J Inf Process Syst, Vol.15, No.5, pp.1192~1200, October 2019
https://doi.org/10.3745/JIPS.04.0139

ISSN 1976-913X (Print)
ISSN 2092-805X (Electronic)

JOURNAL OF INFORMATION PROCESSING SYSTEMS JIPS

Intelligent Resource Management Schemes for Systems, Services, and Applications of Cloud Computing Based on Artificial Intelligence

JongBeom Lim*, DaeWon Lee**, Kwang-Sik Chung***, and HeonChang Yu****

Abstract

Recently, artificial intelligence techniques have been widely used in the computer science field, such as the Internet of Things, big data, cloud computing, and mobile computing. In particular, resource management is of utmost importance for maintaining the quality of services, service-level agreements, and the availability of the system. In this paper, we review and analyze various ways to meet the requirements of cloud resource management based on artificial intelligence. We divide cloud resource management techniques based on artificial intelligence into three categories: fog computing systems, edge-cloud systems, and intelligent cloud computing systems. The aim of the paper is to propose an intelligent resource management scheme that manages mobile resources by monitoring devices' statuses and predicting their future stability based on one of the artificial intelligence techniques. We explore how our proposed resource management scheme can be extended to various cloud-based systems.

Keywords

Artificial Intelligence, Cloud Computing, Edge-Cloud Systems, Fog Computing, Resource Management

1. Introduction

Recent advances in artificial intelligence and its related techniques are introduced [1,2] with emphasis on how these techniques affect resource management on various computing environments—Internet of Things [3], fog computing systems [4], edge-cloud systems [5], and intelligent cloud systems [6]. One of the benefits of using artificial intelligence techniques in the computing environments is that no human intervention is required when managing computing resources (resource monitoring, task assignments, virtual machine scheduling in virtualized computing, and task & virtual machine migration), while optimizing resource consolidation by running the iterations and multiplexing many logical components in the data center or the system [7,8].

Of the systems, we consider cloud-based computing systems and architectures in fog computing, edge-cloud systems, and intelligent cloud systems. We introduce recent studies on these computing systems based on artificial intelligence techniques and propose an intelligent resource management scheme that

Manuscript received July 12, 2019; first revision July 26, 2019; accepted September 3, 2019.
Corresponding Author: HeonChang Yu (yuhc@korea.ac.kr)
* Dept. of Game & Multimedia Engineering, Korea Polytechnic University, Siheung, Korea (jblim@kpu.ac.kr)
** Dept. of Computer Engineering, Seokyeong University, Seoul, Korea (daelee@skuniv.ac.kr)
*** Dept. of Computer Science, Korea National Open University, Seoul, Korea (kchung0825@knou.ac.kr)
**** Dept. of Computer Science & Engineering, Korea University, Seoul, Korea (yuhc@korea.ac.kr)

manages mobile resources by monitoring devices' status and predicting their stability information based on one of the artificial intelligence techniques (i.e., the hidden Markov model).

Although the hidden Markov model is developed in the concept of the statistics and pattern theory, it has received significant attention, especially in the artificial intelligence field, where an inference model is used for estimating future states [9-11]. Notable applications of the hidden Markov chain are the temporal pattern recognition and the reinforcement learning, including voice-to-text and text-to-speech applications, handwritten text recognition, gesture recognition, grammatical tagging, word-category disambiguation, score following for music, and bioinformatics. Recently, the hidden Markov models have been generalized for complex data structures and nonstationary data with pairwise/triplet Markov models.

The remainder of the paper is organized as follows: After reviewing artificial-intelligence-related techniques in the Internet of Things and fog computing in Section 2, we describe the edge-cloud systems and examine recent results of resource management schemes in Section 3. The proposed intelligent resource management scheme based on the hidden Markov chain for mobile devices is presented in Section 4. Finally, Section 5 concludes the paper with future research directions.

