and *Andrey Koucheryavy*[1]

[1] St. Petersburg State University of Telecommunication, 22 Prospekt Bolshevikov, St. Petersburg, Russia
[2] Peoples' Friendship University of Russia (RUDN University), 6 Miklukho-Maklaya St, Moscow, 117198, Russia

a_ashraf@zu.edu.eg, a.vybornova@gmail.com, akouch@mail.ru

Abstract. With the realization of 5G system which will become a fact by 2020, there is a great demand to achieve the Tactile Internet system. Tactile Internet system should handle a 1ms communication latency, which is the main problem of the system realization. One of the proposed system structures to achieve such latency is to build the system based on the multilevel cloud architecture and the 5G network structure. In this work, we build a system model for a multi-level cloud based Tactile Internet system. The model is used to find the system latency and evaluate the system performance. The proposed system is simulated and the results show that the system will achieve a lower latency than other known architectures. The proposed model also reduces the overall network congestion. It can be used to optimize the number of clouds in the system to achieve the best system performance.

Keywords. Tactile Internet, Cloud, System model, Mobile Edge Computing, Ultra-low Latency, 5G.

1 Introduction:

Unlike the existing cellular networks, the future 5G cellular system will support new machine type communication services. By achieving the waited cellular system in 2020, Tactile Internet may become a reality. Tactile Internet will enable human -to-machine (H2M) communication and interaction, which consequently will enable a new era for the communication networks [1]. Tactile systems will have massive applications in many fields such as smart cities, education, health care, augmented reality and smart grid [2]. The main challenge with the design and development of the Tactile Internet is the 1ms round trip delay.

Mobile edge computing (MEC) is one of the key features that will enable the development and realization of the 5G system. MEC merge the three technologies of the mobile Internet, mobile computing and cloud computing [3]. The first technology is the mobile Internet which represents the wireless communication network or the cel-

lular network. The second technology is the mobile computing that is represented by the techniques used for executing wireless communications. Mobile computing includes both hardware and software involved in the communication process such as protocols, user equipment and the network infrastructure. The last part is the cloud computing that provides a way for resources sharing or in other word deliver everything for the user as a service at the time and place the user need it.

Moving clouds closer to the user (approximately one hope away from the user) will allow to achieve lower latency and thus enable the realization of real time haptic communication that is known as Tactile Internet, which becomes a very promising area of research. MEC replaces the large and expensive data centers with small distributed cloud units connected to the cellular network [4]. These small cloud units have limited capabilities in terms of processing and storage. There are a lot of studies that suggest places for the edge computing unit in order to achieve better latency.

In [5] we suggest a multi-level cloud based Tactile Internet system. The system consists of three cloud levels: Micro-cloud, Mini-cloud and Core network cloud level as shown in Fig. 1. Micro-clouds are employed in each cellular cell and connected to the radio access network (RAN), and thus they are one hop away from users. Each group of Micro-clouds is connected to a Mini-cloud unit which has higher capabilities and can process much complex tasks. The second level of clouds (Mini-clouds) act as the controller for the first level (Micro-clouds) connected to it. They also can perform tasks that exceed the workload of the Micro-clouds connected to it and tasks that need processing capabilities greater than that of the Micro-cloud. Mini-clouds are connected to the core network cloud that represents the third level of clouds [5].

Presenting a new level of higher capability clouds in the way between core network and RAN's clouds reduces the communication latency and the throughput. Thus, multi-level cloud based Tactile Internet system reduces the round trip delay by applying multilevel hierarchical of cloud units. In this paper, we build a mathematical model for the multi-level cloud based Tactile Internet system. The model is used to find the latency of the system and evaluate the system performance. In Sec. 2 the mathematical model of the system is discussed and the total latency is calculated. In Sec. 3 the system model is simulated over Java environment. Sections 4 concludes the paper and describes the future work.

