Dynamic Reconfiguration of Computing Resources to Support NaaS Technology

Larysa Globa, Svitlana Sulima, Oleksandr Romanov and Mariia Skulysh
Institute of Telecommunication Systems, Igor Sikorsky Kyiv Polytechnic Institute, Industrialna Lane 2, Kyiv, Ukraine
lgloba@its.kpi.ua, itssulima@gmail.com, a_i_romanov@ukr.net, mskulysh@gmail.com

Keywords: Virtualization, Network Function Virtualization (NFV), Machine Learning (ML), Network as a Service (NaaS), Traffic Management System, Resource Configuration, Network Reconfiguration.

Abstract: Future communication systems and networks, namely smart networks, are becoming the foundation of the human-centric Internet. They provide an energy-efficient and high-performance infrastructure where NGI (Next Generation Internet) and other digital services can be developed and deployed. The development of new services provided to the end user led to the technology of providing a network as a service NaaS (Network as a Service) implementation, which should be an intelligent, decentralized and highly automated network. The architecture of the intelligent network is already software-defined and provides functions that highly exceed the connectivity capabilities: multi-service and mobile edge computing will allow data to be stored and processed locally at the edges of the network to guarantee quick response and efficient use of network resources. But assignment the required amount of computing resources and determining their location remains an incompletely resolved issue. This paper proposes a method of dynamic reconfiguration of computing resources to support NaaS technology. The method allows you to manage the placement and determination of the required amount of network functions resources to optimize the amount of computing resources allocated to the network function in the telecom operator infrastructure, and also binds the performance of the network virtualized function with the amount of resources assigned to it, takes into account the hybridity of the communications environment.

1 INTRODUCTION

Network as a Service (NaaS) enables tenants and cloud users to connect their distributed services across multiple clouds without relying on their own networking services and resources [1]. Today, CSPs implementing NaaS are focused on increasing efficiency and flexibility in business and network operations while improving the service experience for business customers. The main goals are to reduce the complexity of the network and support system, to give customers more control over their services.

NaaS is a framework for building network services that decouples the customer's service request, along with the operational and business support systems (OSS/ BSS) used to create and manage the service, from the network resources required to provide that service. The decoupling occurs through an abstraction layer that hides the details and complexity of the network, which is central to simplifying the entire service lifecycle.

So, if a customer requests a service with a defined bandwidth with certain characteristics between specific points, the CSP provider's OSS/BSS system, as well as network systems and their elements, autonomously figure out how to perform the service, and how much network and computing resources are required to operate services.

To cope with a significant increase in signaling traffic, mobile operator is developing network virtualization and cloud computing technologies to build scalable and flexible mobile networks and offer them as cloud service.

Our research falls into the category of advanced placement algorithms, other placement algorithms range from ILP formulations, such as those described in [2], to evolutionary algorithms, like [3], and heuristic methods.

In the design of virtualized networks, the trade-off between the cost of resource deployment and the effective provisioning of services must be considered [4].

Current descriptions of NaaS services lack determining the location and capacity of reserved computing resources of virtual network functions, taking into account the state of computing and network resources, requirements to the Quality of Service for signaling flows and heterogeneity of the network. Due to the dynamic provision of resources proposed in the paper method can reduce the amount of resources used, implementing calculation of topology and capacity of resources at the same time.

2 PEOPLE-ORIENTED SERVICES

Telecommunication networks are critical and strategic infrastructures that must be protected with appropriate tools. TM Forum considers these main trends that significantly affect the development of modern technologies in various industries and citizens' lives, in particular:

1) Robotic automation, where advances in robotics and artificial intelligence accelerate the introduction of interconnected and autonomous machines in many sectors of human life [5–6].
2) Mass monitoring and remote control, which is necessary for life support processes to optimize productivity and increase their efficiency in the economy [7–8].
3) Autonomous and hyper-connected urban transport on demand [9].
4) Industrial IoT (IIoT) with cloud computing technologies that form industrial networks that must provide productivity, secure and reliable communication in real time in the enterprise where these technologies are implemented.
5) Tactile IoT, which will make human life more comfortable.

All these services can be provided as NaaS layers within one physical network using NFV technology. The challenge is to place these virtual layers on the network effectively.

2.1 Requirements for Parameters of Network Operation

Table 1 shows five abstract dimensions for evaluating the effectiveness of the functioning of services with the disclosure of a set of parameters that characterize the corresponding requirements for the network, namely: bandwidth, time, security, the level of application of artificial intelligence (AI) and the application of ManyNets technology (technology of inter-network interaction).