2. Fog Computing Systems

Fog computing is a decentralized computing infrastructure in which resource-intensive functionalities are located somewhere between the cloud and the data source to reduce resource burdens (computation, network bandwidth, storage capacity). Fig. 1 shows the fog computing architecture with virtualization enabled. There are two features in the fog computing architecture, that is, the quality of services and the energy-aware deployment in virtualized computing environments.

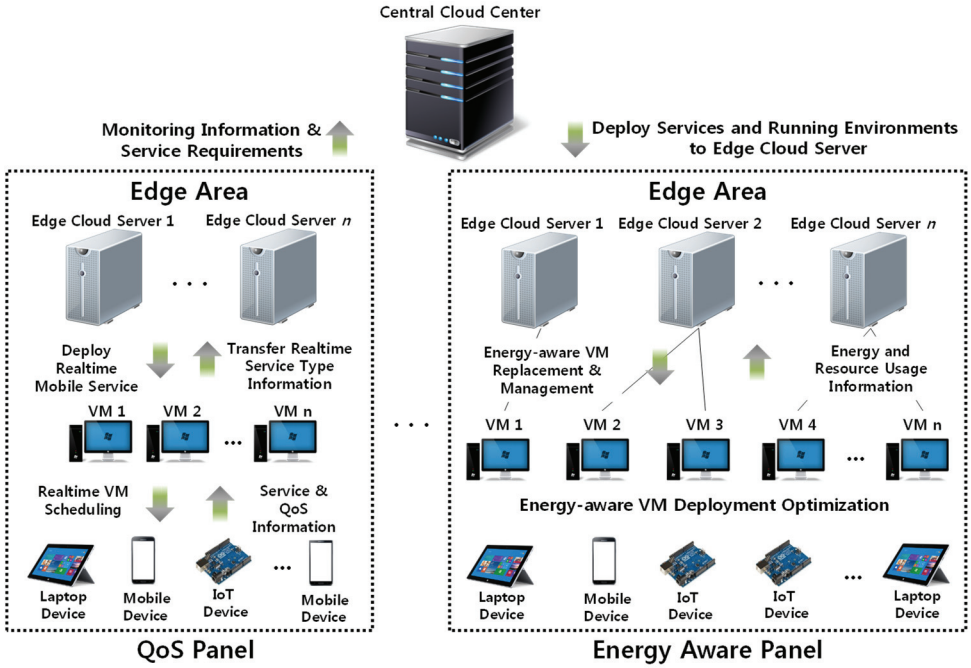

Fig. 1. The fog computing architecture with virtualization enabled.

For the quality of services, the central cloud center interacts with the edge-cloud servers by retrieving resource monitoring information and service requirements. The edge-cloud servers deploy real-time mobile services to the virtual machines for the Internet-of-Things devices. At the same time, the deployed virtual machines transfer real-time service information, including device types. The Internet of Things devices (laptop, smartphone, and sensor devices) directly interact with the deployed virtual machines by delivering service and quality of service information. According to the received data, the virtual machines are scheduled in real-time.

On the energy-aware panel in Fig. 1, the architecture is similar to the quality of service panel, but the resource management scheme is different. In other words, the edge-cloud servers consider the trade-off between energy consumption and performance of the virtual machines and the Internet of Things devices. Therefore, the virtual machines send energy and resource usage information to the edge-cloud servers, which then optimize the deployment of virtual machines in terms of energy consumption.

In the fog computing context, the authors of [12] proposed an intelligent algorithm for offloading decisions when multiple Internet of Things devices are present nearby. The focus of the proposed algorithm is twofold: device-driven intelligence and human-driven intelligence for network objectives (energy consumption, latency, network bandwidth, network availability, and security/privacy preservation).

Another research approach for fog computing in this context is healthcare [13–15]. Healthcare applications with fog computing technologies are promising because such technologies enable us to develop prediction techniques for our daily lives, for which real-time and low latency are of importance.