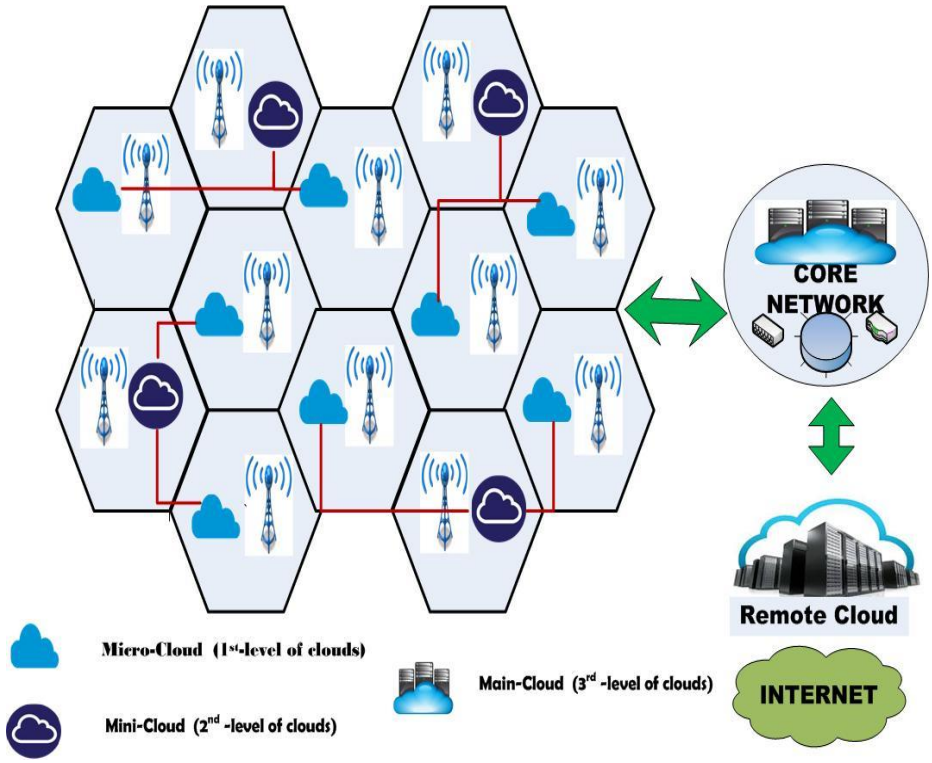

Fig. 1. Tactile Internet system [5].

2 Mathematical model

In this section, we introduce a mathematical model for the multi-level cloud based Tactile Internet system introduced in [5]. The model is used to find the latency and evaluate the system performance. In order to design a low latency Tactile Internet system with the desired performance, a mathematical model for this case is introduced and all the important parameters are defined. Figure 2 illustrates the system model for the Tactile Internet based on the multi-levels of cloud units.

In our system, each cellular cell eNB is connected to a small cloud unit (Micro-cloud) $C_{micro}(i)$ with acceptable processing elements, where, $i \in \{1, 2\ldots M\}$ and M is total number of Micro-clouds in the network. Each group of Micro-clouds are connected to larger cloud unit known as Mini-cloud $C_{mini}(j)$, where, $j \in \{1, 2\ldots N\}$ and N is the total number of Mini-clouds in the network. The Mini-cloud has higher processing and storage capabilities, and used to handle higher performance tasks that cannot be handled by Micro-clouds. Each Mini-cloud also acts as a controller for Micro-clouds connected to it. Mini-clouds represent the gateway between Micro-

clouds and the core network. In our model, we assume that each Mini-cloud unit is connected to a fixed number S of Micro-cloud units.

The rate of tasks offloaded to the Micro-cloud unit changes based on the cell users demands. Thus, we assume the tasks randomly arrived based on the Poisson process with a Poisson rate of λ_i. Each cell produces a workload W_i to the connected edge computing unit $C_{micro}(i)$ with a Poisson rate λ_i. Every Micro-cloud can handle tasks offloaded by the corresponding eNB, but in case the processing demands of the current tasks is equal or higher than the maximum workload $W_{cmax}(i)$, new tasks are moved to the Mini-cloud unit, until the resources of the Micro-cloud are released. Therefore, each Micro-cloud unit holds $W_{micro}(i)$ workloads and other non-handled tasks are shifted to the Mini-cloud unit. The computing time of Micro-cloud unit depends on the delivered work load $W_{dmicro}(i)$.

Each Mini-cloud unit can handle up to $W_{mmax}(j)$ of the work load, where, $W_{mmax}(j)$ is the maximum workload of the Mini-cloud unit $C_{mini}(j)$. Tasks that require higher processing capabilities than current free processing resources of the Mini-cloud unit are shifted to the core network cloud.

