





Efficient Slice Allocation Method using Cluster Technology in Fifth-Generation Core Networks

Sang-Myeon Park  and Young-Song Mun , *Member, KIICE*

Department of Computer Science and Engineering, Soongsil University, Seoul 06978, Korea

Abstract

The explosive growth of data traffic and services has created cost challenges for networks. Studies have attempted to effectively apply network slicing in fifth generation networks to provide high speed, low latency, and various compatible services. However, in network slicing using mixed-integer linear programming, the operation count increases exponentially with the number of physical servers and virtual network functions (VNFs) to be allocated. Therefore, we propose an efficient slice allocation method based on cluster technology, comprising the following three steps: i) clustering physical servers; ii) selecting an appropriate cluster to allocate a VNF; iii) selecting an appropriate physical server for VNF allocation. Solver runtimes of the existing and proposed methods are compared, under similar settings, with respect to intra-slice isolation. The results show that solver runtime decreases, by approximately 30% on average, with an increase in the number of physical servers within the cluster in the presence of intra-slice isolation.

Index Terms: 5G Core Networks, Cluster Technology, Slice Allocation, Network Functions, Virtual Network Functions

I. INTRODUCTION

The smart revolution has led to the widespread use of mobile-based high-definition video and audio multimedia services through smart devices, and the emergence of the Internet of Things, a core technology in the fourth industrial revolution. These have induced an exponential growth of mobile network traffic owing to the deployment of multiple small sensors that provide a wide variety of information, such as on temperature and humidity [1]. Such an explosive growth is detrimental, leading to loss of efficiency and overloaded existing networks. These challenges prevent the existing networks from achieving ultra-low latency, ultra-high speed, and hyperconnectivity, which are required by next generation technologies, eventually prompting the emergence of fifth generation (5G) networks.

The 5G networks [2-3] are expected to efficiently handle the

tremendous growth in data traffic caused by a wide variety of devices. Furthermore, with 5G networks, which can offer a peak data rate of 20 Gbps and a latency of 0.001 ms, innovative services using various devices are expected to emerge in different fields, apart from smart device-enabled services. Moreover, 5G networks have attracted attention because effective service customization can be achieved through software-defined networking (SDN) [4] and network function virtualization (NFV) [5]. Unlike the existing networks, which only serve conventional smart devices, 5G networks are expected to have broader service coverage and include a wider array of devices; thus, they must establish network functions that are suitable for the corresponding services. However, creating different network functions not only incurs high costs but also lowers network efficiency. Consequently, this has drawn attention toward network slicing with SDN and NFV. Network slicing is a technology that utilizes network virtualization to


Received 28 August 2019, Revised 25 September 2019, Accepted 26 September 2019

*Corresponding Author Young-Song Mun (E-mail: mun@ssu.ac.kr, Tel: +82-2-820-0676)

Department of Computer science and Engineering, Soongsil University, Seoul 06978, Korea.

Open Access <https://doi.org/10.6109/jicce.2019.17.3.185>

print ISSN: 2234-8255 online ISSN: 2234-8883

 This is an Open Access article distributed under the terms of the Creative Commons Attribution Non-Commercial License (<http://creativecommons.org/licenses/by-nc/3.0/>) which permits unrestricted non-commercial use, distribution, and reproduction in any medium, provided the original work is properly cited.

Copyright © The Korea Institute of Information and Communication Engineering

divide a physical network into multiple virtual networks and allocate a network function (NF) suitable for the required service, aiming at providing appropriate services [6]. Network slicing helps in reducing capital expenditures. This can be attributed to its cost-effective network architecture, ability to provide prompt services without constructing a new physical network when introducing a new service, and ability to efficiently utilize physical network resources. Several studies have been conducted to achieve efficient virtual network function (VNF) allocation through network slicing [7-10]. Moreover, Dietrich et al. [11] and Sattar et al. [12] proposed an efficient network slicing method based on mixed integer linear programming (MILP). Dietrich et al. proposed an approach focusing on effective NFV allocation in long-term evolution (LTE) core networks, and Sattar et al. adjusted this proposed method to fit the 5G core networks. However, in both the methods, the operation count increases exponentially with an increase in the number of physical servers and NFs/VNFs to be allocated when selecting a physical server based on the MILP formulation.

