Wang, Xiaofei; Ren, Xiaoxu; Qiu, Chao; Cao, Yifan; Taleb, Tarik; Leung, Victor C.M.

Net-in-AI

Published in:
IEEE Network

DOI:
10.1109/MNET.011.2000319

Published: 01/03/2021

Net-in-AI: A Computing-Power Networking Framework with Adaptability, Flexibility and Profitability for Ubiquitous AI

Xiaofei Wang, *Senior Member, IEEE,* Xiaoxu Ren, *Student Member, IEEE,* Chao Qiu, *Member, IEEE,* Yifan Cao, *Student Member, IEEE,* Tarik Taleb, *Senior Member, IEEE,* Victor C.M. Leung, *Life Fellow, IEEE*

***Abstract*—Along with the unprecedented development of artificial intelligence (AI), a considerable number of intelligent applications are universally recognized to significantly facilitate the evolution of anthropogenic activities. The abundant AI computing power is one of the main pillars to fuel the booming of ubiquitous AI applications. As the computing power proliferates to a multitude of network edges, even end devices, the networking function bridges the gap, on the one hand, among ends-edges-clouds, on the other hand, between the multiple AI computing power and the heterogeneous AI requirements. The emerging new opportunities have spawned the deep integration between computing and networking. However, the complete development of the integrated system is under-addressed, including adaptability, flexibility, and profitability. In this paper, we propose a computing-power networking framework for ubiquitous AI by establishing Networking in AI computing-power pool, denoted as Net-in-AI. We design the framework to enable the adaptability for computing-power users, the flexibility for networking, and the profitability for computing-power providers. We then formulate a computing-networking resource allocation problem, with the joint perspective of these three aspects. Experimental results prove the superior performance of the proposed framework in comparison to the current popular schemes.**

***Index Terms*—Computing-networking integration, artificial intelligence, adaptability, flexibility, profitability.**

I. INTRODUCTION

The past few years have witnessed the unparalleled development of artificial intelligence (AI). Coupled with the rise of AI, a number of intelligent applications are universally recognized to significantly propel the social evolution, such as speech recognition, natural language generation, and virtual agents, etc. The abundant computing power, as one of the main pillars of AI, fuels the booming of ubiquitous AI applications [1]. As such, the burgeoning AI computing technology and framework, featured by heterogeneous, accelerated and programmable computing, open up countless possibilities for meaningful research and applications of AI.

Before the emergence of fifth-generation (5G) networks, mega-scale data were generally generated and cached at a single or a few data centers, which spawned a great deal of traditional cloud-centric approaches for efficient computing and resource management. However, the rising of the beyond 5G (B5G, alternatively termed as 6G) [2] has brought forth zillions of bytes of data to network edges. Consequently, it would accelerate the proliferation of computing power from a few centers to a multitude of network edges, even end devices [3]–[5]. The problem of management and allocation of computing power at a couple of data centers is moderately non-trivial, while for multi-mode and multi-level computing power, it is prohibitively difficult. Additionally, the networking function bridges the gap among end devices, network edges, and clouds, and also, between various computing power and diversified requirements. Obviously, the emerging new opportunities have spawned the deep integration between computing and networking for ubiquitous AI.

However, there are many issues that need to be addressed before integrated computing-networking system becomes a mature technology, including 1) how to provide users with adaptable computing services, so as to satisfy users' diverse requirements. These varied needs of users have opened up intense and adaptive demands of AI computing power; 2) how to support flexible networking service, so as to achieve rapid response. To be specific, networking provides access to multifarious AI services by a multitude of collaborative end devices, edge nodes, and clouds. Such cooperation needs to overcome network congestion, low resource utilization, etc; 3) how to ensure the welfare for computing-power providers, so as to maximize the profitability of the computing-networking system. It is well known that one of the main factors to motivate the wide deployment of such systems is the incentive mechanism and business model. Therefore, it is necessary to deeply consider the incentive to help others perform tasks, via contributing one's own computing power. Overall, *adaptability, flexibility* and *profitability* are three main metrics for the integrated AI computing-networking system.

