J. Internet Comput. Serv.
ISSN 1598-0170 (Print) / ISSN 2287-1136 (Online)
http://www.jics.or.kr

멀티 액세스 엣지 컴퓨팅을 위한 Mobility-Aware Service Migration (MASM) 알고리즘☆

Mobility-Aware Service Migration (MASM) Algorithms for Multi-Access Edge Computing

하 지 크[1] 리 덕 타이[2] 김 문 성[3*] 추 현 승[1*]
Haziq Hamzah Duc-Tai Le Moonseong Kim Hyunseung Choo

요 약

5G 목표 중 하나인 초신뢰성 저지연 통신에 도달하기 위해 멀티액세스 엣지 컴퓨팅 패러다임이 탄생했다. 이 패러다임은 클라우드 컴퓨팅 기술을 네트워크 엣지에 더 가깝게 하며 서비스 지연 시간을 줄이기 위해서는 네트워크 엣지에 있는 여러 Edge Cloud에서 서비스 호스팅된다. 모바일 사용자의 경우 서비스 품질 유지를 위해 서비스를 가장 적합한 Edge Cloud로 마이그레이션하는 것은 중요하고 고이동성 시나리오에서는 서비스 마이그레이션 문제가 더욱 복잡해진다. 고정 이동 경로에서 사용자 이동성과 Edge Cloud 선택에 대한 어떤 영향을 미치는 건지 관찰하는 것이 이 연구의 목표다. Mobility-Aware Service Migration (MASM)은 고이동성 시나리오 동안 라우팅 비용과 서비스 마이그레이션 비용이라는 두 가지 주요 매개변수를 기반으로 서비스 마이그레이션을 최적화하기 위해 제안된다. 제안된 알고리즘을 기존의 그리디 알고리즘과 비교하여 평가한다.

☞ 주제어 : 서비스 마이그레이션, 다중 액세스 엣지 컴퓨팅, 고이동성, 엣지 클라우드, QoS

ABSTRACT

In order to reach Ultra-Reliable Low-Latency communication, one of 5G aims, Multi-access Edge Computing paradigm was born. The idea of this paradigm is to bring cloud computing technologies closer to the network edge. User services are hosted in multiple Edge Clouds, deployed at the edge of the network distributedly, to reduce the service latency. For mobile users, migrating their services to the most proper Edge Clouds for maintaining a Quality of Service is a non-convex problem. The service migration problem becomes more complex in high mobility scenarios. The goal of the study is to observe how user mobility affects the selection of Edge Cloud during a fixed mobility path. Mobility-Aware Service Migration (MASM) is proposed to optimize service migration based on two main parameters: routing cost and service migration cost, during a high mobility scenario. The performance of the proposed algorithm is compared with an existing greedy algorithm.

☞ keyword : Service migration, Multi-Access Edge Computing, High mobility, Edge Cloud, QoS

1. Introduction

Multi-Access Edge Computing (MEC) [1], a distributed cloud computing paradigm at the edge of the network, is a "one-hop" communication to the end-user that has been introduced to reduce the latency of a network. The aims are especially to real-time services that require Ultra-Reliable Low-Latency (URLL) connections. An Edge Cloud (EC) [2] is an infrastructure that supports the MEC paradigm which provides resources similar to conventional cloud datacenter but

1 College of Computing, Sungkyunkwan University, Suwon, 440-330, Korea.
2 Department of Electrical and Computer Engineering, Sungkyunkwan University, Suwon, 440-330, Korea.
3 Department of Liberal Art, Seoul Theological University (Computer Engineering), Bucheon, 422-742, Korea.
* Corresponding author (choo@skku.edu and moonseong@stu.ac.kr)
[Received 05 June 2020, Reviewed 15 June 2020, Accepted 15 July 2020]
☆ This work is partly supported by the Ministry of Education, IITP, and NRF, Korea, under the GITRC support program (IITP-2020-2015-0-00742), High-Potential Individuals Global Training Program)(2019-0-01579), AI Graduate School Support Program (No.2019-0-00421), and mid-career support program (NRF-2020R1A2C2008447).
☆ A preliminary version of this paper was presented at ICONI 2019.