Therefore, this paper proposes an efficient method of allocating network slices using cluster technology. Cluster technology employed during server selection helps in reducing the number of physical servers that are calculated using the MILP formulation. Firstly, the physical servers are clustered within the constraints so that an appropriate cluster can be located when receiving a request for VNF allocation. Subsequently, MILP formulation is performed on the physical servers within the previously identified clusters to select an appropriate physical server for VNF allocation.

The rest of the paper is organized as follows: Section II describes the system model using cluster technology and its application. Section III compares the measured solver run-times between the existing and proposed methods in a 5G core network to demonstrate the efficiency of the proposed method. Section IV presents the conclusions.

II. SYSTEM MODEL AND METHODS

This paper extends the works of Dietrich et al. and Sattar et al., and proposes a novel method to reduce the processing time for slice allocation by using cluster technology in 5G core networks. The proposed method is designed as follows:

- 1) Cluster technology is applied to the physical servers.
- 2) An appropriate cluster is selected using the proposed formulation.
- 3) The most appropriate server within the chosen clusters is selected.

A. Cluster Configuration

To apply the proposed method in 5G core networks, firstly,

cluster technology must be applied to the physical servers. A cluster comprises a cluster leader and members, where the cluster leader satisfies the following conditions:

- The cluster leader and members are well connected.
- The maximum number of hops from the cluster leader to any member is two.
- The cluster leader can store the status information of the members.

The hop limit constraint is applied because distance affects physical link delay. If the distance exceeds two hops, the physical server might exit its home subnet. By limiting the maximum hop distance to two, the impact on the physical link delay of other clusters can be minimized. Additionally, the cluster leader identifies cluster availability for VNF allocation based on member status information, whenever a slice request for a new VNF occurs.

B. Cluster Selection

Once cluster configuration is completed for all the physical servers, each cluster leader must determine the average CPU utilization (U^n) for all the servers (V_{cs}^n) present in the n-th cluster. U^n denotes the average CPU utilization for V_{cs}^n and is obtained using Equation (1).

$$U^n = \frac{\text{Avg} \left(\sum_{j \in V_{cs}^n} r_j \right)}{r_{j,\max}} \tag{1}$$

According to Dietrich et al. and Sattar et al., the number of VNFs that can be allocated to a single physical server is determined by intra-slice isolation (K_{rel}). Before assigning VNF_i , a check needs to be performed to see if it can be actually assigned to server m . Equation (2) provides the ratio of servers in the n-th cluster that can allocate a VNF_i . The parameter γ_m^i prevents the mapping of infeasible $VNF/server$ combinations ($\gamma_m^i \in \{1, \infty\}$).

$$P_\gamma = \frac{\text{Number of } (\gamma_m^i = 1)}{nV_{cs}^n}, \quad \forall m \in V_{cs}^n \tag{2}$$

Thus, when assigning a new VNF_i , the cluster leader acquires the status value (C_n) of the cluster, which is obtained from the product of the computing demand value (g^i), average utilization for servers within the cluster, and percentage of servers without VNF_i allocation (P_γ). C_n can be obtained using Equation (3). Each time a new VNF allocation request is generated, the requested VNF is allocated to the cluster with the smallest C_n .

$$C_n = (1 - U^n) g^i P_\gamma \tag{3}$$

C. Server Selection

Once the cluster for a new VNF allocation has been chosen, the most appropriate server within the cluster is selected. A suitable server for VNF allocation can be selected using the MILP formulation proposed by Sattar *et al.* as follows:

Minimum

$$\sum_{i \in V_F} \sum_{u \in V_{CS}} \left(1 - \frac{r_u}{r_{u,max}}\right) g^i x_u^i \gamma_u^i + \sum_{(i,j) \in E_F} \sum_{(u,v) \in E_{CS}} L_{uv} f_{uv}^{ij}, \quad (4)$$

subject to:

$$\sum_{i \in V_F} x_u^i \leq K_{rel}, \quad \forall u \in V_{CS}, K_{rel} = 1, 2, 3, \dots, \quad (5)$$

$$\sum_{i \in V_F} g^i \leq \sum_{u \in V_{CS}} r_u, \quad (6)$$

$$\sum_{(i,j) \in E_F} g^{ij} \leq \sum_{(u,v) \in E_{CS}} r_{uv}. \quad (7)$$

Unlike the method proposed by Sattar *et al.*, the target servers of the objective function shown in Equation (4) and constraints shown in Equations (5)-(7) are limited to the clusters selected in the 2-B (Cluster Selection) process. The objective function from Equation (4) is allocated to the server with the least utilization and delay among the servers in the selected cluster. The formulation proposed in this paper differs from that of previous studies in terms of the scope of operation in the first term. Dietrich *et al.* and Sattar *et al.* established the MILP formulation with all the physical servers as target servers to identify the least utilized server. In contrast, this study uses a cluster as a unit for the target servers, which in turn, decreases the operation count for the MILP formulation.