Blockchain, as the underlying technology of crypto-currencies, has been a relatively recent technological trend. It is an open, cryptographic, and decentralized system, maintaining immutable ledgers that are accessible but tamper-proof for all users. Specifically, blockchain systems are hun-

Xiaofei Wang, Xiaoxu Ren, Chao Qiu and Yifan Cao are are with the College of Intelligence and Computing, Tianjin University, Tianjin, China. E-mails: xiaofeiwang@tju.edu.cn, xiaoxuren@tju.edu.cn, chao.qiu@tju.edu.cn, yifancao@tju.edu.cn.

Tarik Taleb is with the Department of Communications and Networking, School of Electrical Engineering, Aalto University, Finland, with the Department of Computer and Information Security, Sejong University, Seoul, South Korea, and with the Center for Wireless Communications, The University of Oulu, Oulu, Finland. E-mail: tarik.taleb@aalto.fi.

Victor C.M. Leung is with the College of Computer Science and Software Engineering, Shenzhen University, Shenzhen, China, and also with the Department of Electrical and Computer Engineering, The University of British Columbia, Vancouver, Canada. E-mail: vleung@ieee.org.

Corresponding author : Chao Qiu (chao.qiu@tju.edu.cn)

gry for computing power. For example, one of the most commonly used consensus protocols, Proof of Work (PoW), usually requires large amounts of computing power to break a cryptography puzzle through brute-force [6]. As such, some works study efficient consensus mechanisms to replace PoW, including replica-based and learning-based consensus protocols. However, a replica-based consensus protocol, such as Practical Byzantine Fault Tolerance (PBFT), has been proved ineffective in dealing with the problem of resource wastage [7]. In this paper, we utilized Proof of Learning (PoL), one of the learning-based protocols, to replace the cryptography puzzle. PoL treats the process of training neural networks (NNs) as a working puzzle to unlock the potential of sharing more advanced intelligence among edges [8].

Therefore, the incorporation of blockchain into integrated computing-networking systems is synergistically beneficial. For integrated computing-networking systems, blockchain promises several benefits of incentive mechanism, reliable and traceable sharing of computing power. Moreover, the efficient consensus protocol-based blockchain economically consumes computing power to achieve a win-win situation, constructing the blockchain, while training NNs at the same time. For the blockchain, the integrated computing-networking system is functionally necessary due to its large volumes of computing and networking supplement. The main contributions of this article can be summarized as follows.

- We propose a computing-power networking framework for ubiquitous AI by establishing blockchain and multi-level Networking in AI computing-power pool, denoted as Net-in-AI, this framework enables incentive and intelligent computing-networking amalgamation.
- We introduce an AI Computing-Power Allocation Mechanism and consider the adaptability, flexibility and profitability as three metrics to evaluate the performance of the integrated AI computing-networking system.
- We formulate a computing-networking allocation problem, with the joint perspectives of users, networking and computing-power providers. Experimental results prove the superior performance of the proposed framework to the current popular ones.

The rest of this article is organized as follows. The proposed framework, along with a use case of the framework, is described in Section II. Section III then presents the problem formulation of computing-networking allocation. Simulation results are presented and discussed in Section IV. Finally, concluding remarks are given in Section V.

II. Proposed Framework of the Amalgamated System

In this section, we first present the proposed framework, followed by a use case of the proposed framework.

A. *Framework*

The complete development of the Net-in-AI framework is no silver-bullet due to the *Trilemma*, e.g., the adaptive computing demands from users, the flexible networking management requisitions, and the incentive requirements from providers. In most existing works [9], [10], the computing demands, the networking managements and the incentive are studied separately. However, they are all the underlying promoters enabling the integrated system. How to abstract and allocate resources, as well optimize these three problems have significant impacts on the performance of the framework, while the role of blockchain and application scenarios should not be overlooked yet. Therefore, in this article, we propose a computing-power networking framework with adaptability, flexibility, and profitability for ubiquitous AI, as shown in Fig. 1. This framework is composed of the following layers.

1) Infrastructure Layer: The comprehensive constructions of B5G and edge computing are accelerating the proliferation of AI computing power from clouds, to network edges and end devices, characterized by economical mobile broadband, low latency, and high privacy. *End devices*, e.g., monitors and sensors, *network edges*, e.g., base stations and gateways, as well *clouds* are jointly considered in this framework.