http://dx.doi.org/10.7472/jksii.2020.21.4.1

on a smaller scale. The resources provided by EC are computational power, storage, and network connectivity. Using the provided resources, EC can host a different type of services by using virtualization technologies. In other words, EC can host cloud-based services at the edge of the networks by using Virtual Network Function (VNF) technologies [3]. VNF lets the services to be represented as virtual instances which later can be migrated to a new host easily using a certain approach.

Service migration on EC can be separated into two parts, the migration process and migration strategy. Migration process mostly is how the migration takes place. The performance parameter is the time taken to migrate a Virtual Machine (VM) from the current host to another host. Live Migration [4] is one of the most promising migration process. It is a process of moving a running VM between different physical machines without disconnecting the client. Memory, storage, and network connectivity of the VM are transferred from the original guest machine to the destination. Migration strategy is to use different performance parameters such as vehicle mobility or network quality on selecting the best candidate EC to host a migrated VM. In another word, it is called EC selections. This paper covers the scope of EC selections which uses mobility data as the performance parameter.

Deployment of EC is in a distributive manner, meaning a bunch of EC is connected to a central cloud (larger scale datacenter). Because EC is closer to end-users, latency is lower compare to the connection between end-users and central cloud. But with the reduction of latency, other factors must be taken into consideration such as user mobility, VM migration, Quality of Service (QoS) maintenance and etc.

This paper targets the service migration for the connection between EC and an emergency vehicle with a more specific scenario, ambulance mobility. Our idea is that future generation ambulances will be equipped with an advanced medical haptic technology, thus it will allow specialized doctors to deliver a real-time interactive treatment to on-board patients during critical situations. The on-board pieces of equipment will create the QoS requirement for data streaming which all the data can be represented as a different type of flows. The flows that are defined in our study are haptic, audio, video and some common required medical information. Each flow comes with its own set of QoS requirements. In our study, we defined the QoS requirements in terms of data rate and latency. Furthermore, we assume that the path of the ambulance is fixed, considering statistical data [5].

All service migration schemes must follow QoS requirements of services to maintain reliability even though they aim to reduce the latency of a connection. Mobility-aware Service Migration (MASM) is an offline algorithm that aims to minimize total network cost: routing cost and migration cost when migrating services between ECs considering the mobility path of the ambulance.

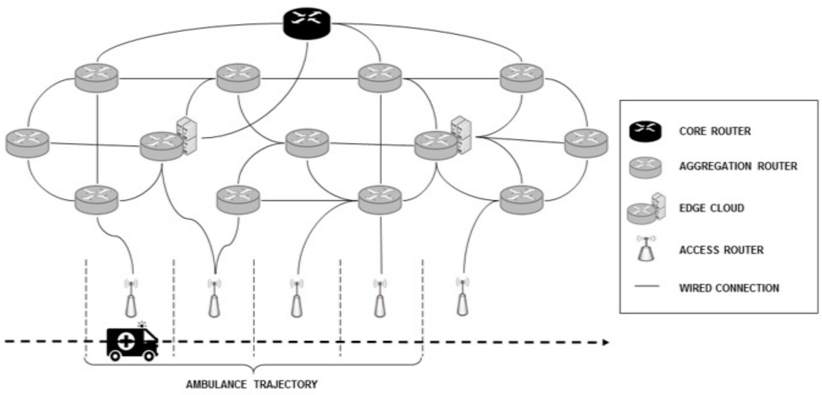
The remainder of this paper is organized as follows. Section 2 reviews the related work. The problem statement is presented in Section 3. In Section 4, we discuss about the proposed algorithm, MASM. In Section 5, an analysis of the proposed approach is provided. Finally, the paper is concluded in Section 6.

