EiF: Toward an Elastic IoT Fog Framework for AI Services

JongGwan An, Wenbin Li, Franck Le Gall, Ernoe Kovac, Jaeho Kim, Tarik Taleb, and JaeSeung Song

The authors believe that the hyper-connected IoT ecosystem on fog platforms with contextual AI technologies is a promising solution. In this work, they introduce the EiF, a flexible fog computing framework that runs on IoT gateways with adaptive AI services fostered on the cloud. Their approach can be viewed as an integration of three emerging technologies, namely IoT, fog, and AI.

Abstract

The first generation of IoT was developed and deployed all over the world by connecting devices with common functionalities that were not sufficiently efficient or reliable for use in dynamic situations that require adaptive solutions. However, these fundamental IoT functions and services mainly targeted stable environments; there is consequently a strong need for the next generation of IoT services to be smarter, faster, and more reliable. We believe that the hyper-connected IoT ecosystem on fog platforms with contextual AI technologies is a promising solution. In this work, we introduce the EiF, a flexible fog computing framework that runs on IoT gateways with adaptive AI services fostered on the cloud. Our approach can be viewed as an integration of three emerging technologies, namely IoT, fog, and AI. Generally, EiF virtualizes an IoT service layer platform for fog nodes, and provides functions to manage and orchestrate various fog nodes; upon service virtualization and orchestration, AI services are fostered within both the federated cloud and distributed edge side and are deployed on fog nodes. We demonstrate the feasibility of EiF via the example of intelligent traffic flow monitoring and management.

Introduction

While cloud computing has laid a foundation for providing computing resources to end users and Internet of Things (IoT) service providers, fog computing is taking its first steps toward making IoT clouds more elastic ("Elastic IoT") by leveraging the computation capabilities at the cloud and edge level. Nevertheless, the current fog computing paradigm still faces challenges, particularly when dealing with situations that require context awareness and autonomous decision making under real-time constraints. Artificial intelligence (AI) technologies, as largely adopted by industry, promise to boost fog computing by providing distributed AI services in fog nodes. IoT, fog computing, and AI are the three most important technologies to drive the next-generation computing ecosystem. Each technology has been developed independently [1–3].

IoT is currently re-drawing our daily lives by enabling access to basically every physical object around the world. These things can deliver information and make the services of those physical objects available on the cloud [4]. However, many IoT platforms, supporting basic functions, do not satisfy the requirements of many industries and domains that are searching for smarter, faster, and more reliable IoT services. Many IoT manufacturers and service providers are paying more attention to fog computing because of its advantages, that is, its greater computational power, knowledge, and analytics placed as close to users as possible [5–7]. Fog computing typically virtualizes network resources and places them at fog nodes. However, network resources and IoT platforms, services, and knowledge are all expected to be located at fog nodes when using IoT. AI technologies are rapidly progressing and have recently accelerated in a manner that could not have been anticipated a few years ago. For example, deep learning technology is taking the task of processing unstructured data such as video or audio out of the hands of highly experienced experts and has enabled a larger workforce of programmers to use AI-provided cognitive services [8].

However, these technologies are only useful in specific fields. The efforts and research put into integrating these key technologies is currently still in its early stages. This article intends to find the synergy path through integrating IoT, fog computing, and AI to tackle the current limitations. We propose the Elastic Intelligent Fog (EiF) concept, an enhanced IoT service layer platform for fog nodes with advanced features such as semantics enablement, the virtualization of common IoT service functions, and the enablement of machine learning. This enhanced platform can be dynamically instantiated at the fog nodes and filter unnecessary data to make quick decisions. In addition, since the fog node is embedded with a lightweight AI engine that can derive context and operational information through managing sensors and adjacent intelligent fog nodes, many value-added IoT services can be introduced to users such as an intelligent traffic flow monitoring service.

The main contributions of this article are as follows:

- Define a framework for a global AI-enabled fog computing paradigm and IoT-fostered AI services.
- Define AI and IoT interactions in a distributed architecture along with the technology interoperability in fog computing environments.
- Provide network-user- and user-mobility-aware technologies located in the terminal to reduce network overhead and round-trip time.