Table 1: Abstract dimensions with corresponding network requirements [10].

Abstract dimensions	Appropriate network requirements
Capacity	Bandwidth, QoE, QoS, flexibility
Time	Latency, timing, jitter, accuracy, scheduling, coordination, geolocation accuracy
Security	Security, confidentiality, reliability, sustainability, traceability (in the sense of monitoring) of authorized access, lawful interception
The level of application of artificial intelligence	Edge computing, storage, modeling, collection and analytics for network configuration and management
The level of application of inter-network interaction technology (ManyNets)	Addressing, Mobility, Network Interface, Multiple Radio Access Technologies, Heterogeneous Network and Computing Convergence

In the coming years, there is a clear trend towards the maximum increase in the level of automation in all processes and interactions (except for cases that may negatively affect the processing and interaction process).

Future networks will increase the level of requirements for real-time management of network services and the level of automation, which will be crucial for improving the economic efficiency of the work performed.

The emergence of the concept of Network Functions Virtualization (NFV) has opened up new opportunities for telecommunication systems, at the same time, new questions arise, that is, the need for new approaches, models and methods of organizing the process of servicing telecommunication flows in such systems.

The general idea behind this technology is that network functions are implemented as software entities that can run on a standard hardware platform. According to the ETSI specification, a network function is a functional unit within the network infrastructure that has well-defined external interfaces and well-defined functional behavior. Network functions in the LTE EPC network are, for example, MME, HSS, PGW, SGW, which for the NFV case will be deployed on server resources (CPU, memory, NIC).

Service delivery in the telecommunications industry has traditionally been based on network operators deploying static physical equipment that is located at fixed points and has fixed service characteristics.

Thus, the task of effective distribution of computing resources (CPU, storage, bandwidth) of the basic network based on NFV virtualization technology is relevant in order to increase the utilization rate of computing resources during the operation of the core network (LTE EPC) to guarantee the appropriate level of quality of service provision due to the dynamic redistribution of service flows in the network and reconfiguration computing resources of a service.

2.2 Provision of Services by a CSP-Provider in a Multi-Layered NaaS Architecture

Modern NaaS implementations provide clear benefits for CSP-providers and their customers. Customers benefit from a wide choice of data transfer options, taking into account the cost of such transfer, simplification of interaction processes, integration, control and delivery. These NaaS features are described as "smoothing at the edges" between the components needed to create end-to-end services. Customers receive a fully automated process of execution and activation of services. CSP providers gain efficiency and agility benefits from increased operational productivity due to a weaker connection of the network that interacts with the systems that support its operation (Figure 1, Figure 2).

2.3 Determining the Location and Required Capacity of Reserved Virtual Computing Resources

To solve the problem of dynamic reconfiguration of computing resources to support NaaS technology, it is necessary firstly to choose which of the possible data centers should be rented for the subsequent placement of virtualized networks on them, as well as to determine how much processor, memory and network bandwidth resources should be reserved for virtual machines for the worst case.

For this purpose, it is proposed to determine the number of virtual networks to be deployed, and to actually link each network function of the traditional network to the data center and the number of resources allocated to the corresponding virtualized network function, with the objective function of cost minimization [11].

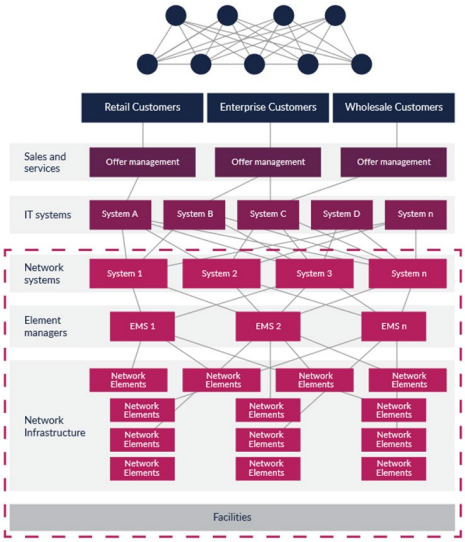

Figure 1: The network side of the services provision by a CSP-provider using a multi-layered NaaS architecture.