Moreover, in the second term, the formula for L_{uv} proposed by Sattar *et al.*, which considers the minimum delay path, is modified. $L_{uv,init}$ (initial delay assigned to the link $(u, v) \in E_{CS}$) value is identical owing to the hop limit (two hops) set during cluster configuration. Therefore, L_{uv} can be computed using Equation (8), where $L_{uv,init}$ is removed from the existing L_{uv} formula.

$$L_{uv} = \beta \cdot \left(1 - \frac{r_{uv}}{r_{uv,max}}\right),$$

$$\beta = 2.5ms, \forall (u, v) \in E_{CS}. \quad (8)$$

Similar to the existing method, constraint in Equation (5) represents the maximum number of NFs (K_{rel}) that can be allocated to a single server in a cluster through intra-slice isolation. The constraints in Equations (6) and (7) emphasize that the remaining computing and bandwidth capacity of the servers and links in the cluster are greater than the comput-

Table 1. Notations used in system model and method

Symbol	Description
U^n	Average of server utilization in the n-th cluster
C_n	Status value of the n-th cluster
P_γ	Percentage of servers without VNF_i allocation in cluster
nV_{CS}^n	Number of servers present in the n-th cluster
V_{CS}^n	A group of servers present in the n-th cluster
V_F	The set of VNF elements to be allocated
V_{CS}	The set of servers in the selected cluster
E_F	The set of edges between the allocated VNF elements
E_{CS}	The set of edges between the servers in the selected cluster
x_n^i	Assignment of VNF_i to server n
γ_n^i	Feasibility indicator of the mapping of VNF_i to server n
g^i	Computing demand of VNF_i in GHz
g^{ij}	Bandwidth demand of edge (i, j) in Mbps
r_u	Residual capacity of server u in GHz
$r_{u,max}$	Maximum capacity of server u in GHz
r_{uv}	Residual capacity of link (u, v) in Mbps
$r_{uv,max}$	Maximum capacity of link (u, v) in Mbps
f_{uv}^{ij}	Amount of bandwidth assigned to link (u, v) for edge (i, j) in Mbps
L_{uv}	Physical link delay

ing demand and bandwidth capacity, respectively, when allocating VNF_i . Table 1 summarizes the notations used in the proposed method.

III. SIMULATION AND RESULTS

To evaluate the proposed method, a 5G core network with slicing requests was implemented in Java with i5-4670 CPU and 8.0 GB RAM, and CPLEX 12.9.0 was used as the MILP solver. Further, the MILP formulation by Sattar *et al.*, which considers intra-slice isolation, was compared with the proposed method and analyzed under simulation settings. Fig. 1 depicts the simulation architecture, and Table 2 lists the simulation parameters.

In the simulation architecture, the link bandwidth was set to 20 Gbps between data center switches, 10 Gbps between aggregation switches, 1 Gbps between aggregation and edge switches, and 250 Mbps between an edge switch and a phys-

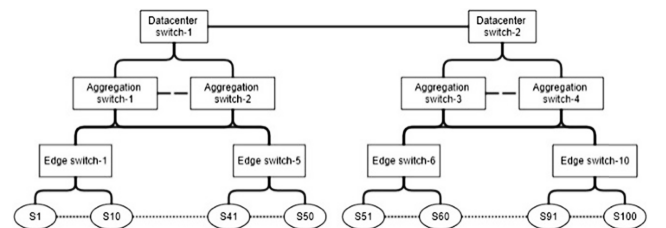


Fig. 1. Simulation architecture

Table 2. Simulation parameters

Parameters	Values
CPU capacity/server	12.0 GHz
Total servers	100
VNF/slice	10
Intra-slice isolation	1-10
Total slice request	100
VNF CPU requests/slice	0.5-2.0 GHz
Bandwidth requests/slice	30-70 Mbps