The next section describes the challenges and motivation of the EiF framework, and then presents an overview of the EiF reference architecture, followed by detailing essential techniques. Then

Digital Object Identifier:
10.1109/MCOM.2019.1800215

JongGwan An, Tarik Taleb, and JaeSeung Song are with Sejong University; Wenbin Li and Franck Le Gall are with Easy Global Market; Ernoe Kovac is with NEC Europe Ltd.; Jaeho Kim is with Korea Electronics Technologies Institute; Tarik Taleb is also with Aalto University and Oulu University.

0163-6804/19/$25.00 © 2019 IEEE

we discuss the feasibility of EiF through introducing detailed procedures of EiF migration and simulation results. The last section summarizes this article and suggests future work.

Challenges and Motivation

In this section, we discuss five challenges (C1–C5) of IoT, cloud, and AI to motivate the need for an adaptive IoT service layer platform with the intelligence to run on fog IoT nodes (Fig. 1)

C1. No Architecture Framework: The huge amount of data and knowledge from the IoT world is the most promising field to drive cross-domain AI research and enable machine learning to create highly developed AI services [9]. However, existing IoT platforms require the intelligence to improve resource management and support cross-domain interactions. Defining common reference architecture for the cloud, IoT, and AI interactions helps us understand the behaviors of future IoT platforms and AI services.

C2. Virtualization and Softwarization: Many research activities focus on the virtualization of network access and core networks [10]. However, the network functions for IoT service layers and AI engines that execute AI algorithms can also be virtualized, bringing similar benefits such as optimization and quality of service (QoS) mediation. This limits the flexibility and on-demand ability, but the processing capability is close to the target objects. Having a flexible and self-optimizing software layer at the IoT fog nodes will greatly extend the applicability of AI/IoT services for different execution environments.

C3. Real-Time Dynamic Configuration: Fog IoT systems are still mostly manually configured and run in a static configuration [11]. Only a few systems have an explicit programming model that makes programming cloud/fog systems more effective and easier. In addition, neither current cloud computing nor the IoT service layer platform supports dynamic configuration, which would be required to migrate software instances that run on a fog node between data centers following their respective users.

C4. Semantics and Context for Operation: Modern AI technology is severely limited by two factors: domain information models without semantic descriptions, and lack of understanding of the operational conditions under which AI-based services operate. The former problem severely increases the effort of using available information by introducing labor-intensive manual tasks for data scientists. The latter limits the use of AI to well-prepared and understood contexts, which contradicts the idea of using AI in real-world operational systems. Automatically incorporating domain models into AI engines and training AI engines for an exploding set of contexts greatly increase the computational load and complexity of AI engines.

C5. Data Availability and Integration: AI engines are built into traditional applications, spanning a wide range from banking to home automation. In such cases, the AI uses predefined datasets that are typically trained while being designed and that only support a specific information model. However, AI will increasingly be deployed in dynamic environments with previously unknown sets of available information. AI/IoT systems will need to continually adapt and learn new skills as their environments change. Unfortunately, integrating various data from different information systems that have not been designed for integration necessitates the handling of subtle differences in the information model.

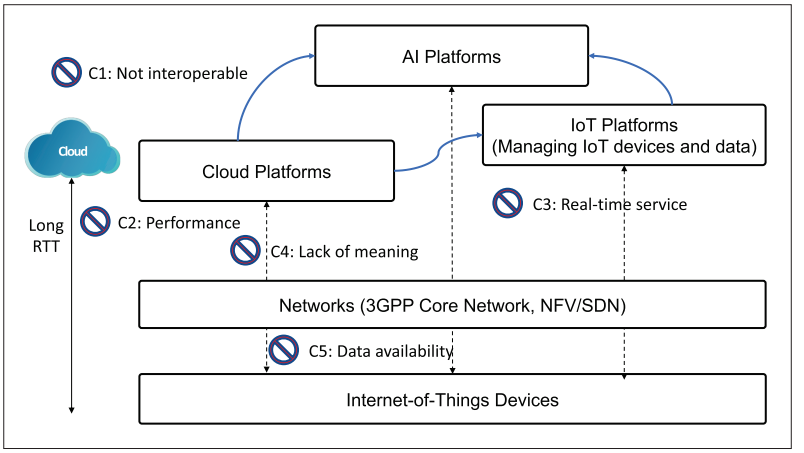

Figure 1. Challenges for three trendy technologies.

Technical Approach

This section provides a summary of the technical approach to build an EiF framework.