3. Edge-Cloud Systems

In the edge-cloud systems, resource capabilities such as computing, network, and storage are distributed throughout the system, taking it closer to where traffic originates. Fig. 2 shows the edge-cloud architecture with mobile devices. Assume that a mobile device's edge cloud is associated with the edge-cloud server A in the figure. The tasks of the mobile device (partial service) from the central cloud server are copied to the edge-cloud server A.

At this stage, if the mobile device moves to another location closest to the edge-cloud server B, the associated tasks (partial service) are scheduled for migration from the edge-cloud server A to B. Note that the user of the mobile device does not acknowledge the processes in the central cloud server and the edge-cloud servers. Again, if the mobile device moves to another location closest to the edge-cloud server C, the associated tasks (partial service) are scheduled for migration from the edge-cloud server B to C. In this manner, mobile and real-time applications can benefit from the edge-cloud systems with latency reduced.

For intelligent edge-cloud systems, the authors of [16] proposed a stochastic online machine learning technique that learns from the dynamism of the system. The study summarizes applications to wireless communications (traffic classification, channel coding/decoding, channel estimation, scheduling, and cognitive radio) and suggests an online learning framework for mobile edge computing systems for big data analytics.

The offloading technique is widely used since it significantly reduces the burdens of computation and communication of mobile devices. For this matter, the authors of [17] proposed a two-way initiative scheme for offloading decision making. It detects network congestion and solves the offloading problem

by integrating the random early detection algorithm.

For the resource management with respect to energy consumption in edge cloud computing environments, Liu et al. [18] proposed an energy management system for the Internet of Things devices based on the deep reinforcement learning. The proposed framework allows agents to schedule tasks, considering energy consumption. Unlike the basic method, Liu et al.'s approach considers the capacity limitation of edge servers.

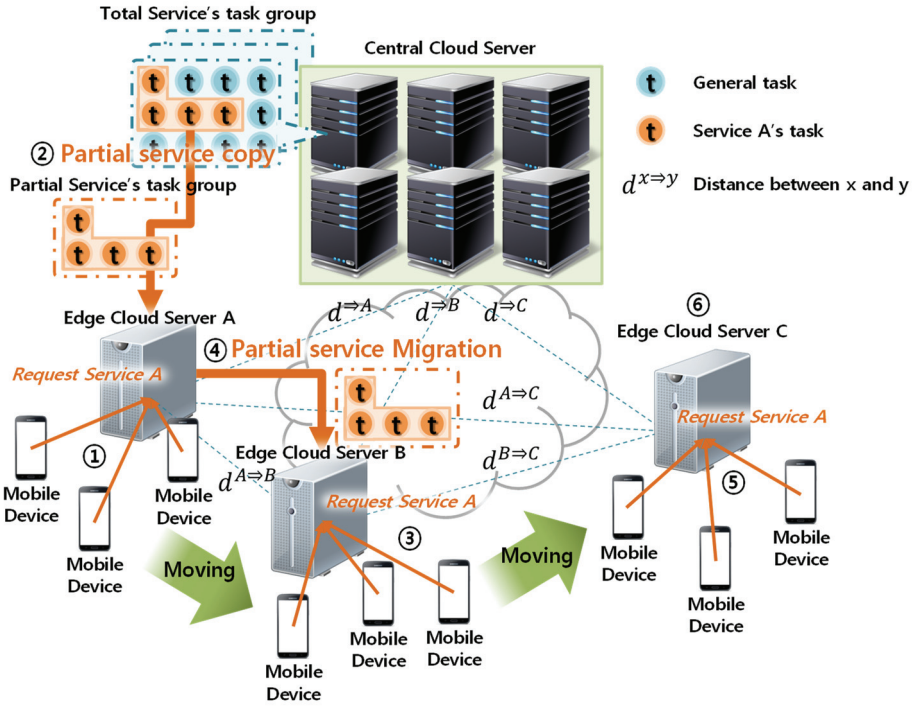

Fig. 2. The edge-cloud architecture with mobile devices.