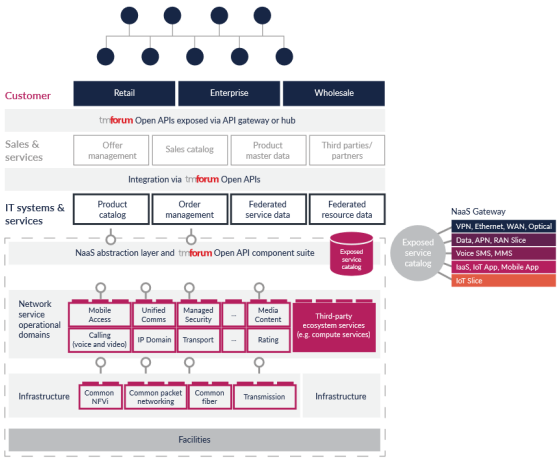

Figure 2: The client side of the services provision by a CSP-provider using a multi-layered NaaS architecture.

We will build a mathematical model of nonlinear optimization with integer variables, which takes into account the available resources on the nodes and their cost, the intensity of the excess service load, and permissible delays as constraints.

The method of solving such a problem is based on the joint location of individual virtualized services of the basic network on the physical network. We assume that the virtual network functions of the mobile core network have the same functionality and interfaces as the network

components of the 3GPP LTE Evolved Packet Core (EPC) architecture. The number of service chains must be determined in advance. A partial case is the consideration of one service chain per cell/eNodeB. Since realistic scenarios for mobile networks are 10000 eNodeBs, the resulting optimization model will be very large and require a long computational time to solve. In this regard, we choose reasonably large clusters of eNodeBs and assume that each of these eNodeB clusters addresses one service chain of the core network.

The case is considered when the provider of telecommunication services already has an existing topology of base stations. It is necessary to define a subset of network nodes where load aggregation units will be placed, which will form requests to one virtualized EPC service. After that, an aggregation node (Traffic Aggregation Point – TAP) is assigned to each base station site.

Let x_i be a binary decision variable that takes the value 1 if a TAP must be placed at point i and 0 otherwise. In addition, we define y_{ji} as a binary variable that takes the value 1 if base station j directs the load to the ith TAP, and 0 otherwise. It is necessary to determine the values of x_i and y_{ji} so as to find the optimal value of the objective function.

Objective function (1) seeks to minimize network delays. The objective function (2) represents the general cost of establishing traffic aggregation nodes and the cost of establishing links between base stations and the corresponding TAPs serving them. Objective function (3) aims to leave more free bandwidth on each physical channel. Residual bandwidth on all channels is maximized, as heavily loaded channels can lead to network congestion, so it is desirable to get a solution where more channel capacity remains.

These optimization goals can be useful for network operators to plan the best deployment strategy.

$$\min_{x_i,y_{ji}} \left(\sum_i \sum_j y_{ji} \cdot L_{ji}\right), \tag{1}$$

$$\min_{x_i,y_{ji}} \left(\sum_i x_i \cdot cost_i + \sum_i \sum_j y_{ji} \cdot costl_{ji}\right), \tag{2}$$

$$\max_{x_i,y_{ji}} \left(\sum_i y_{ji} \cdot (c_{ji} - B_{ji})\right), \tag{3}$$

where L_{ji} is the delay of the communication channel between site j and TAP i; $cost_i$ is a cost that consists of two parts: a fixed initial cost f_i, which is responsible for fixed investments such as space and equipment installation, and additional costs $costN_i$ per unit of processing power installed on a computing node, where d_i is the amount of computing resources used: $cost_i = f_i + costN_i \cdot d_i$;

$costl_{ji}$ is the cost of establishing a connection between site j and TAP i, defined as a linear combination of the initial fixed cost fl_{ji} and the variable part depending on the bandwidth B_{ji} required by the channel and the cost of a unit of bandwidth $costL_j$: $costl_{ji} = fl_{ji} + costL_i \cdot B_{ji}$; c_{ji} – available bandwidth.