Fog Clouds Following Users: The concept of Follow-Me Cloud (FMC) was proposed to support the smooth migration of ongoing IP services between data centers to follow their respective users [12]. While several works have addressed service migration, EiF aims to consider this principle for mobile fog computing (MFC) [13]. Pushing computation processes to the edge server, in the vicinity of end users, reduces delays and enables applications that require millisecond-range response times. Similar in spirit to FMC, the concept of Follow-Me Fog (FMF) ensures that IoT applications/services follow the mobility of users through multi-access edge computing (MEC). A smooth implementation of FMF requires a full consideration of the MEC architecture. In addition, EiF uses AI to perform decisions about service migrations. Based on monitoring data related to user movement patterns, preferences, and requested services, efficient decisions can be made that improve the quality of experience.

Semantic and Context for AI: AI needs to solve two problems: meaning and context. EiF solves these using semantic and AI techniques. For the semantic task, EiF adopts a mechanism to translate natural language mechanically into data through word-generation-based logic to understand language through understandable principles and thus express language structure and thoughts [14]. For context, EiF fully understands the meaning of text and extends existing automated knowledge discovery functions by supplementing the text of a document with a repository containing global knowledge that can be retrieved from the context.

Virtualization and Slicing: In addition to basic network slicing features such as isolation and customization, EiF also introduces an AI-based predictive network resource allocation mechanism that can tremendously impact the scalability and performance of network slices. Moreover, EiF enables greater intelligence and network slic-

Traditional cloud and IoT platforms are not designed to support recent emerging needs for intelligent real-time applications. However, EiF enhances existing cloud and IoT platforms with a framework that enables dynamically extending the smart things managed by IoT with AI capabilities to intelligent things.

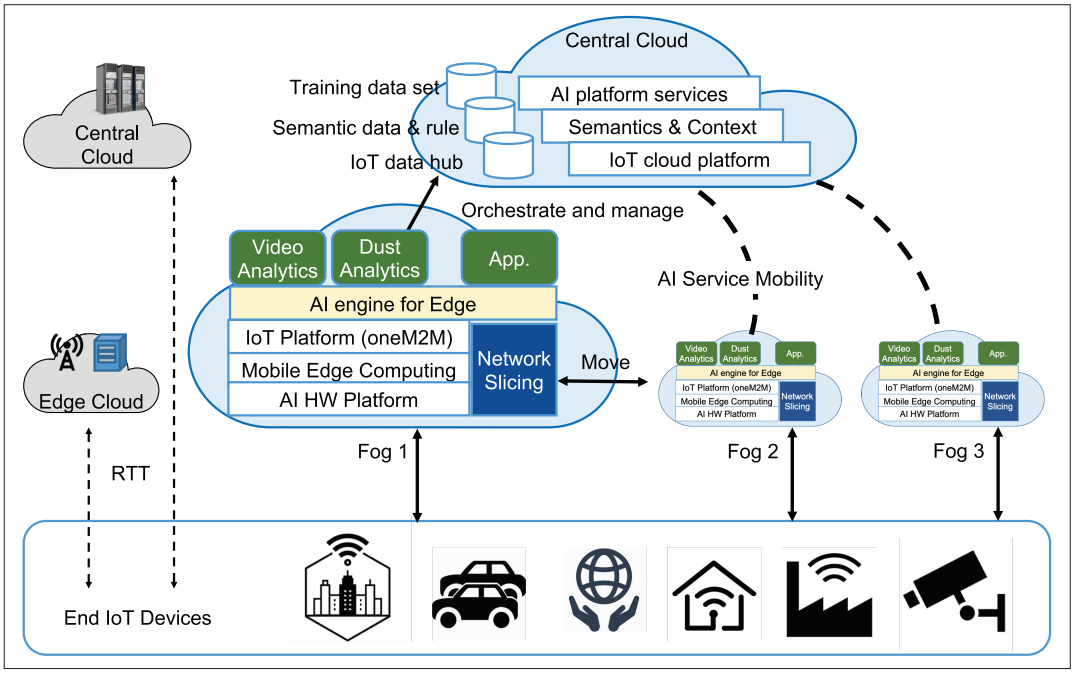

Figure 2. Conceptual view of EiF.

ing capabilities, particularly in the service mobility aspect, by using smart algorithms to determine and trigger network slice mobility and its implementation. In addition, network functions for IoT service layers are virtualized in EiF [15], which means that only necessary common IoT service functions, such as device management, group data handling, and discovery, can be dynamically deployed to a fog cloud node.