4. Intelligent Cloud Computing Systems

Cloud computing is a pay-as-you-go and on-demand model for delivering computing resources (CPU, memory, storage, and network) based on the virtualization technology. When a user requests a certain amount of computing resources, the cloud data center schedules for provisioning as requested, and the requested virtual machine (or container) can be used within a minute.

To provide computing resources, the cloud data center pools a significant number of physical machines. Hence, the resource consolidation of the cloud data center affects performance and management costs. A well-managed cloud data center can save energy consumption and reduce carbon dioxide emissions.

For managing computing resources, artificial intelligence techniques can be integrated into the cloud computing systems. The authors of [19] proposed dynamic resource prediction and allocation techniques in the cloud radio access network for 5G. The method uses long- and short-term memory for predicting throughput and employs a genetic algorithm for allocating resources.

Chien et al. [20] proposed an intelligent architecture for beyond 5G heterogeneous networks. The research aim of the architecture is to improve network performance in edge cloud computing environments based on artificial intelligence techniques. For maintaining the quality of services, the authors use the packet forwarding technique, and they recommend appropriate deep learning methods for different network themes.

Zhang et al. [21] proposed a multiple algorithm service model to support heterogeneous services and applications. The model is designed to reduce energy consumption and network latency/delay by consolidating virtual machines in the cloud computing system. To solve the optimization problem, the authors use the tide ebb algorithm that finds robust results by assessing the relationship between computation speeds and energy costs.

For managing mobile devices in the cloud computing environments, we propose an intelligent resource monitoring scheme that predicts their future stability based on the hidden Markov model. Fig. 3 shows the proposed workflow of artificial intelligence applications with the hidden Markov chain model. The workflow in the figure is based on the iterative model. Note that other task models can also be applied in the proposed model.

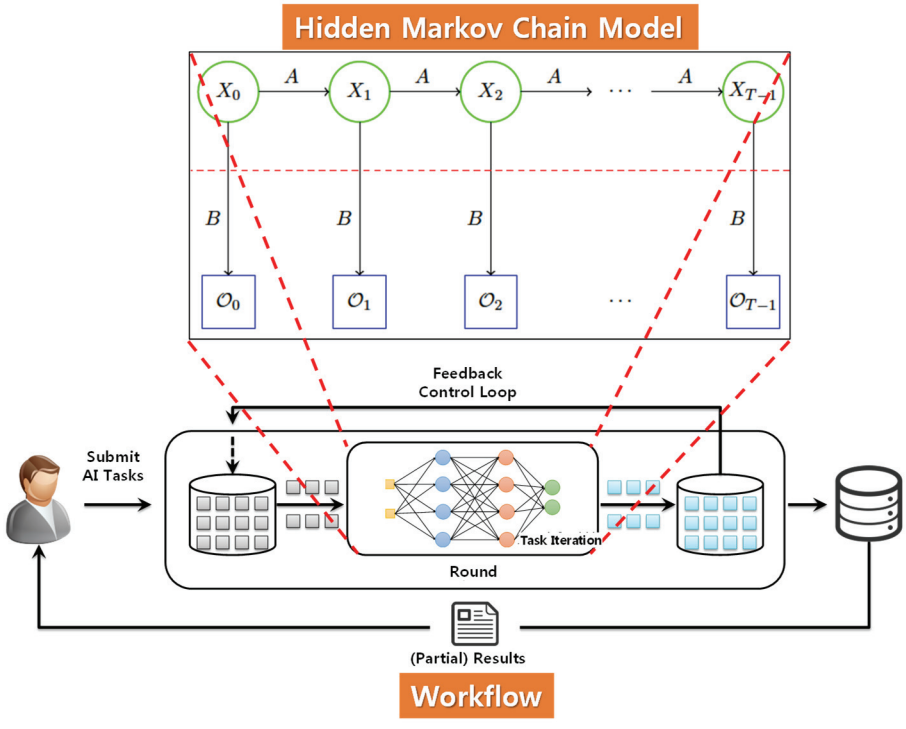

Fig. 3. The workflow of artificial intelligence applications with hidden Markov chain model.