We consider the multi-server queuing model M/M/s [6] to model the Micro- and Mini-clouds. For Micro-clouds the model is $M/M/S_{mic}$ and for Mini-clouds $M/M/S_{min}$, where S_{mic} and S_{min} are the numbers of servers in the Micro and Mini-cloud unit respectively.

The total latency consists of the task's response time and the communication delay. The average response time for the tasks at the Micro and Mini-clouds is the sum of the queuing time and the processing time of the tasks. The average processing time of the tasks in the Micro and Mini-cloud units can be calculated as a function of the arrival rate λ, based on the M/M/S queuing model and the Erlang's C formula as deduced in [7][8].

$$T_{micro-i}(\lambda) = \frac{C\left(s_i, \frac{\lambda_i}{\mu_i}\right)}{s_i\mu_i - \lambda_i} + \frac{1}{\mu_i} \quad (1)$$

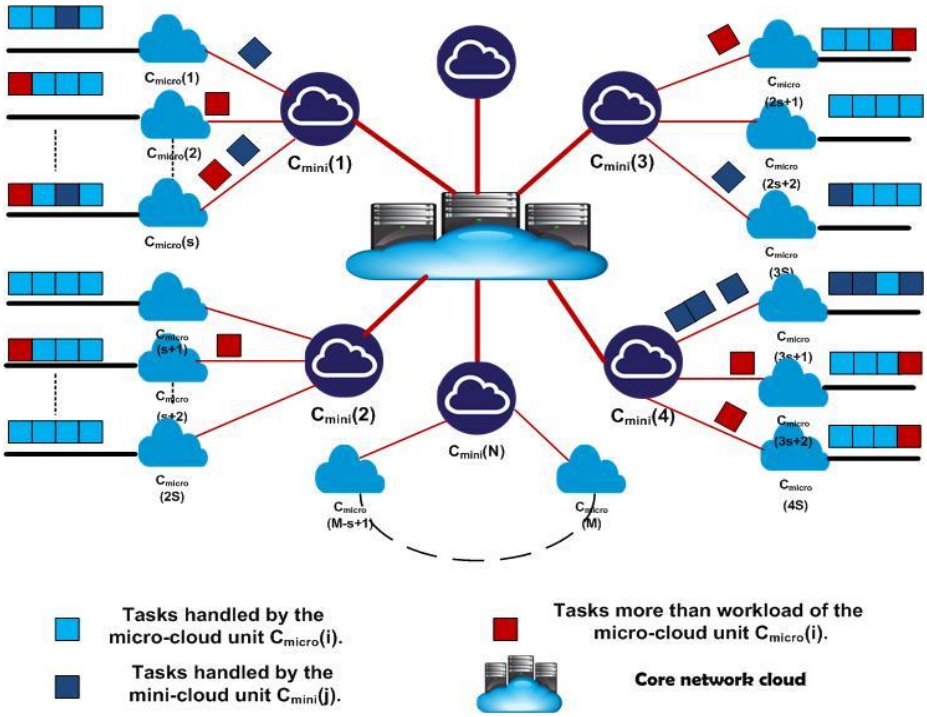

Fig. 2. System model for the Tactile Internet system.

$$T_{mini-j}(\lambda) = \frac{C\left(s_j, \frac{\lambda_j}{\mu_j}\right)}{s_j\mu_j - \lambda_j} + \frac{1}{\mu_j} \tag{2}$$

$$C(n,\rho) = \frac{\left(\frac{(s\rho)^c}{n!}\right)\left(\frac{1}{1-\rho}\right)}{\sum_{k=0}^{n-1}\frac{(n\rho)^k}{k!} + \left(\frac{(n\rho)^c}{n!}\right)\left(\frac{1}{1-\rho}\right)} \tag{3}$$

Where $T_{micro\text{-}i}$ is the average processing time of tasks in Micro-cloud unit i, $T_{mini\text{-}j}$ is the average processing time of tasks in Mini-cloud unit j, S_i is the total number of servers in the Micro-cloud unit i, S_j is the total number of servers in the Mini-cloud unit j, λ_i and λ_j are the arrival rates of the Micro-cloud unit i and Mini-cloud unit j and μ_i and μ_j are the corresponding service rates.