Orchestration and Management: Orchestration functions in EiF helps in managing all dependencies and relationships between services that comprise a particular application, similar to typical cloud services. In addition to this basic orchestration function, EiF applies machine learning to benefit coordinated fog and cloud service orchestration, focusing on proactive QoS enactment and service adaptation mechanisms. For this, EiF provides preemptive service management to predict the failure and QoS degradation of services based on the monitoring data that the monitoring system makes available, and can reconfigure or replace services that are predicted to fail before their failure occurs, which improves the availability and overall QoS.

Distributed AI and AI Reasoning: EiF benefits from existing AI technologies and provides machine learning techniques for distributed AI that can run on fog nodes. Adaptive and federated algorithms that can offer cognitive functions are developed for this with increased performance and efficiency, including characteristics from the federated learning approach, by implementing and expanding the techniques used in automated machine learning (https://automl.info/). Distributed AIs exchange data with each other through the fog cloud, and the central IoT cloud develops a federation of machine-learning and deep-learning models. In addition, both deductive and inductive reasoning are addressed to extract critical information from large sets of structured and unstructured data. This allows EiF to visualize and examine problems that need solving from different perspectives.

EiF System Overview

This section provides an overview of the EiF architecture. EiF's main goal is to provide cross-border AI services using highly distributed, reliable, and precise fog-cloud-based IoT systems.

Figure 2 shows the conceptual view of the EiF system. Traditional cloud and IoT platforms are not designed to support recent emerging needs for intelligent real-time applications. However, EiF enhances existing cloud and IoT platforms with a framework that enables dynamically extending the smart things managed by IoT with AI capabilities to intelligent things. Modern software methodologies such as virtualization and intention-driven software composition can help combine AI and IoT into a single system. The central cloud is designed to accommodate various functions that enable intelligent AI services using IoT data management and semantic technologies. This central cloud also provides the application programming interfaces (APIs) used by AI applications to utilize pre-defined AI and IoT components provided by a platform-as-a-service (PaaS) layer with a home in various data hubs. The fog cloud virtualizes network functions, IoT functions, and AI functions to provide required services to places that are close to users. As this fog cloud for services requires real-time feedback and processing, it has management function orchestration and manages required virtual functions in real time. Both layers communicate via a standardized interface to synchronize services and data, manage virtual resources, provision security and trust, train AI algorithms from distributed knowledge, and so on. AI engines such as face recognition, anomaly detection, and real-time situation analysis are integrated into the data-stream analysis process.

The EiF system is characterized by:

- Dynamic loading of virtual AI/IoT functions from trusted cloud servers

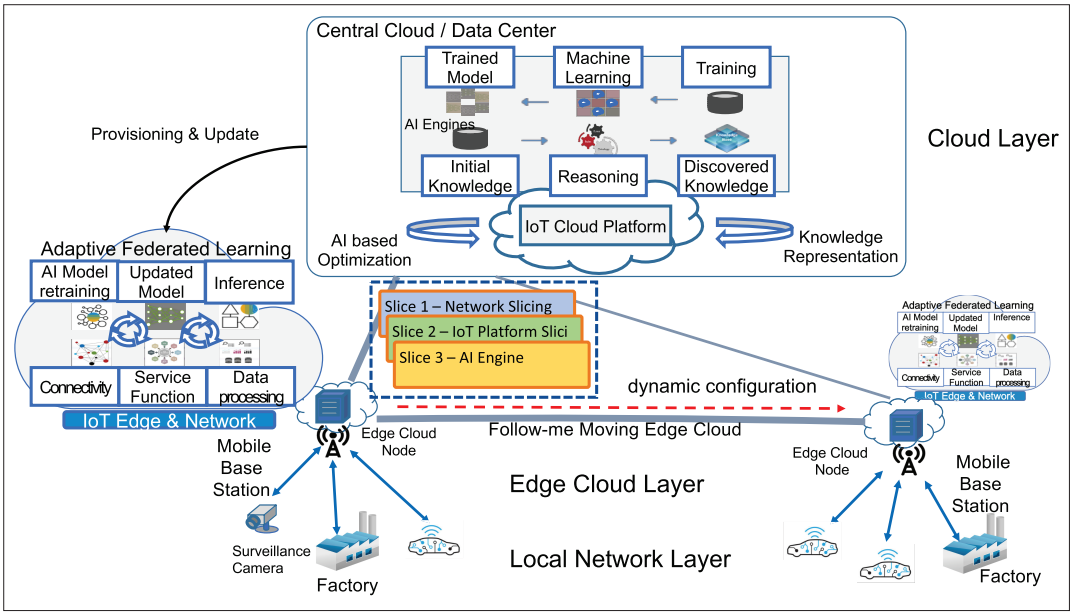

Figure 3. Detailed EiF functional architecture.