Additionally, with the proliferation of the skyrocketing number and types of these ends-edges-clouds computing devices, networking equipment is placed great expectations on bridging the gap among them. The ubiquitous access to these computing devices and computing requirements will be supported by wireless or wired access networks, such as WiFi, smart routers, gateways, base stations, etc.

2) Resource Pooling Layer: In this layer, multi-level computing and ubiquitous networking are abstracted and pooled, where the pooling hypervisor is the top-drawer component. Generally, the pooling hypervisor is responsible for perceiving physical computing and networking resources from the infrastructure layer, while pooling and grouping the scattered resources into computing pools and networking pools, respectively. Due to the fact that the computing power is crowdfunded from decentralized computing providers, the traceable usage of the computing pools will be a major concern. Meanwhile, the reliability and privacy of the networking pools are also considered to be especially necessary for the system.

3) Scheduling Optimization Layer: Different demands placed on the system have various requirements. The demands are grouped into multiple classes according to their computing demands, networking demands, and payment amounts for providers. The class of computing demands involves *'computing-intensive'* demands that require extensive computing power, *'computing-moderate'* demands that need moderate computing power, and *'scavenger'* demands that are not desirable in the system. Similarly, the class of networking demands include *'fast networking'* demands that requires non-congested networking, *'moderate networking'* and *'scavenger'*. Since the use of computing and networking resources in the system is incentive-driven, i.e., pay-to-use, the class of payment amounts can be grouped into *'high cost'*, *'moderate cost'* and *'scavenger'*. For a comprehensive perspective, the whole classes of services are summarised in Table I.

The classified demands will be processed by scheduling optimization algorithms [11], [12], such as reinforcement learning (RL), auction mechanism, convex optimization, etc. The optimized decisions will be made in this layer, while the optimization objectives are to achieve the adaptability for AI