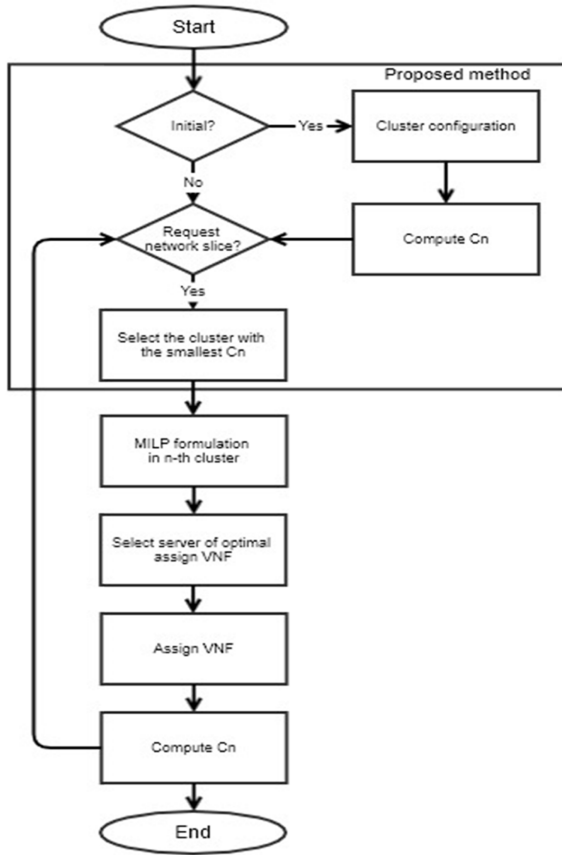


Fig. 2. Simulation algorithm

ical server.

Additionally, Fig. 2 depicts the simulation algorithm. Prior to optimal server selection, cluster technology is applied to 100 physical servers and the cluster status value (C_n) for each cluster is computed. Subsequently, an optimal server, with the smallest C_n , is selected by using the MILP formulation. Finally, we assign the VNF to the optimal server.

The simulation was divided into two to measure the solver runtimes. In the first simulation, intra-slice isolation was not considered ($K_{rel}=1$). Here, the solver runtimes using the proposed method were measured while varying the number of clusters as 1, 2, 4, 10, and 20. The measured runtimes were

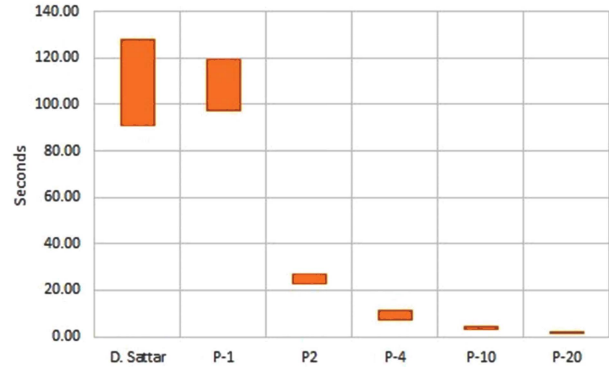


Fig. 3. Result of the first simulation without intra-slice isolation

Table 3. Result of the first simulation without intra-slice isolation

	Minimum	Maximum	Average
D. Sattar	90.98	127.71	109.95
P-1	97.14	119.15	109.51
P-2	23.11	26.84	25.33
P-4	7.15	11.29	9.31
P-10	3.32	4.17	3.78
P-20	1.41	2.08	1.84

compared to those of the existing method. The results showed that the average solver runtime measured using the existing method was approximately 109.95 s, which is similar than that using the proposed method for one cluster (109.51 s). When the number of clusters was 2 or more, the average solver runtime of the proposed method sharply decreased compared to that of the existing method, requiring approximately 25.33 s, 9.31 s, 3.78 s, and 1.84 s for 2, 4, 10, and 20 cluster members, respectively. Fig. 3 and Table 3 show the results of the first simulation.