- The use of virtualization techniques for function shipping (e.g., data transformation routines) from trusted repositories or cloud servers and isolating different applications and processing flows as slices
- Smart-object orchestration with intelligent AI functions to form Intelligent Things
- Dynamic utilization of fog gateways and cloud servers based on the interpretation of abstract descriptions of processing flows while also profiling components and flows to achieve AI-based optimization

The predicted deployment of billions of devices and the global access required by cross-border AI services necessitate a layered real-time management approach, from fog computing models to cloud computing, and the ability to cope with common service functions such as discovery mechanisms and end-to-end security.

Realization of EiF

EiF provides a management system for AI services via an adaptive federation learning process as depicted in Fig. 3. EiF provides a management system for AI services via an adaptive federation learning process, and collects data from infrastructure resources, application components, data functions, and AI operations. Furthermore, the framework supports the dynamic provisioning of slicing units and respective resource elasticity. A scale-in/out model for slicing units permits the reservation of only necessary resources to deployed slices. In that framework, slicing algorithms are triggered together to support the aforementioned technique.

Additionally, the monitoring framework is designed to be adaptive as per the metrics being collected, which allows the incorporation of application-specific monitoring probes that follow identified metrics and AI operations-specific metrics based on stakeholder-defined parameters. AI reasoning is used to discover additional knowledge from federated data in the central cloud, while inference is applied to the edge cloud for local knowledge discovery. In particular, interactions between fog inference and cloud reasoning enable mutual enrichment and knowledge sharing.

In our previous FME work [12], an average downtime was experienced for lightweight service migration (e.g., around 2 s), and launching new services took some time, which can affect service functionality and quality of experience (QoE). EiF overcomes this problem by leveraging AI to anticipate user mobility patterns and traffic generation, and then accordingly customize services so that the QoE is not jeopardized and service interruption can be prevented. Thanks to EiF, if a user's mobility pattern can be predicted, the times and locations (i.e., fog or cloud) at which the services should be placed can be optimally determined, which can reduce the service relocation frequency (i.e., migration) while ensuring QoE.

EiF adopts a distributed AI with a two-layer structure to analyze the relationship between situations that occur in environments that require distributed AI. At the lowest layer, the components are detected in each environment. The relationships between the fog results obtained from the bottom layer are analyzed in the upper cloud layer. The system collects video images of the target situation to be analyzed from each environment to learn about the lowest layer of distributed AI. Through classifying objects' appearances and positions that constitute video images, the system analyzes them into a learnable AI dataset from which the system can start to advance its AI using deep learning architectures.

In the upper layer, a probability graph model is used to judge the situation in real time by considering the relationships between generated results. The system returns results as formatted metadata.

Use Case and Simulation Results

This section describes a use case and presents simulation results to show the feasibility of the EiF concept.

Use Case

One example of utilizing EiF is warning pedestrians about fine dust levels and providing environment information based on machine learning executed on a fog device. Unlike traditional fine dust warn-

AI reasoning is used to discover additional knowledge from federated data in the central Cloud, while inference is applied to the edge Cloud for local knowledge discovery. In particular, interactions between fog inference and Cloud reasoning enables mutual enrichment and knowledge sharing.