2. Related Work

To solve the service migration problems considering the high mobility scenario, some related studies must be taken into consideration. The related work is categorized based on the different approaches on how to select the best MEC server, another word for EC, to host a service. The parameters that are taken into consideration are vehicles mobility, MEC server capacity, VM migration cost, and certain QoS requirements related parameters such as latency, bandwidth and throughput.

Some paper related to using MEC in a vehicular network scenario focuses more on the mobility part but lack of discussions on the MEC selections. In paper [6], the authors proposed using a vehicular cloud radio access network (vCRAN) in the automotive field with edge computing infrastructure where communication and networking resources were centrally controlled. The proposed vehicular network consisted of remote radio heads (RRHs) located at roadsides, MEC servers responsible for signal processing and service management, a cloud server managing MEC servers, and a software-defined network (SDN). Although the paper provides a vehicular network mobility support method for the allocation of services in the MEC servers, the paper only considers the delay minimization without considering the cost of migrating VMs before selecting a new MEC server to migrate a service.

Besides focusing towards mobility, some paper focuses



(Figure 1) Example of topology generated during simulation.

more on inter-MEC handoffs for the VNFs of the corresponding SFC. The paper in [7] mentions that without inter-MEC handoffs for the VNFs of the corresponding SFC can affect the user satisfaction of the service due to unacceptably long service delay because of user mobility. To solve this problem, the author proposed an on-line algorithm called Follow-Me Chain which includes two key components: SFC placement and SFC migration. Both of the components is to determine, upon a handoff, which VNFs of the SFC needs to be migrated and with how many resources of each MEC server to be allocated to the user so that the service delay constraint of the SFC will not be violated. The paper does not discuss in details about MEC server selections in an ultra-dense cellular network of an urban-like environment.

The other approach is targeting the problems with maintaining QoS during service migration. In paper [8], the authors propose a potentially alternative way called Carry-on State Service Handover (CaSH) to resume the ongoing service at a new appropriate MEC host (MEH) during user equipment (UE) movement by having the UE carry the service state, and present the advantages of this approach. The proposed scheme is promising but it might not be suitable in a real-life scenario because, without VM migration, MEC server capacity must be higher due to the pressure of user traffics.

Lastly, there is some approach that takes into consideration of a specific high mobility scenario and discusses the MEC server selection strategies while maintaining QoS. In paper [9], a scenario of emergency medical services is considered. In emergency medical services, the lag time between injury and treatment is one of the most critical parameters concerning patient survivability. In this paper, a novel reliable multi-modal e-health high mobility service optimization framework was proposed for ambulances utilizing MEC servers to efficiently transport real-time patient information to the hospital.

3. System Model

There are two models in the system, which are the network model and the dataflow model.

3.1 Network model

Figure 1 shows one of the network topologies that is used in the study. A wireless backbone network is modelled as graph $G(V, E)$, where V is set of nodes/vertices, and E is the set of edges/transmission-links, which can be wireless or wired connections. An edge has a capacity and will cause a delay to the flows that go through it. Without loss of generality, one gateway node (core router) is assumed.

The network is divided into three layers. There are core routers, aggregation routers, and access routers layer. Edge Clouds is responsible for handling VNF chaining for supporting virtual instances of flows. ECs are deployed on the aggregation router layer. Mobility is taken into consideration by an ambulance, which is assumed to have pre-defined mobility that is decomposed into location epochs. In the other word, the trajectory of the ambulance is partitioned into epochs according to Access Router (AR) as in Figure 1. Each location epoch is a period when an ambulance is within an AR connection. Handovers in the research are the inter-handover between ARs. A Base Station (BS) is assumed to have a one-to-one connection with the access router, thus is omitted in the research.