To simplify the calculation of the total latency we can assume that the latency function is a linear function and as indicated in [9] this assumption is acceptable and verified. The total latency can be calculated simply as following:

a- If the task is handled by Micro-cloud unit:

$$T_{T-micro-i}(w_i) = f_c(w) + d_{cell} = [\alpha(w_{dmicro-i}) + \beta] + d_{cell} \tag{4}$$

Where $T_{T\text{-micro-i}}$ is the total latency for the offloaded tasks of the Micro-cloud unit i, f_c is the linear function that is used to determine the processing delay for the current work load and d_{cell} is the communication latency inside the cellular cell.

b- If the task is moved and handled by the Mini-cloud unit:

$$T_{T-mini-j}(w_j) = f_c(w) + d_{cell} = \left[\alpha\left(w_{dmini-j}\right) + \beta\right] + d_{cell} + d_{C_{Micro-i},C_{Mini-j}} \quad (5)$$

Where $T_{T\text{-mini-j}}$ is the total latency for the offloaded tasks of the Mini-cloud unit j and $d_{C_{Micro-i},C_{Mini-j}}$ is the communication delay between micro-cloud unit i and Mini-cloud unit j.

3 Simulation and results

In this section, we analyze the suggested system model for the Tactile Internet system in a simulation environment, after the mathematical model is defined in the previous section.

a- Simulation environment and simulation parameters

There are many simulation environments that are used to simulate and deploy Micro-cloud and Clouds with different facilities and capabilities [10], [11]. These environments are able to create virtual machines (VM), remote procedure execution and web services with different capabilities. We developed a tool kit based on the CloudSim framework to analyze the system. The simulator is based on Java language and on the IDE NetBeans. The simulation is run on Window 7 basic (64-bit) and i7 Processor with 3.07 GHz of speed and memory of 8GB.

We build a system consists of 300 Micro-cloud units distributed randomly and connected to 20 Mini-cloud units. Each Micro-cloud unit represents the offload to a cellular cell. Each Mini-cloud unit is connected and controls 15 Micro-cloud units. This number is a design parameter and should be optimized in terms of achieving the lowest latency and best performance of the system. We assume that all Micro-clouds have equal capabilities and also all Mini-clouds are the same. The application tasks are sent and distributed to the Micro-clouds randomly. All important simulation parameters are illustrated in table1.

b- Simulation results and analysis

We consider three simulation cases with three different values for Wcmax. In each case the latency for each Micro and Mini- cloud unit is calculated two times. The first is for the simulated system and the second is the theoretical one. The theoretical latency of Micro and Mini-cloud units is calculated using equations (4) and (5) based on the amount of workload delivered to the cloud unit.

In the first case, we assume a workload of 20 events per second and tasks higher than this workload will be directed to Mini-cloud unit. This will put a load on the Mini-cloud units. Figure 3 shows the total latency for each Micro-cloud unit for both

theoretical and simulation models. The average latencies for the theoretical and simulation cases are 0.698 and 0.70 milliseconds respectively, and it seems to be very near for all Micro-clouds because they are of the same parameters. Theoretical latency varies from one cloud unit to another based on the amount of delivered workload since the tasks are distributed randomly. Figure 4 indicates the total latency for each Mini-cloud unit and the average total latencies for theoretical and simulation cases are 1.127 and 1.13 milliseconds. It is clear that the latency for Mini-cloud units is much higher as there is an additional communication hop between the Mini and Micro-cloud units. Without Mini-clouds, the latency is supposed to be much higher as the tasks would be delivered to the core network and the core network would be loaded with more workload.

In the second case, we increase the maximum work load of the Micro-cloud units to 30 and thus it should reduce the total load on the Mini-cloud units. This is because Micro-cloud units in this case handle many tasks and therefore reduce the number of tasks moved to Mini-cloud units. Figure 5 shows the total latency for each Micro-cloud unit compared to the delay according to the theoretical model. The average latencies for the theoretical and simulation cases are 0.838 and 0.84 milliseconds respectively and it increased because of the increased workload. Figure 6 indicates the total latency for each Mini-cloud unit. The average latencies are 0.978 and 0.98 milliseconds and this is less than the first case as the tasks moved from Micro-cloud units are less than that in the first case.

In the third case, the maximum work load of all Micro-cloud units is set to 40 and the delay for Micro and Mini- clouds is recorded and compared to the delay according to the theoretical model as indicated in Fig. 7 and 8.