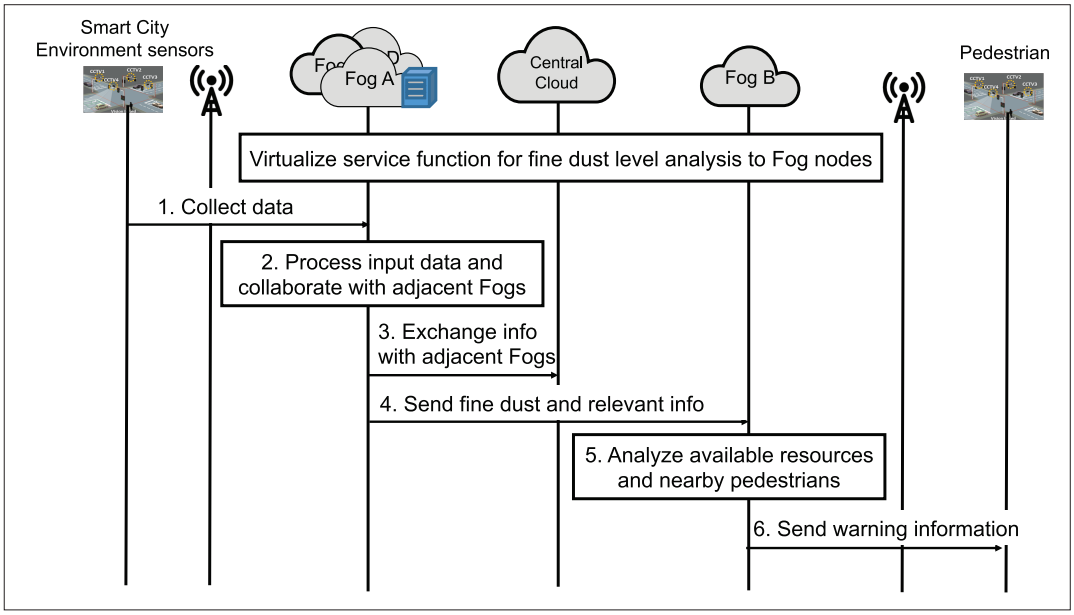

Figure 4. Procedure for fine dust warning service.

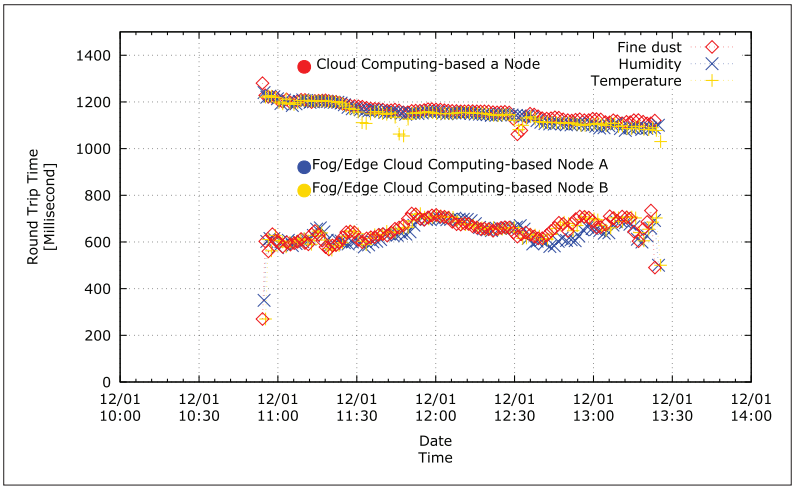

Figure 5. RTT latency.

ing forecast, the forecast service on EiF is focused on fine dust and other relevant information (e.g., temperature and humidity) 20 m from where the sample should be collected to analyze the air pollution at the ground and is available for any point or route around the globe thanks to the EiF's dynamicity and FMC. The forecast is created from a network of observational road edge/fog nodes, fine-dust models, and machine-learning algorithms. The service can provide hyper-local insights into current and forecasted road/ground fine dust conditions, pavement temperatures, and relevant information at a geo-location-level granularity.

Procedures

Figure 4 shows the procedures of how EiF supports the fine dust warning service through orchestrating various network entities. In this procedure, we assume that a virtualized IoT platform and distributed AI engine are deployed at fog nodes.

In our scenario, various input data that are helpful for analyzing the fine dust level from distributed sensors are analyzed by fog node A (Fog-A). The set of input data includes the temperature, humidity, local weather conditions, wind, and surveillance cameras deployed on the road. A distributed AI engine in Fog-A processes collected data, and if Fog-A's processing capability is insufficient, pool-based capability sharing/scaling with other nearby fog nodes can be utilized. If Fog-B, which is adjacent to Fog-A, knows the weather conditions but does not have other information required to analyze the fine dust level, Fog-B tries to get this information from Fog-A and/or the central IoT server platform. Then Fog-B analyzes all the information and retrieves the fine dust level for its local area. Fog-B also performs analysis on its video input streams to send the appropriate fine dust level (whether high- or low-quality) to pedestrians who are heading toward its locality.