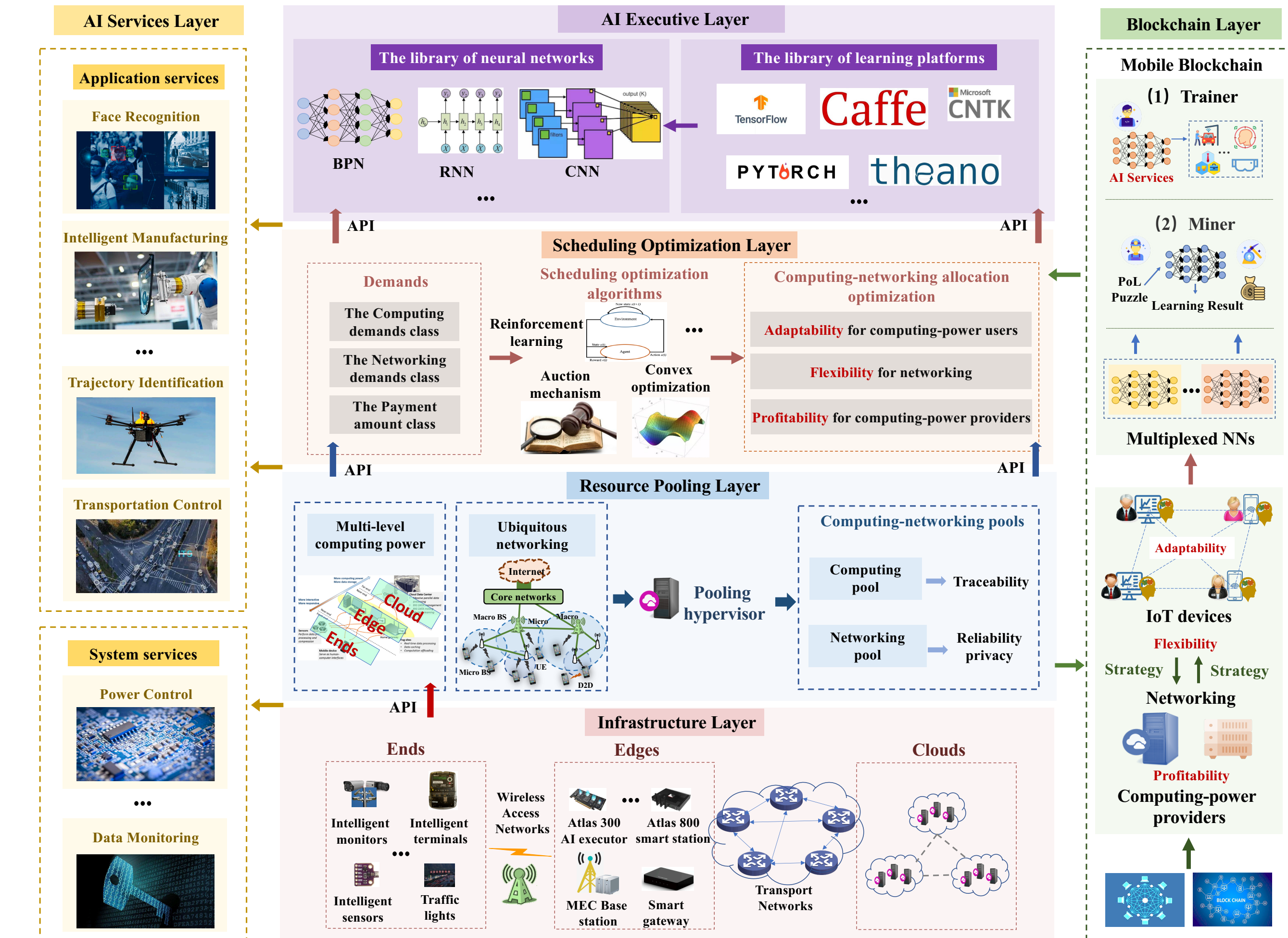


Fig. 1: The Net-in-AI framework.

TABLE I: The classes of service

Classes	Items
The class of computing demands	Computing-intensive
	Computing-moderate
	Scavenger
The class of networking demands	Fast networking
	Moderate networking
	Scavenger
The class of payment amounts	High cost
	Moderate cost
	Scavenger

computing-power users, the flexibility for networking, and the profitability for AI computing-power providers.

4) AI Executive Layer: In order to efficiently accomplish the AI services, the framework could realize the optional and pluggable NNs, as well the learning executive platforms in this layer. According to AI services' requirements, the framework selects the proper NNs, such as text recognition using back propagation networks (BPNs), sequential voice recognition using recurrent neural networks (RNNs), and image recognition using convolution neural networks (CNNs). Furthermore, a multitude of learning platforms are available in this layer, including Tensorflow, Caffe, PyTorch, Theano, CNTK and so on. With the allocated computing-networking resource from the below three layers, the proper NNs, and the learning executive platform, the AI services will be carried out in this layer.

5) Blockchain layer: Currently, the heterogeneous, decentralized and crowd-funded AI computing power from ends-edges-clouds is utilized by users in an uncompensated manner. Therefore, a trusted platform is required to support the reliable management and assure the service credibility for autonomous members in computing-power networking [7]. The framework that integrates blockchain and multi-level networking in AI computing-power pool enables the incentive and intelligent computing-networking amalgamation. Therefore, we introduce the blockchain layer, featuring security, transparency, and decentralization, as a valid solution to impart credibility among network members in a tamper-proof and traceable manner.

Specifically, the resources-constrained users can request computing power from the providers to run mobile blockchain applications and support diversiform intelligent tasks sustained by NNs. Actually, the NNs referred in tasks are multiplexed.