It is possible to use a linear combination (4) of expressions (1)-(3) with weighting coefficients a, b, c, which can be used not only to set a greater weighting of one or another component, but also to scale the values of the expressions in order to reduce them to comparable values and have a meaningful composition:

$$\min_{x_i,y_{ji}} \begin{pmatrix} a \cdot \sum_i \sum_j y_{ji} \cdot L_{ji} + b \cdot \left(\sum_i x_i \cdot costN_i + \sum_i \sum_j y_{ji} \cdot costL_{ji} \right. \\ \left. -c \cdot \left(\sum_i y_{ji} \cdot (c_{ji} - B_{ji})\right) \right. \end{pmatrix}. \tag{4}$$

Constraints:

$$\sum_i y_{ji} = 1 \quad \forall j, \tag{5}$$

$$y_{ji} \le x_i \quad \forall j \ \forall i, \tag{6}$$

$$\sum_i x_i \le p, \tag{7}$$

$$\sum_j y_{ji} \cdot d_j \le p_i \quad \forall i, \tag{8}$$

$$\sum_i y_{ji} \cdot (c_{ji} - B_{ji}) \ge 0 \quad \forall j, \tag{9}$$

$$\sum_i y_{ji} \cdot L_{ji} \le T_j \quad \forall j. \tag{10}$$

Constraint (5) ensures that each base station will be attached to one TAP. Constraint (6) ensures that a link is established between base station site j and TAP i only if i has been placed.

Constraint (7) ensures that the maximum number of TAPs does not exceed the budget p, while (8) is a capacity constraint that ensures that the total processing requirements of all base stations assigned to a particular TAP do not exceed the actual installed physical resources p_i. Constraint (9) guarantees the sufficiency of channel resources for establishing channels, and (10) the admissibility of the delay value, i.e. not exceeding the threshold T_j.

The physical network is given in the form of a graph $SN = (N, NE)$, where N is the set of physical nodes and L is the set of channels. Each channel $l = (n_1, n_2) \in NE$, $n_1, n_2 \in N$ has a maximum bandwidth $c(n_1, n_2)$ and each node $n \in N$ is associated with certain resources c_n^i, $i \in R$, where R is the set of types resources (CPU, memory, disk space, network interface). The set of all Traffic Aggregation Points (TAP), i.e. eNodeB clusters, in the network is denoted by $K \subseteq N$. For each node n $\in$ N, $suit_n^{k,j}$ is a binary parameter that indicates whether it is administratively possible to deploy a function of type $j \in V$ on the node, where V is the set of types of network functions, k-th service, where $k \in K$.

The virtual basic mobile network is represented by a set of services (one service per TAP) that are built into the physical network.

The channel bandwidth requirements between two functions, *j1* and *j2*, *(j1,j2)* $\in E$ related to the TAP service $k \in K$ is denoted as $d_k^{(j1,j2)}$. $d_k^{j,i}$ is the amount of computing resource of type *i* allocated for network function *j* of service *k*. $s_{n,i}^{k,j}$ denotes the time of processing a request on resource type *i* of virtual network function *j* of service *k* by one resource unit of node *n*. The requirements for the allowable processing time of the request by the network function *j*, which is related to the service *k*, are denoted as P_k^j. T_k is the maximum delay for $k \in K$, $L(n_1,n_2)$ is the network delay for the channel $(n_1,n_2) \in NE$.

The goal of the optimization is to find the location of the virtualized services of the core network (that is, the placement of network functions and the distribution of resources, as well as the determination of traffic transmission paths between them), so as to minimize the conditional costs of the occupied resources of channels and nodes in the physical network, while satisfying the load requirements $\lambda_{k,j}$. We formulate the objective function (expression (11)) in the form of a linear combination of two cost expressions: the occupied amount of computing nodes resources, where the conditional cost of a unit of resource *i* on node *n* is denoted as *costN(i,n)*, and the occupied bandwidth of channels, where $costL(n_1,n_2)$ is the conditional cost of a physical channel bandwidth unit $(n_1,n_2) \in NE$.

The following expressions (11)-(20) represent the formulation of the nonlinear programming optimization problem. Boolean variables $x_n^{k,j}$ indicate whether network function *j* associated with service *k* is located on physical node *n*. For *j*=TAP, $x_n^{k,TAP}$ are not variables, but input parameters that indicate where TAP *k* is located, i.e.

$$x_n^{k,TAP} = \begin{cases} 1 \text{ if } k=n, \\ 0 \text{ otherwise} \end{cases}$$

Similarly, the boolean variables $f_{(n1,n2)}^{k,(j1,j2)}$ indicate whether the physical channel $(n1,n2) \in NE$ is used for the path between *j1* and *j2* for service *k*.