Simulation Results

This section presents preliminary simulation results regarding the performance of EiF. We used the scenario addressed in Fig. 4 in the simulation with a real IoT dataset collected from the Santander Smart City in Spain and a portrait image dataset from Google. We showed the feasibility of EiF by measuring the round-trip time (RTT) of IoT data and machine learning results exchange. Here, we consider no congestion in the used links between entities. We compare EiF against the case where all processing was performed at the central cloud server.

Figure 5 shows how IoT sensor data (i.e., temperature, humidity, and fine dust concentration) can be exchanged and processed in both the EiF and the central cloud environment. The central cloud server receives all the measured data information directly from the sensors. In the case of EiF, Fog-1 and Fog-2 share measured data information with each other so that even if one of them encounters a problem, the service can work through the other nodes. This result clearly shows that EiF achieves lower data latency than the central cloud environment as the service and processing are always placed at the nearest fog entities. In contrast, if no elastic edge/fog concept

is used, the data latency and IoT data processing time obviously increase as the central cloud is only reachable via long communication paths. Figure 6 shows RTT for high-volume image data. Regarding data exchange time for machine learning, EiF is 1.5 times better than the conventional cloud environment.

These experimental results clearly show that EiF is more efficient than the central cloud system for both general IoT sensor data processing and high-volume data processing. However, the gain of EiF has a cost associated with signaling and coordination between the number of fog entities, which is obviously higher than the central system. This observation clearly indicates a need for more sophisticated distributed machine learning algorithms.

Conclusions

This article introduces EiF, a novel architectural framework that integrates recent trendy technologies, namely AI, IoT, and the cloud, to support cross-border AI services that use highly distributed, reliable, and precise cloud-based IoT systems and technologies. EiF virtualizes an IoT service layer platform such that only necessary functions can be sent to fog nodes in the vicinity of end users. Both lightweight IoT service functions and AI engines can be placed at fog nodes. The orchestration and management of various fog nodes are also key features of EiF. Finally, distributed AI technologies are deployed at fog nodes to train machine learning models and make decisions as rapidly as possible. We validated the EiF concept through simulations, and the results clearly show the benefit of using EiF compared to a conventional centralized cloud platform. For future work, we plan to apply EiF in various AI environments, such as recommendation models, which usually involve embedding tables requiring distributed inference.

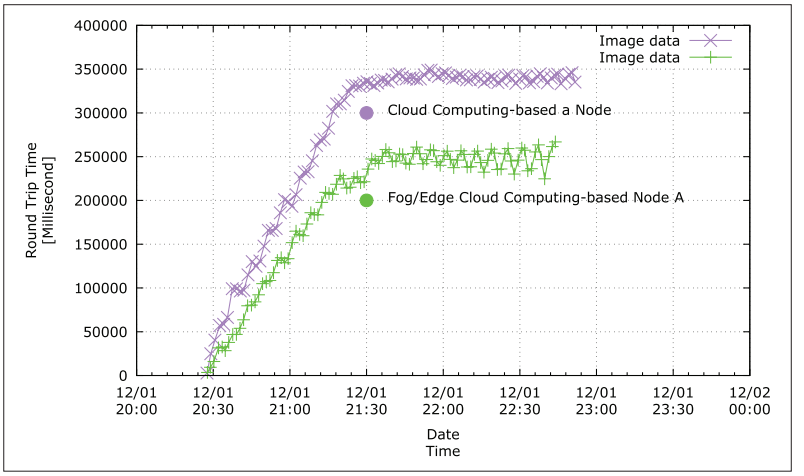

Figure 6. Machine learning latency.

Acknowledgment

This work was supported by the Institute for Information & Communications Technology Promotion(IITP) grant funded by the Korea government (MSIT) (No.2018-0-01456). This work was also supported in part by the Academy of Finland 6Genesis project under Grant No. 318927. Prof. Song and Prof. Taleb are co-corresponding authors of this work.

References

[1] S. Sarkar, S. Chatterjee, and S. Misra, "Assessment of the Suitability of Fog Computing in the Context of Internet of Things," *IEEE Trans. Cloud Comp.*, vol. 6, no. 1, Jan 2018, pp. 46–59.

[2] S. J. Russell and P. Norvig, *Artificial Intelligence: A Modern Approach*, Pearson Education Ltd, 2016.

[3] A. V. Dastjerdi *et al.*, "Fog Computing: Principles, Architectures, and Applications," *Internet of Things*, Elsevier, 2016, pp. 61–75.