The main costs in the network are routing cost and migration cost. The routing cost occurred when a flow is assigned to a path at the current location epoch. The migration cost occurred when a set of flows is migrated from the previous EC to another at a different location epoch. Total routing cost will be the sum of the routing cost for every location epochs depending on how many location epoch is needed for the ambulance to complete its mission.

The research aims to reduce the overall network cost containing routing cost and migration cost. The path selection decision determines the routing cost outcome, while the EC migration decision determines the migration cost outcomes. Several constraints will limit feasible choices of paths such as the capacity and latency of a path, number of paths, number of ECs, and prioritization of flows.

3.2 Dataflow model

As services are deployed onto EC using VNF, flows of data are present. In the paper, the type of flows is specified, as shown in table 1. Data rate and latency requirements [10, 11, 12] are further specified to complete the QoS requirements table. Haptic, audio, and video is the core services that must be reliable to the ambulance. Thus, to ensure the reliability for these core services, there are prioritized into synchronized flows and other flows is consider as normal flows. By doing this, we can ensure that the QoS requirement is not validated.

The haptic flow provides necessary information for the doctors to treat a patient remotely or if the situation is worse, doctors can do on-board surgery. We consider the haptic communication system based on multiple Omni. Each Omni represents a virtual finger with a rate of 24bytes/sec. For audio, it is assumed that one to three channels of rich audio of 64kbps for each channel. For video, one to three streams of high-quality 4K video with rate requirement for every stream to be approximately 10Mbps.

(Table 1) QoS requirements

Flow type	Rate	Latency
Haptic	192kbps	30ms
Audio	64kbps	100ms
Video	10Mbps(4K stream)	
ECG	24kbps	
EMG	100kbps	
Heart Rate	5kbps	60s
HQI	100Mbytes(per request)	10s

Normal flows is consisting of Electrocardiogram (ECG), Electromyogram (EMG), Heart rate signal, and High Quality Imaging (HQI) for purpose such as x-ray or ultra-sound scans. The following ECG, EMG, Heart Rate and HQI is refered in the paper [11].

4. Proposed Idea

This section explains our proposed heuristic algorithms on minimizing the routing and migration costs. There are two approaches in the research when minimizing the routing and migration costs. Firstly, the selection of ECs is made by EC selection algorithm. Then, after the EC is selected, the flow assignment algorithm will be executed to minimize the routing cost further in the scheme.

4.1 Flow assignment

The flow assignment algorithm [9] responsible for assigning each flow to a path. A path here is the path that contains the chosen EC for each location epoch. The algorithm also checks whether the QoS of flows can be satisfied before assigning it to a paths. The algorithm implements the idea of flow prioritizing, which it will initially assign the synchronized flows before assigning other flows.

As the flow assignment algorithm can be executed after an EC is chosen; thus, we need to consider some EC selection algorithm. The EC selection algorithm in the paper consists of target heuristic, Greedy Migration, and our proposed heuristic, Mobility-Aware Service Migration.

4.2 Greedy Migration n (GM-n)

This heuristic is as what the name stated, a greedy-based migration algorithm. The idea is choosing a static EC to serves the ambulance for the n location epochs such that the overall network cost within larger n location epoch has a smaller impact from the result of migrations. The result of migrations could generate significant cost in the scenario when the migration cost is high.

The greedy way for selecting the best EC would be by selecting the available ECs with minimum routing cost, EC_set, for each location epoch, t. From the closest available ECs, EC_set, GM-n selects the EC with the overall minimum routing cost. The algorithm is as follows:

Algorithm 1: Greedy Migration n (GM-n)

```
Input: G, T, n
Output: data_flows, EC_selected
EC_set ← ∅
t ← current location epoch
while t ≤ T do
    for each t in the [t, t+n) epochs do
        EC_t ← the closest EC to the current AR
        EC_set ← EC_set ∪ {EC_t}
    end
    min ← ∞
    for each EC_i in EC_set do
        cost ← sum of routing costs from EC_i to
        each AR of [t, t+n) epochs
        if cost < min then
            min ← cost
            EC_selected ← EC_i
    end
    data_flows ← Flow assignment [9]
    t ← t+n //next n location epochs
end
```