When a user submits one or more artificial intelligence tasks, computing resources for the tasks are allocated via the cloud portal system. The allocated resources can be virtual machines, containers, or edge cloud servers according to the cloud computing environment. Then, the artificial intelligence tasks are performed by iterating the feedback control loop. After one round finishes, the (partial) results are forwarded to the input for the next round.

In the feedback control loop stage, the hidden Markov chain model is applied. The hidden Markov chain model uses current and past mobile devices' stability information and predicts future stability. More specifically, we regard the monitoring information as observable states and calculate a probability for hidden states. By computing the probability for hidden states, we predict the future stability of mobile devices. The predicted stability information can be used for cloud consolidation and cloud resource scheduling.

Table 1 shows the comparison and summary of resource management schemes based on artificial intelligence techniques in cloud computing environments. With respect to the categories of the cloud-based systems, our scheme is closely related to intelligent cloud computing systems. For the characteristics, our scheme is differentiated from other studies. In the proposed intelligent resource management scheme, mobile devices in the cloud-based system (including fog computing and edge-cloud) are periodically monitored, and the monitored information is used for predicting future stability and mobility based on the hidden Markov model. Thus, our scheme can be used for general cloud applications such as task scheduling, resource consolidation, and computation offloading while optimizing the overall system performance.

Table 1. Comparison and summary of resource management schemes based on artificial intelligence

Category	Study	Characteristic	Technique/consideration	Application
Fog computing	[12]	Latency reduction for the Internet of Things systems	Computation offloading, load balancing, and interoperability	Healthcare, Internet of Things
	[13]	Energy and latency reduction	Machine learning, task offloading	Body sensor network, health monitoring
	[14]	Achieve overall system performance	Geo-distributed systems between sensor nodes and cloud	Healthcare, smart home
	[15]	Sensitive data protection, delay reduction	Patient-driven healthcare architecture	Healthcare (individual, clustered)
Edge-cloud	[16]	Decoupling of tasks between time slots and edge devices	Machine learning for wireless communications	Mobile edge computing, big data analytics
	[17]	Avoidance of network congestion	Computation offloading, wired/wireless communication	Task scheduling in edge-cloud systems
	[18]	Improvement of the energy management performance, reduction of the execution time	Energy-aware scheduling scheme with deep reinforcement learning	Smart cities (smart building, smart power grid, multi-energy networks)
Intelligent cloud computing	[19]	Improvement of chip assembly and production efficiency	Cognitive manufacturing, intelligent manufacturing	Robot-factory
	[20]	Implementation of intelligent system architectures and network function	Heterogeneity of beyond 5G	Resource allocation, integrated packet forwarding
	[21]	Optimization of energy consumption and delay	Workload weights and the computation capacities	Artificial intelligence applications
	Ours	Predict future stability and mobility	Hidden Markov model	Mobile and artificial intelligence application

5. Conclusions

Management of mobile devices is not a trivial task since it is challenging to predict movement and faults. The proposed intelligent resource management scheme predicts mobile devices' stability based on the hidden Markov model. With monitoring information, the future stability information of mobile devices can be obtained. We divided cloud resource management techniques into three categories (i.e., fog computing systems, edge-cloud systems, and intelligent cloud computing systems), and analyzed various schemes for cloud resource management and its requirements (quality of services, service level agreements, and availability of the system) based on artificial intelligence techniques. Future work includes further improvement of algorithms, performance evaluation of various cloud services (e.g., backup, replication, checkpoint, task, and virtual machine migration).

Acknowledgement

This research was supported by Basic Science Research Program through the National Research Foundation of Korea (NRF) funded by the Ministry of Education (No. NRF-2018R1D1A1B07045838).