Expression (12) ensures that only one network function of each type is placed for each TAR/service. Expression (13) guarantees that the placement of resources is carried out on physical nodes that have the administrative ability to place the corresponding network functions. Expressions (14), (15) and (16) represent limitations on the available resources of physical nodes and channels. Expression (17) is a flow conservation constraint for all paths in the physical network. Expression (18) guarantees that the variables in the network function placement and path mapping problem are Boolean.

$$\min_{f_{(n_1,n_2)}^{k,(j_1,j_2)}, d_t^{j,i}} \left(\begin{array}{l} \sum_{k\in K}\sum_{i\in V}\sum_{n\in N}\sum_{i\in R} x_n^{k,j}\cdot d_k^{j,i}\cdot costN(i,n) + \\ + \sum_{(n_1,n_2)\in L} costL(n_1,n_2)\cdot \sum_{k\in K}\sum_{(j_1,j_2)\in E} f_{(n_1,n_2)}^{k,(j_1,j_2)}\cdot d_k^{(j_1,j_2)} \end{array} \right) \quad (11)$$

$$Constraints \sum_{n\in N} x_n^{k,j} = 1 \quad \forall k\in K, j\in V \quad (12)$$

$$x_n^{k,j} \le suit_n^{k,j} \quad \forall k\in K, j\in V, n\in N \quad (13)$$

$$\sum_{(w,n)\in NE}\sum_{k\in K}\sum_{(j_1,j_2)\in E} f_{(w,n)}^{k,(j_1,j_2)}\cdot d_k^{(j_1,j_2)} \le c_n^{bdw} \quad \forall n\in N \quad (14)$$

$$\sum_{k\in K}\sum_{j\in V} x_n^{k,j}\cdot d_k^{j,i} \le c_n^{i} \quad \forall n\in N, i\in\{R\backslash bdw\} \quad (15)$$

$$\sum_{k\in K}\sum_{(j_1,j_2)\in E} f_{(n_1,n_2)}^{k,(j_1,j_2)}\cdot d_k^{(j_1,j_2)} \le c(n_1,n_2) \quad \forall (n_1,n_2)\in NE \quad (16)$$

$$\sum_{(w,n)\in NE} f_{(w,n)}^{k,(j_1,j_2)} - \sum_{(n,w)} f_{(n,w)}^{k,(j_1,j_2)} = x_n^{k,j_1} - x_n^{k,j_2} \quad \forall k\in K, n\in N, (j_1,j_2)\in E \quad (17)$$

$$x_n^{k,j}, f_{(n_1,n_2)}^{k,(j_1,j_2)} \in \{0,1\} \quad \forall k\in K, j\in V, n\in N, (j_1,j_2)\in E, (n_1,n_2)\in NE \quad (18)$$

$$\sum_{(j_1,j_2)\in E}\sum_{(n_1,n_2)\in L} f_{(n_1,n_2)}^{k,(j_1,j_2)}\cdot L(n_1,n_2) \le T_k \quad \forall k\in K \quad (19)$$

$$\sum_{n\in N} x_n^{k,j}\sum_{i\in R}\left(\frac{1}{\frac{j,i}{d_k}\lambda^{k,j}}\right) \le P_k^j \quad \forall t\in T, j\in V \quad (20)$$

To limit channel delays, the delay constraint shown in expression (19) is also added. And in order to take into account the required performance of the virtual network function in the model, restrictions on the value of the application processing time defined in expression (20) are necessary.

The problem (11)-(20) is supposed to be solved offline at the initial stage. According to the solution, each network function is reserved a certain amount of resources of the virtual network function, based on the assessment of its greatest need for resources; the instantaneous needs of various network functions are dynamically met by activating the necessary configuration of virtual machines at runtime in such a way as to satisfy the guarantees provided for each network function.

The solution to the optimization problem can be found using a real coded genetic algorithm MI-LXPM.

2.4 Evaluation of the Method of Computing Resources Dynamic Reconfiguration

The cost of NaaS deployment consists of the cost of launching and using a server, using server resources, using communication channels and using migration resources. Quantitative and qualitative analysis (Figure 3) of the proposed method showed a reduction of costs associated with reserved resources by up to 15%, which contributes to increasing the efficiency of load maintenance and saving computing resources.

Thus, it is possible to draw a conclusion about the effectiveness of the proposed method, which

allows you to link the performance of the network virtualized function with the amount of resources allocated to it, as well as take into account the hybridity of the telecommunications environment. The method can be applied dynamically when managing the network functions deployment in a hybrid hardware environment to minimize the costs for the communication operator and improve the quality of service for subscribers.