4.3 Mobility-Aware Service Migration n (MASM-n)

MASM-n is the improvement of GM-n. The biggest difference between GM-n and MASM-n is that MASM-n guarantees that the chosen candidate EC has minimum routing cost and meets the QoS requirement (data rate and latency) when the flow is assigned to the network paths. Our proposed algorithm, MASM-n are as follows:

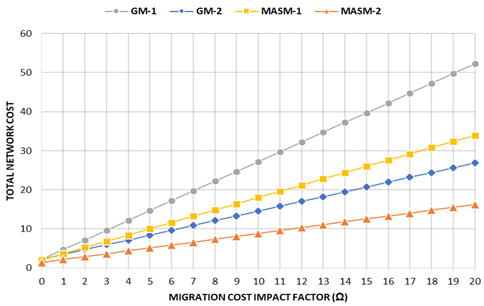
Algorithm 2: Mobility-Aware Service Migration n (MASM-n)

```
Input: G, T, n, k
Output: data_flows, EC_selected
EC_set ← ∅
t ← current location epoch
while t ≤ T do
    for each t in the [t, t+n) epochs do
        EC_topk_t ← the top k closest EC to the
        current AR [13]
        EC_set ← EC_set ∪ EC_topk_t
    end
    min ← ∞
    for each EC_i in EC_set do
        cost ← sum of routing costs from EC_i to
        each AR of [t, t+n) epochs
        if cost < min then
            min ← cost
            EC_selected ← EC_i
    end
    data_flows ← Flow assignment [9]
    t ← t+n //next n location epochs
end
```

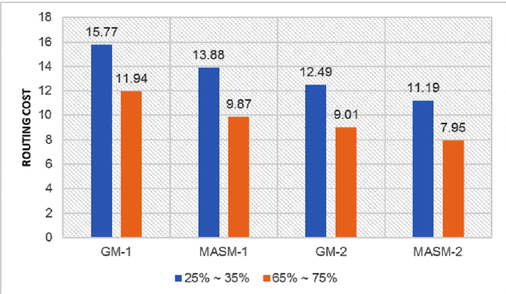
The way MASM-n guarantees the minimum routing cost is by first considering the top k subset paths, EC_topk_t, that the flows can be assigned later. If we have a set of paths from the EC to the current AR, MASM-n calculates the routing cost of all the paths. Then, MASM-n chooses the top k subset paths, EC_topk_t, with minimum routing cost based on Yen's k-th shortest paths [13]. After we have EC_topk_t with the minimum routing cost MASM-n repeats the process of GM-n on finding the overall closest EC, and this EC would be the one that is selected, $EC_{selected}$, for the current location epoch.

When the flow assignment algorithm is executed for assigning flows, the minimum routing cost of subset paths will not be changed due to validation of QoS requirement. For GM-n there is a tendency that the chosen subset paths will be changed due to validation of QoS requirement.

Both GM-n and MASM-n only consider the routing cost when finding the best EC, $EC_{selected}$. As this scheme is still in the preliminary phase; in the future, the scheme will conclude better strategies that consider both minimizing the routing cost and migration cost. At the meantime, migration cost in the current scheme is defined similarly as the routing



(Figure 2) Total network cost vs. migration impact factor. No. of Edge Clouds = 4, No. of Handovers = 10.



(Figure 3) EC proportion vs. routing cost. No. of Handovers = 10.

cost. Technically during simulation, we use a different random number generator to interpret them as an independent entity.

5. Performance Evaluation

The simulation was developed using C++ language and mostly involved Boost graph library to generate the network topology. The simulation follows the respective settings as in Table 2. All results are obtained from 100 Monte-Carlo iterations. To make a comparison, GM-1 is defined as the baseline algorithm in the research.