References

[1] M. Yao, M. Sohul, V. Marojevic, and J. H. Reed, "Artificial intelligence defined 5G radio access networks," *IEEE Communications Magazine*, vol. 57, no. 3, pp. 14-20, 2019.

[2] M. H. ur Rehman, I. Yaqoob, K. Salah, M. Imran, P. P. Jayaraman, and C. Perera, "The role of big data analytics in industrial Internet of Things," *Future Generation Computer Systems*, vol. 99, pp. 247-259, 2019.

[3] Y. Zhang, X. Ma, J. Zhang, M. S. Hossain, G. Muhammad, and S. U. Amin, "Edge intelligence in the cognitive Internet of Things: improving sensitivity and interactivity," *IEEE Network*, vol. 33, no. 3, pp. 58-64, 2019.

[4] A. H. Sodhro, S. Pirbhulal, and V. H. C. de Albuquerque, "Artificial intelligence-driven mechanism for edge computing-based industrial applications," *IEEE Transactions on Industrial Informatics*, vol. 15, no. 7, pp. 4235-4243, 2019.

[5] Y. Dai, D. Xu, S. Maharjan, G. Qiao, and Y. Zhang, "artificial intelligence empowered edge computing and caching for Internet of vehicles," *IEEE Wireless Communications*, vol. 26, no. 3, pp. 12-18, 2019.

[6] R. Donida Labati, A. Genovese, V. Piuri, F. Scotti, and S. Vishwakarma, "Computational intelligence in cloud computing," in *Recent Advances in Intelligent Engineering*. Cham: Springer International Publishing, 2020, pp. 111-127.

[7] M. Satyanarayanan and N. Davies, "Augmenting cognition through edge computing," *Computer*, vol. 52, no. 7, pp. 37-46, 2019.

[8] A. Kaplan and M. Haenlein, "Siri, Siri, in my hand: who's the fairest in the land? On the interpretations, illustrations, and implications of artificial intelligence," *Business Horizons*, vol. 62, no. 1, pp. 15-25, 2019.

[9] H. Gacanin and M. Wagner, "Artificial intelligence paradigm for customer experience management in next-generation networks: challenges and perspectives," *IEEE Network*, vol. 33, no. 2, pp. 188-194, 2019.

[10] W. C. Chien, C. F. Lai, and H. C. Chao, "Dynamic resource prediction and allocation in C-RAN with edge artificial intelligence," *IEEE Transactions on Industrial Informatics*, vol. 15, no. 7, pp. 4306-4314, 2019.

[11] Z. Li, L. Liu, and D. Kong, "Virtual machine failure prediction method based on AdaBoost-Hidden Markov model," in *Proceedings of 2019 International Conference on Intelligent Transportation, Big Data & Smart City (ICITBS)*, Changsha, China, 2019, pp. 700-703.

[12] A. A. Mutlag, M. K. Abd Ghani, N. Arunkumar, M. A. Mohammed, and O. Mohd, "Enabling technologies for fog computing in healthcare IoT systems," *Future Generation Computer Systems*, vol. 90, pp. 62-78, 2019.

[13] Q. D. La, M. V. Ngo, T. Q. Dinh, T. Q. S. Quek, and H. Shin, "Enabling intelligence in fog computing to achieve energy and latency reduction," *Digital Communications and Networks*, vol. 5, no. 1, pp. 3-9, 2019.

[14] A. M. Rahmani, T. N. Gia, B. Negash, A. Anzanpour, I. Azimi, M. Jiang, and P. Liljeberg, "Exploiting smart e-Health gateways at the edge of healthcare Internet-of-Things: a fog computing approach," *Future Generation Computer Systems*, vol. 78, pp. 641-658, 2018.