The time delay for each of Micro and Mini-clouds in the previous cases can be found to be relatively near to the theoretical results. The average delays for all Micro-cloud units and Mini-cloud units in the three cases are compared with that of the theoretical model in Table 2. Finally, the average workload for all Micro and Mini-clouds for each of the previous cases is illustrated in Figure 9.

Table 1. Simulation parameters.

parameter	Description	value
M	Number of Micro-clouds in the network	300
N	Number of Mini-clouds in the network	20
S	Number of Micro-cloud units connected to each Mini-cloud unit	15
W_{mmax}	Maximum work load of the Mini-cloud unit per second	100 events/s
W_{cmax}	Maximum work load of the Micro-cloud unit per second	(20,30,40) events/s
λ_i	Arrival rate of the Micro-cloud unit	15

μ_i	Service rate of the Micro-cloud unit	5 Mbps
μ_j	Service rate of the Mini-cloud unit	8 Mbps
d_{cell}	The communication latency inside the cellular cell	1 ms/hop
$d_{C_{Micro-i},C_{Mini-j}}$	The communication delay between micro-cloud unit and Mini-cloud unit	1.5 ms
α	Gradient of computing function	10
β	Constant of computing function	0
RAM,HDD	Micro-cloud RAM, Storage Mini-cloud RAM, Storage	1024Mb,1Gb 2048 Mb, 5Gb

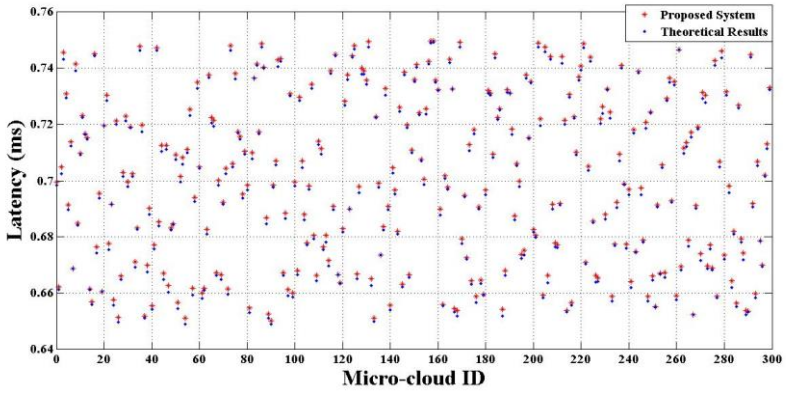

Fig. 3. Latency of Micro-cloud units in case (1).

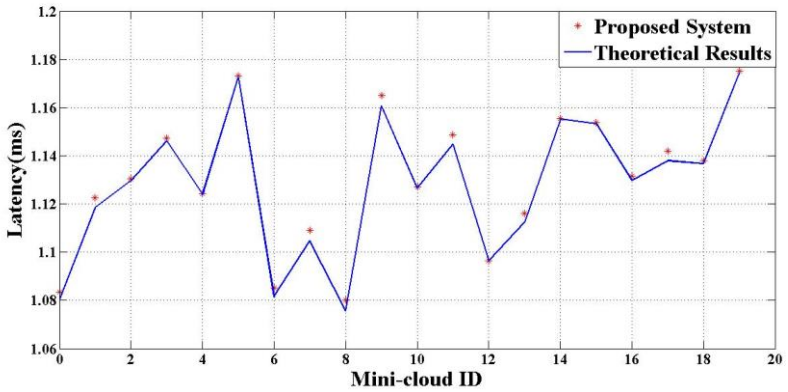

Fig. 4. Latency of Mini-cloud units in case (1).

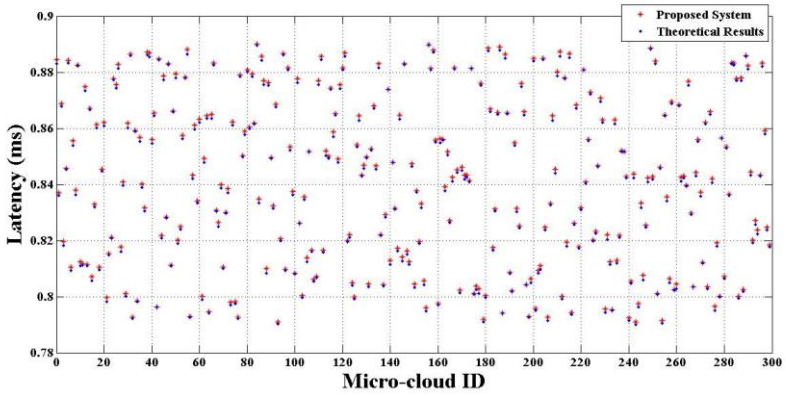

Fig. 5. Latency of Micro-cloud units in case (2).