Total network cost is the minimum of routing cost and migration cost. But in order to observe the migration cost impact on the total network cost, we add one more parameter which is impact factor, Ω. Migration cost $(C_{m,m'})$ in the current scheme is defined as the routing cost from the previous EC to the EC that the service is migrated.

(Table 2) Simulation parameters

Parameter	Value
Ambulance mission duration	8 minutes
Average ambulance speed	67 km/h
Macro cell coverage radius	500 m
No. of location epochs, T	10
No. of AR	24
No. of Aggregation Router	2 -10
No. of EC	2 - 6
No. of flows	7
No. of synchronized flows	3
Capacity per link	2 Gbps
Propagation delay per link	1 - 1 ms
Routing cost per flow per link	1 - 100
Migration cost per flow	1 - 100

Figure 2 shows the migration cost impact factor, $\Omega*(C_{m,m'})$. In a lower migration impact scenario, MASM-2 received up to 70% gain compared to GM-1 and received up to 40% gain compared to GM-2. From the research observation, MASM-2 receive better result compare to target heuristic because of its nature on finding the overall lowest routing cost when finding the best EC.

Note that overall network cost also can be affected by the no. of ECs deployed in the aggregation layers and. Figure 3 shows the routing cost of different EC proportion. EC proportion is defined as followings:

$$\frac{No.of\ Edge\ Clouds}{No.of\ Aggregation\ Routers}\times 100\%$$

As the EC proportion increases, the no. of EC selection also increased. An increased no. of EC selections means that there are other subset path that has lower routing cost. For every heuristic we can see that routing cost decreases along the no. of ECs. To be remined, the proportions is not the only case that affect the routing cost. The other factor is the location of the ECs. Theoratically, the further the location of EC is from the end-user, the higher the routing cost will be.

6. Conclusion

In this research, an EC service migration issues were discussed. The paper also presented some preliminary scheme to solve the service migration problem between ECs.

More specifically, the research focused on a domain related to e-health services, and simulation was implemented. The performances were evaluated by comparing our proposed heuristic with a baseline heuristic, and trade-offs between parameters were discussed.

Reference

[1] K. Gray, L. Ibanez, S. Aylward et al., "MEC in 5G Networks," ETSI White Paper No. 28, first edition, Jun 2018. https://www.etsi.org/images/files/ETSIWhitePapers/etsi_wp28_mec_in_5G_FINAL.pdf

[2] D. Sabella, V. Sukhomlinov, L. Trang et al., "Developing Software for Multi-Access Edge Computing," ETSI White Paper No. 20, second edition, Feb 2019. https://www.etsi.org/images/files/ETSIWhitePapers/etsi_wp20ed2_MEC_SoftwareDevelopment.pdf

[3] J. Erfanian, B. Smith et al., "Network Function Virtualization," ETSI White paper on NFV priorities for 5G, Issue 1, October 2012. http://portal.etsi.org/NFV/NFV_White_Paper.pdf

[5] K. PetzÃd'll, J. PetzÃd'll, J. Jansson and G. NordstrÃum, "Time saved with high speed driving of ambulances," Accident Analysis & Prevention, vol. 43, no. 3, pp. 818-822, 2011. https://doi.org/10.1016/j.aap.2010.10.032

[6] Y. Kim, N. An, J. Park and H. Lim, "Mobility Support for Vehicular Cloud Radio-Access-Networks with Edge Computing," 2018 IEEE 7th International Conference on Cloud Networking (CloudNet), pp. 1-4, Tokyo, 2018. https://doi.org/10.1109/CloudNet.2018.8549365

[7] Y. Chen and W. Liao, "Mobility-Aware Service Function Chaining in 5G Wireless Networks with Mobile Edge Computing," ICC 2019 - 2019 IEEE International Conference on Communications (ICC), pp. 1-6, Shanghai, China, 2019. https://doi.org/10.1109/WPMC.2018.8712928