[15] A. Kumari, S. Tanwar, S. Tyagi, and N. Kumar, "Fog computing for Healthcare 4.0 environment: opportunities and challenges," *Computers & Electrical Engineering*, vol. 72, pp. 1-13, 2018.

[16] Q. Cui, Z. Gong, W. Ni, Y. Hou, X. Chen, X. Tao, and P. Zhang, "Stochastic online learning for mobile edge computing: learning from changes," *IEEE Communications Magazine*, vol. 57, no. 3, pp. 63-69, 2019.

[17] Z. Yin, H. Chen, and F. Hu, "An advanced decision model enabling two-way initiative offloading in edge computing," *Future Generation Computer Systems*, vol. 90, pp. 39-48, 2019.

[18] Y. Liu, C. Yang, L. Jiang, S. Xie, and Y. Zhang, "Intelligent edge computing for IoT-based energy management in smart cities," *IEEE Network*, vol. 33, no. 2, pp. 111-117, 2019.

[19] L. Hu, Y. Miao, G. Wu, M. M. Hassan, and I. Humar, "iRobot-Factory: an intelligent robot factory based on cognitive manufacturing and edge computing," *Future Generation Computer Systems*, vol. 90, pp. 569-577, 2019.

[20] W. C. Chien, H. H. Cho, C. F. Lai, F. H. Tseng, H. C. Chao, M. M. Hassan, and A. Alelaiwi, "Intelligent architecture for mobile HetNet in B5G," *IEEE Network*, vol. 33, no. 3, pp. 34-41, 2019.

[21] W. Zhang, Z. Zhang, S. Zeadally, H. C. Chao, and V. C. M. Leung, "MASM: A multiple-algorithm service model for energy-delay optimization in edge artificial intelligence," *IEEE Transactions on Industrial Informatics*, vol. 15, no. 7, pp. 4216-4224, 2019.

JongBeom Lim https://orcid.org/0000-0001-8954-2903

He received B.S. degree in information and communication from Baekseok University, Korea in 2009. In 2011 and 2014, he received M.S. and Ph.D. degrees in computer science and education from Korea University, Korea, respectively. From 2015 to 2017, he was a visiting professor with the IT Convergence Education Center, Dongguk University, Korea. Since March 2017, he is with the department of game and multimedia engineering, Korea Polytechnic University, Korea as an assistant professor. His research interests fall within the general fields of computer science and its applications including distributed computing and algorithms, cloud computing and virtualization, artificial intelligence and big data analytics, mobile and sensor networks, and fault-tolerant and resilient techniques.

DaeWon Lee https://orcid.org/0000-0001-7089-8205

He received his B.S. in the division of Electricity and Electronic Engineering from Soonchunhyang University, Asan, Korea in 2001. He received his M.E. and Ph.D. degrees in Computer Science Education from Korea University, Seoul, Korea in 2003 and 2009, respectively. He is currently an assistant professor in the Department of Computer Engineering at Seokyeong University in Korea. His research interests are in IoT, mobile computing, distributed computing, cloud computing, and fault-tolerant systems.

Kwang-Sik Chung https://orcid.org/0000-0002-9631-8579

He received a B.S., M.E., and Ph.D. degrees from Korea University, in 1993, 1995, and 2000, respectively. His major was distributed mobile computing. Currently, he has interesting in M-learning and cloud computing for smart learning. He is an assistant professor at Korea National Open University.

HeonChang Yu https://orcid.org/0000-0003-2216-595X

He received the B.S., M.S., and Ph.D. degrees in computer science and engineering from Korea University, Seoul, Korea, in 1989, 1991, and 1994, respectively. He has been a Professor of computer science and engineering with Korea University since 1998. From February 2011 to January 2012, he was a Visiting Professor of electrical and computer engineering in Virginia Tech. Since 2015, he has been the Vice President of Korea Information Processing Society, Korea. He was awarded the Okawa Foundation Research Grant of Japan in 2008. His research interests include cloud computing, virtualization, distributed computing, and fault-tolerant systems.