[4] Y. Yang, "Multi-Tier Computing Networks for Intelligent IoT," *Nature Electronics*, vol. 2, no. 1, 2019, p. 4.

[5] A. V. Dastjerdi and R. Buyya, "Fog Computing: Helping the Internet of Things Realize its Potential," *Computer*, vol. 49, no. 8, Aug. 2016, pp. 112–16.

[6] N. Abbas *et al.*, "Mobile Edge Computing: A Survey," *IEEE Internet of Things J.*, vol. 5, no. 1, 2018, pp. 450–65.

[7] N. Chen *et al.*, "Fog as a Service Technology," *IEEE Commun. Mag.*, vol. 56, no. 11, Nov. 2018, pp. 95–101.

[8] B. Tang *et al.*, "Incorporating Intelligence in Fog Computing for Big Data Analysis in Smart Cities," *IEEE Trans. Ind. Informat.*, vol. 13, no. 5, 2017, pp. 2140–50.

[9] P. O'Donovan *et al.*, "A Fog Computing Industrial Cyber-Physical System for Embedded Low-Latency Machine Learning Industry 4.0 Applications," *Manufacturing Letters*, vol. 15, 2018, pp. 139–42.

[10] Y. He *et al.*, "Software-Defined Networks with Mobile Edge Computing and Caching for Smart Cities: A Big Data Deep Reinforcement Learning Approach," *IEEE Commun. Mag.*, vol. 55, no. 12, Dec. 2017, pp. 31–37.

[11] C. Perera *et al.*, "Fog Computing for Sustainable Smart Cities: A Survey," *ACM Comp. Surveys*, vol. 50, no. 3, 2017, p. 32.

[12] T. Ouyang, Z. Zhou, and X. Chen, "Follow Me At The Edge: Mobility-Aware Dynamic Service Placement for Mobile Edge Computing," *IEEE JSAC*, vol. 36, no. 10, Oct. 2018, pp. 2333–45.

[13] T. Taleb *et al.*, "Mobile Edge Computing Potential in Making Cities Smarter," *IEEE Commun. Mag.*, vol. 55, no. 3, Mar. 2017, pp. 38–43.

[14] A. I. Maarala, X. Su, and J. Riekki, "Semantic Reasoning for Context-Aware Internet of Things Applications," *IEEE Internet of Things J.*, vol. 4, no. 2, 2017, pp. 461–73.

[15] S. Husain *et al.*, "Mobile Edge Computing with Network Resource Slicing for Internet-of-Things," *Proc. 2018 IEEE 4th World Forum Internet of Things*, IEEE, 2018, pp. 1–6.

Biographies

JongGwan An (cftn3212@sju.ac.kr) received his B.Sc. degree in computer science at Woosong University and is currently working toward a Ph.D. degree in computer science at Sejong University. He is an assistant researcher at the Korea Electronics Technology Institute (KETI).

Wenbin Li (wenbin.li@eglobalmark.com) is a research engineer at Easy Global Market (EGM) and received his Ph.D. degree in computer science from the National Institute of Applied Science Lyon with a special focus on the resilient service-oriented computing paradigm in dynamic environments.

Franck Le Gall (franck.le-gall@eglobalmark.com) received his Ph.D. degree from the University of Rennes. He is CEO at EGM, where he is driving company development of advanced testing technologies.

Ernoe Kovac (Ernoe.Kovacs@neclab.eu) holds a Ph.D. from the University of Stuttgart. At NEC Laboratories Europe, he is a senior manager for Cloud Services and Smart Things. His group works on cloud computing, IoT analytics, and context-aware services.

Jaeho Kim (jhkim@keti.re.kr) is a managerial researcher and a team leader in the IoT Platform Research Center at KETI. He received his B.S. and M.S. degrees from Hankuk University of Foreign Studies. He received his Ph.D. in electrical and electronic engineering from Yonsei University.

Tarik Taleb (tarik.taleb@aalto.fi) is currently a professor at the School of Electrical Engineering, Aalto University, Finland. Prior to his current position, he worked as a senior researcher and 3GPP Standards Expert at NEC Europe Ltd, Heidelberg, Germany.

JaeSeung Song (jssong@sejong.ac.kr) is an associate professor in the Computer and Information Security Department at Sejong University. He received a Ph.D. from Imperial College London in the Department of Computing, United Kingdom. He holds B.S. and M.S. degrees in computer science from Sogang University.