[8] K. Yunoki and H. Shinbo, "Carry-on State Service Handover between Edge Hosts for Latency Strict Applications in Mobile Networks," 2018 21st International Symposium on Wireless Personal Multimedia Communications (WPMC), Chiang Rai, pp. 472-477, Thailand, 2018. https://doi.org/10.1109/WPMC.2018.8712928

[9] G. Zheng, C.Y. Wang, V. Friderikos, and M. Dohler, "High mobility multi modal e-health services," 2018 IEEE International Conference on Communications (ICC), pp. 1–7, IEEE, 2018. https://doi.org/10.1109/ICC.2018.8422887

[10] V. Gokhale, S. Chaudhuri and O. Dabeer, "HoIP: A point-to-point haptic data communication protocol and its evaluation," Twenty First National Conference on Communications (NCC). IEEE, pp. 1-6, 2015.

[11] A. Grzech, P. Swiatek and P. Rygielski, "Dynamic Resources Allocation for Delivery of Personalized Services," Software Services for e-World Berlin, vol. 341, pp. 17-28, 2010. https://doi.org/10.1007/978-3-642-16283-1_6

[12] G. Gini, M. Arvetti, I. Somlai and M. Folgheraiter, "Acquisition and Analysis of EMG Signals to Recognize Multiple Hand Movements for Prosthetic Applications," Applied Bionics and Biomechanics, vol. 9, no. 2, pp. 145-155, 2012. https://doi.org/10.3233/ABB-2011-0024

[13] E. Q. V Martins and M. M. Pascoal, "A new implementation of Yen's ranking loopless paths algorithm," 4OR, vol. 1, pp. 121-133, 2003. https://doi.org/10.1007/s10288-002-0010-2

◐ 저 자 소 개 ◑

하 지 크(Haziq Hamzah)
2019년 성균관대학교 컴퓨터공학과(공학사)
2019년～현재 성균관대학교 대학원 소프트웨어학과 석사과정
관심분야 : 소프트웨어 정의 네트워킹, 멀티 액세스 엣지 컴퓨팅, 모빌리티 관리, 네트워크 기능 가상화 등.
E-mail : ajid951125@skku.edu

리 덕 타이(Duc-Tai Le)
2010년 베트남 국립 대학교 대학원 컴퓨터공학과(공학석사)
2016년 성균관대학교 대학원 컴퓨터공학과(공학박사)
2016년～2019년 성균관대학교 박사후 연구원
2019년～현재 성균관대학교 전자전기컴퓨터공학과 연구 교수
관심분야 : 무선 에드혹 및 센서 네트워크, 소프트웨어 정의 네트워킹, 엣지/클라우드 컴퓨팅, IoT 등
E-mail : ldtai@skku.edu

김 문 성(Moonseong Kim)
2002년 성균관대학교 일반대학원 수학과(이학석사)
2007년 성균관대학교 일반대학원 전기전자및컴퓨터공학부(공학박사)
2007년～2009년 미국 미시간주립대학교 컴퓨터과학공학과 연구원
2009년～2018년 특허청 사무관
2018년～현재 서울신학대학교 교양학부 조교수
관심분야 : 유무선 네트워크, 모바일 컴퓨팅, 머신러닝, 지식재산권 등
E-mail : moonseong@stu.ac.kr

추 현 승(Hyunseung Choo)
1990년 댈러스 텍사스 대학교 대학원 컴퓨터공학과(공학석사)
1996년 알링턴 텍사스 대학교 대학원 컴퓨터공학과(공학박사)
1997년～1998년 특허청 사무관
1998년～현재 성균관대학교 컴퓨팅대학 교수, 산학협력단장, 지능형ICT융합연구센터장
관심분야 : 모바일 센서 네트워크, 소프트웨어 정의 네트워킹, 지능형 모바일 컴퓨팅, 멀티 액세스 엣지 컴퓨팅, 머신러닝 및 인공지능 등
E-mail : choo@skku.edu