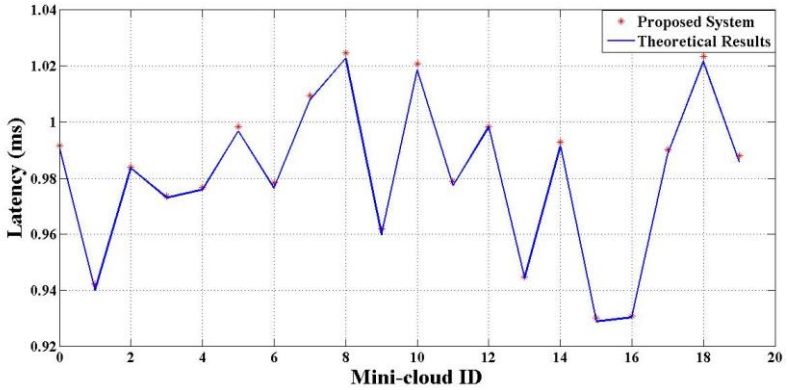

Fig. 6. Latency of Mini-cloud units in case (2).

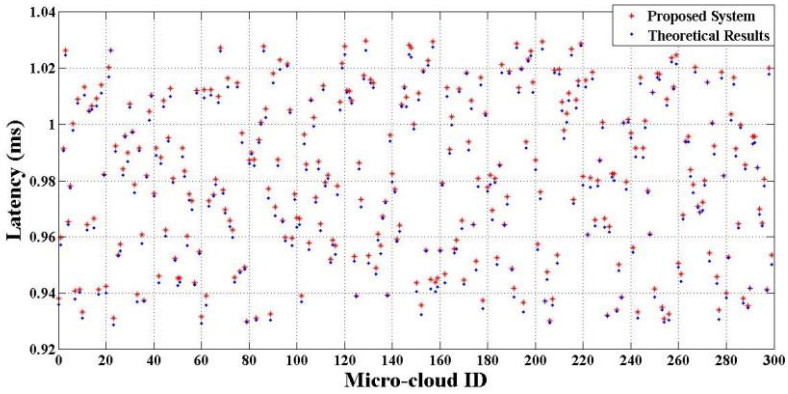

Fig. 7. Latency of Micro-cloud units in case (3).

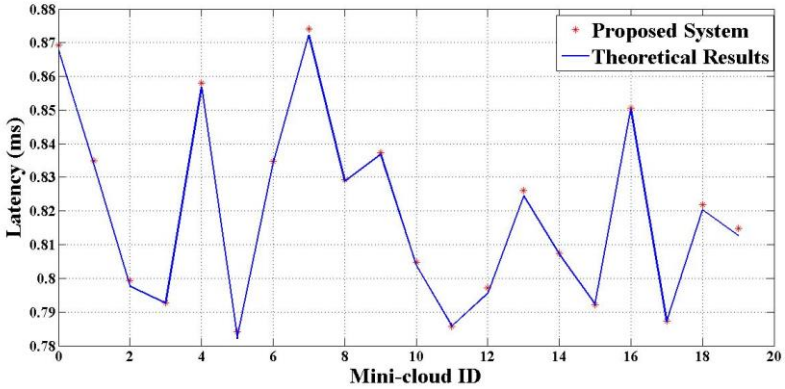

Fig. 8. Latency of Mini-cloud units in case (3).

Table 2. Average delay for Micro- and Mini-clouds.