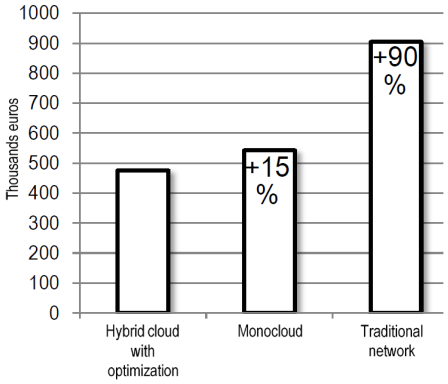

Figure 3: Costs of system resources using the NFV placement method and not using it.

3 CONCLUSIONS

This article considers: the classification of services proposed by TM Forum as the main factors that determined the emergence of NaaS technology, the level of requirements for managing network services in real time and the level of automation, which will be decisive for increasing the economic efficiency of providing modern services to end users.

A method of managing placement and determining the required amount of resources of network functions is proposed to optimize the amount of computing resources allocated to a network function in a telecommunications operator's network.

The proposed method allows you to bind the performance of the network virtualized function with the amount of resources allocated to it, and also takes into account the hybridity of the telecommunications environment.

The method can be applied in managing the deployment of network functions in a hybrid hardware environment to minimize the costs for the communication operator and improve the quality of service for subscribers.

Future research should be related to the development of technology for implementing the proposed method of dynamic reconfiguration of computing resources into the system architecture of the CSP-provider and its adaptation in real conditions..

REFERENCES

[1] Jerbi et al., "Enabling Multi-Provider Cloud Network Service Bundling," 2022 IEEE International Conference on Web Services (ICWS), Barcelona, Spain, 2022, pp. 405-414, doi: 10.1109/ICWS55610.2022.00067.

[2] M. C. Luizelli, L. R. Bays, L. S. Buriol, M. P. Barcellos, and L. P. Gaspary, "Piecing together the NFV provisioning puzzle: Efficient placement and chaining of virtual network functions," 2015 IFIP/IEEE International Symposium on Integrated Network Management (IM), Ottawa, ON, Canada, 2015, pp. 98-106, doi: 10.1109/INM.2015.7140281.

[3] P. T. A. Quang, Y. Hadjadj-Aoul, and A. Outtagarts, "A Deep Reinforcement Learning Approach for VNF Forwarding Graph Embedding," in IEEE Transactions on Network and Service Management, vol. 16, no. 4, pp. 1318-1331, Dec. 2019, doi: 10.1109/TNSM.2019.2947905.

[4] M. Masoumi et al., "Dynamic Online VNF Placement with Different Protection Schemes in a MEC Environment," 2022 32nd International Telecommunication Networks and Applications Conference (ITNAC), Wellington, New Zealand, 2022, pp. 1-6, doi: 10.1109/ITNAC55475.2022.9998347.

[5] R. Hussain and S. Zeadally, "Autonomous Cars: Research Results, Issues, and Future Challenges," in IEEE Communications Surveys & Tutorials, vol. 21, no. 2, pp. 1275-1313, Secondquarter 2019.

[6] S. Iranmanesh, R. Raad, M. S. Raheel, F. Tubbal, and T. Jan, "Novel DTN Mobility- Driven Routing in Autonomous Drone Logistics Networks," in IEEE Access, vol. 8, pp. 13661-13673, 2020.

[7] N. Mhaisen, O. Abazeed, Y. A. Hariri, A. Alsalemi, and O. Halabi, "Self-Powered IoT- Enabled Water Monitoring System," 2018 International Conference on Computer and Applications (ICCA), Beirut, 2018, pp. 41-45.

[8] H. T. Yew, M. F. Ng, S. Z. Ping, S. K. Chung, A. Chekima, and J. A. Dargham, "IoT Based Real-Time Remote Patient Monitoring System," 2020 16th IEEE International Colloquium on Signal Processing & Its Applications (CSPA), Langkawi, Malaysia, 2020, pp. 176-179.

[9] H. Dia, “The real-time city: Unlocking the potential of smart mobility”, In Proceedings of the 38th Australasian Transport Research Forum (ATRF 2016), Melbourne, Australia, 16-18 November 2016.

[10] C. Babb, C. Curtis, S. McLeod, Sam, “The Rise of Shared Work Spaces: A Disruption to Urban Planning Policy?”, 2018, Urban Policy and Research, vol. 36., pp. 1-17, doi: 10.1080/08111146.2018.1476230.

[11] S.V. Sulima “Reconfiguration methods of the computing resources for the core network based on virtualization technology,” qualification scientific paper, manuscript.