The second simulation considered intra-slice isolation and allowed the level of isolation to vary from 1 to 10 ($K_{rel} = 1-10$). It compared the measured solver runtimes of the existing and proposed methods while varying the number of clusters (1, 2, 4, 10, and 20) in the proposed method. The results showed that the solver runtime measured by the existing method differed from that of the proposed method without the application of cluster technology (number of clusters = 1) by approximately 20 – 50% depending on the intra-slice isolation. This indicated that the proposed method had longer runtimes than the existing method owing to the additional operations generated by cluster selection, even when the number of clusters was 1. Furthermore, as the number of clusters increased, the operation count for MILP formulation decreased, yielding shorter solver runtimes than the existing method. However, in the proposed method that relies on cluster technology, a lower extent of solver runtime reduction due to intra-slice isolation was observed. In other words,

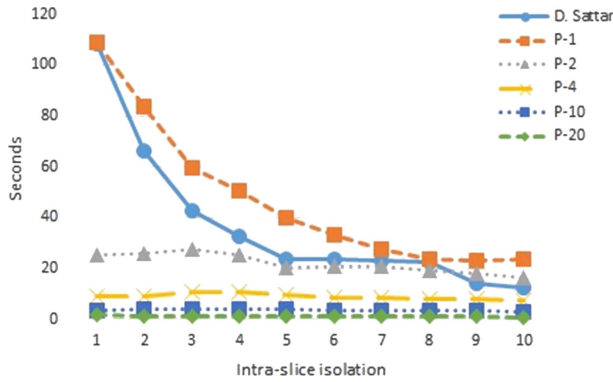


Fig. 4. Result of the second simulation with intra-slice isolation

if the number of clusters exceeded 2 and the intra-slice isolation level was lower than 4 ($K_{rel} \leq 4$), then, the solver runtime hardly decreased or increased. In contrast, if the number of clusters exceeded 2 and the intra-slice isolation level exceeded 5 ($K_{rel} \geq 5$), then, an approximately 40% decrease in solver runtime was observed. Fig. 4 and Table 4 show the results of the second simulation.

IV. DISCUSSION AND CONCLUSIONS

Handling the tremendous growth of data traffic due to various devices incurs high costs for existing networks. Moreover, with the advent of the fourth industrial revolution, networks are expected to provide services to a wider array of devices. This draws focus toward 5G networks, which can handle the problems that remain unsolved for existing networks. The 5G networks can offer effective management solutions for explosive data traffic growth based on SDN and NFV. Recently, several extensive studies on 5G network slicing have been conducted, focusing on its benefits such as ability to handle higher volumes of data traffic compared to existing networks by utilizing virtualization technologies, and ability to provide efficient services without incurring additional costs through customized VNF configurations customized for the requested service. A study on network slicing using the existing MILP formulation indicated that

the operation count increases exponentially with an increase in the number of physical servers and requested VNFs for allocation. This paper proposes an efficient network slicing allocation method using cluster technology. The proposed method shows a reduction in the number of physical servers employed in the MILP formulation, consequently decreasing the total operation count. The proposed method comprises the following three steps: cluster configuration, cluster selection, and physical server selection. To evaluate the proposed method, the measured solver runtimes of the existing and proposed methods, in an identical setting, were compared depending on the presence or absence of intra-slice isolation. The results showed that the solver runtime measured by the proposed method decreased by approximately 30% compared to the existing method. Such differences are due to the reduced number of physical servers in the proposed method. When a VNF makes an allocation request, the existing method performs operations based on the number of physical servers. In contrast, the proposed method first selects a cluster suitable for allocation based on the status information on physical servers and consequently decreases the number of physical servers required to perform the operation. Moreover, if the intra-slice isolation level exceeds 5, the number of physical servers decreases; thus, reducing the time required for an appropriate physical server selection.

REFERENCES

- [1] Q. Ye and W. Zhuang, "Distributed and adaptive medium access control for internet-of-things-enabled mobile networks," *IEEE Internet of Things Journal*, vol. 4, no. 2, pp. 446-460, 2017. DOI: 10.1109/JIOT.2016.2566659.
- [2] A. Gupta and R. K. Jha, "A survey of 5G network: architecture and emerging technologies," *IEEE Access*, vol. 3, pp. 1206-1232, 2015. DOI:10.1109/ACCESS.2015.2461602.
- [3] H. Zhang, S. Vrzic, G. Senarath, N.-D.Dao, H.Farmanbar, J. Rao, C. Peng, and H. Zhuang, "5G wireless network: MyNET and SONAC," *IEEE Network*, vol. 29, no. 4, pp. 14-23, 2015. DOI: 10.1109/MNET.2015.7166186.
- [4] D. Kreutz, F. M. V. Ramos, P. E. Verissimo, C. E. Rothenberg, S. Azodolmolky, and S. Uhlig, "Software-Defined Networking: A comprehensive Survey," *Proceedings of the IEEE*, vol. 103, no. 1, pp. 14-76, 2015. DOI: 10.1109/JPROC.2014.2371999.