		Case (1)	Case (2)	Case (3)
Micro-cloud level	Theoretical result	0.698 ms	0.838 ms	0.978 ms
	Simulation result	0.70 ms	0.84 ms	0.99 ms
Mini-cloud level	Theoretical result	1.127 ms	0.978 ms	0.828 ms
	Simulation result	1.13 ms	0.98 ms	0.83 ms

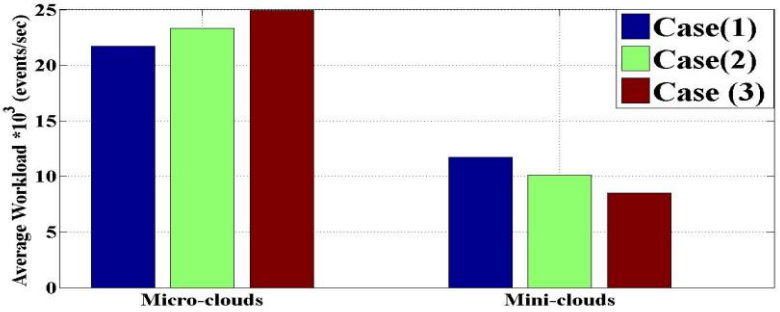

Fig.9. Average workload for each cloud level.

4 Conclusion and future work

One of the efficient ways to reduce the round trip latency of data is to introduce a cloud level in the way between the eNB's cloud and the core network cloud. Multi-level cloud based Tactile Internet system introduces the Micro and Mini-cloud levels before the core network cloud. These two levels, as it was illustrated before, provides

a useful and efficient way to reduce the round trip latency of data and produce away for offloading to reduce the workload delivered to the core network. The paper introduces a mathematical model for the Multi-level cloud based Tactile Internet system that is used to calculate the system latency. Simulation results verify the model and thus it can be used as a valid structure for the Tactile Internet system. The proposed system model can be used to solve the optimization problems of the Tactile Internet in terms of latency and energy efficiency.

Our future vision is to use the mathematical model to optimize the number of first level clouds connected to each Mini-cloud unit to reduce the round trip delay.

Acknowledgement

The publication was financially supported by the Ministry of Education and Science of the Russian Federation (the Agreement number 02.a03.21.0008).

References

1. A. Aijaz, M. Simsek, M. Dohler and G. Fettweis, “Shaping 5G for the Tactile Internet,”2017. *5G Mobile Communications*. Springer International Publishing, 2017. 677-691.
2. ITU-T Technology Watch Report, “The Tactile Internet,” Aug. 2014.
3. K. Gai, M. Qiu, H. Zhao, L. Tao and Z. Zong, “Dynamic energy-aware cloudlet-based mobile cloud computing model for green computing,” Elsevier, Journal of Network and Computer Applications, 59, pp.46-54, Jan.2016.
4. A. Aijaz, M. Dohler, A.H. Aghvami, V. Friderikos and M. Frodigh, “Realizing the tactile internet: haptic communications over next generation 5G cellular networks,” IEEE, Wirel. Comm., 2015.
5. A. Ateya, A. Vybornova, R. Kirichek and A. Koucheryavy, “Multilevel cloud based Tactile Internet system”, IEEE-ICACT2017 international conference, Korea, Feb.2017.
6. A. Nair, M. J. Jacob, and A. Krishnamoorthy, “The multi server M/M/(s, S) queueing inventory system,” Springer US, Ann. Oper. Res., Volume 233, pp 321-333, 2015.
7. Harchol-Balter, Mor., “Performance modeling and design of computer systems: queueing theory in action”, Cambridge University Press, 2013.
8. M. Jia, W. Liang, Z. Xu, and M. Huang, “Cloudlet load balancing in wireless metropolitan area networks”, Computer Communications, IEEE INFOCOM 2016-The 35th Annual IEEE International Conference on. IEEE, 2016.
9. K. Intharawijitr, K. Iida and H. Koga, “Analysis of fog model considering computing and communication latency in 5G cellular networks,” Pervasive Computing and Communication Workshops (PerCom Workshops), 2016 IEEE International Conference, 2016.

10. S. Wang, G. Tu, R. Ganti, T. He, K. Leung, H. Tripp, K. Warr, and M. Zafer, “Mobile micro-cloud: application classification, mapping, and deployment,” in Proc. of Annual Fall Meeting of ITA 2013, Oct. 2013.
11. K. Bahwaireth, L. Tawalbeh, E. Benkhelifa and Y. Jararweh, “Experimental comparison of simulation tools for efficient cloud and mobile cloud computing applications”, Springer EURASIP Journal on Info. Security, June, 2016.