Table 4. Result of the second simulation with intra-slice isolation

	Intra-slice isolation (K_{rel})									
	1	2	3	4	5	6	7	8	9	10
D. Sattar	108.41	66.51	42.52	32.78	23.75	23.48	23.28	22.86	14.44	12.55
P-1	108.89	83.46	59.46	50.64	39.98	32.95	27.77	23.91	23.00	23.60
P-2	25.25	25.69	27.59	25.12	20.61	20.71	20.71	23.91	18.02	16.50
P-4	9.11	9.10	10.59	10.81	9.75	8.82	8.47	9.38	7.98	7.59
P-10	3.75	3.82	4.18	4.05	3.97	3.69	3.41	3.38	3.25	3.10
P-20	1.82	1.56	1.56	1.45	1.22	1.20	1.16	1.12	1.01	1.00

- [5] R. Mijumbi, J. Serrat, J. L. Gorricho, N. Bouten, F. D. Turck, and R. Boutaba, "Network function virtualization: State-of-the-art and research challenges," *IEEE Communications Surveys Tutorials*, vol. 18, no. 1, pp. 236-262, 2016. DOI: 10.1109/COMST.2015.2477041.
- [6] S. O. Oladejo and O. E. Falowo, "5G network slicing: A multi-tenancy scenario," in *Proceeding of 2017 Global Wireless Summit*, 2017. DOI: 10.1109/GWS.2017.8300476.
- [7] S. Kianpishah and R. H. Glitho, "Cost-efficient server provisioning for deadline-constrained VNFs Chains: A parallel VNF processing approach," in *Proceeding of 2019 16th IEEE Annual Consumer Communications & Networking Conference*, 2019. DOI: 10.1109/CCNC.2019.8651799.
- [8] S. Ahvar, H. P. Phyu, S. M. Buddhacharya, E. Ahvar, N. Crespi, and R. Glitho, "CCVP: Cost-efficient centrality-based VNF placement and chaining algorithm for network service provisioning," in *Proceeding of 2017 IEEE Conference on Network Softwarization*, 2017. DOI: 10.1109/NETSOFT.2017.8004104.
- [9] H. Wei, Z. Zhang, and B. Fan, "Network slice access selection scheme in 5G," in *Proceeding of 2017 IEEE 2nd Information Technology, Networking, Electronic and Automation Control Conference*, 2017. DOI: 10.1109/ITNEC.2017.8284751.
- [10] Y. I. Choi and N. I. Park, "Slice architecture for 5G core network," in *Proceeding of 2017 ninth International Conference on Ubiquitous and Future Networks*, 2017. DOI: 10.1109/ICUFN.2017.7993854.
- [11] D. Dietrich, C. Papagianni, P. Papadimitriou, and J. S. Baras, "Network function placement on virtualized cellular cores," in *Proceeding of 2017 9th International Conference on Communication Systems and Networks*, 2017. DOI: 10.1109/COMSNETS.2017.7945385.
- [12] D. Sattar and A. Matrawy, "Optimal Slice Allocation in 5G Core Networks," *IEEE Networking Letters*, vol. 1, no. 2, pp. 48-51, 2019. DOI: 10.1109/LNET.2019.2908351.



Sang-Myeon Park

Received the B.S. degree in Computer Science and Engineering from Soongsil University in 2014, and received the M.S. degree in 2016, and Ph.D. in 2019 from the Department of Computer Science and Engineering at Soongsil University. His research interests include 5G core network, edge network, cloud computing, the internet of things, and blockchain



Young-Song Mun

Received a Ph.D. degree from the University of Texas (Arlington), in 1993. Since 1994, he has been serving as a professor in the Department of Computer Science and Engineering at Soongsil University, Seoul, Korea. His research interests include 5G core network, edge network, cloud computing, the internet of things, and blockchain.