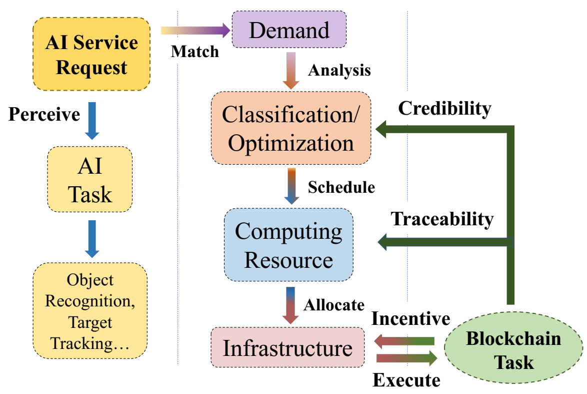


Fig. 2: The workflow of executing an AI task in the Net-in-AI framework

On the one hand, we substitute PoL for cryptography puzzle to overcome the huge waste of computing power, wherein the mining process of PoL is to work out a computational NNs training puzzle. Therefore, a portion of NNs could be regarded as the learning-based training puzzle for miners to compete rewards from blockchain. On the other hand, NNs could be taken as the underlying support for trainers to execute the deep learning algorithms. As such, the NNs in computing-power networking aided by blockchain could be recycled, which provides the powerful support for the sustainable development of computing-networking amalgamation. We will cover the above task model detailedly in the section III-B. Furthermore, in the computing-power transaction process, all the network members are concerned with satisfying their own demands, while the incorporation of blockchain into the computing-networking systems will be of great advantage. In general, the blockchain is conducive to access to optimal strategies and obtain the maximum expected benefits for all parties in the Net-in-AI. Meanwhile, the computing-networking amalgamation is of great importance for the blockchain on account of its abundant computing and networking supplement.

6) AI Service Layer: As shown in Fig. 1, AI services could be divided into two portions. The application services may include face recognition, intelligent manufacturing, trajectory identification, transportation control, etc. On the other hand, system services provide system monitoring and control capabilities, such as power control and data monitoring.

B. Use Cases

Short video platforms, like YouTube, are springing up, which causes increasing amounts of AI tasks, such as object recognition, target tracking, etc. Action recognition is regarded as one tough task in massive data annotation. We set this AI task as an example to show the workflow of the proposed framework.

As shown in Fig. 2, when a user issues an AI service request, the Net-in-AI framework perceives the type of AI task, such as action recognition of a video. The NNs and the learning platforms are then selected in the AI executive layer, ready to match the demand of the AI task. After that, the scheduling optimization layer classifies the requirements of computing, networking, and payment, as shown in Table II, and then deploys the optimization algorithm for scheduling computing and networking resource in the resource pooling layer. Afterwards, the multi-level AI computing power and ubiquitous networking resources are allocated to the computing nodes in the infrastructure layer to execute the specific AI task. Finally, the computing-power networking accomplishes the AI task execution. Additionally, the infrastructures also solve the puzzles in the blockchain layer. The intrinsic features of the blockchain layer, such as incentive, traceability, and credibility, could encourage more facilities to join in the Net-in-AI framework, and keep the framework running in a reliable and efficient manner.

III. Problem Formulation

This section describes the integrated computing-networking system and formulates the AI computing resources management problem, while developing an efficient *AI Computing-Power Allocation Mechanism* for Net-in-AI with heterogeneous resources and multiple demands.

A. System Descriptions

Net-in-AI is a framework, which forms the computing-power into a network. Moreover, it's committed to sharing the AI computing power by flexibly supporting network services, helping computing resources providers to gain profit while dynamically adapting to the multifarious customization needs of users. As shown in the bottom layer of Fig. 1, we consider a set of mobile devices, denoted as $i \in N = \{1, 2, \ldots, n\}$, locate in the vicinity of their corresponding wireless access points connected to the edge nodes. The access points are connected to each other via wired links that also provide access to the cloud servers. The computing nodes, recorded as $j \in M = \{1, 2, \ldots, m\}$, share the computing power by Net-in-AI to perform the multifarious task $l \in K = \{1, 2, \ldots, k\}$. These diversiform infrastructures compose the resource pool to enable computing-power networking to possess a wealth of AI computing power. Furthermore, the above AI computing power information and networking status would be stored in the blockchain as the block information. Accordingly, the Net-in-AI framework could share the computing message and allocate befitting computing power to the requested nodes for providing better service and support.

B. Task Model

In the Net-in-AI, we would like to establish networking in AI computing-power pool to achieve three main factors of the integrated AI computing-networking system, i.e., adaptability, flexibility and profitability. Specifically, the networking includes blockchain and multi-level networking, wherein the blockchain networking is levaraged to record and trade AI computing resources while the multi-level networking is used to schedule and share AI computing power. Spontaneously, the errand of the mobile devices is mainly concentrated in two classes: AI computing power for mining and AI computing power for service, as shown in Fig. 3.

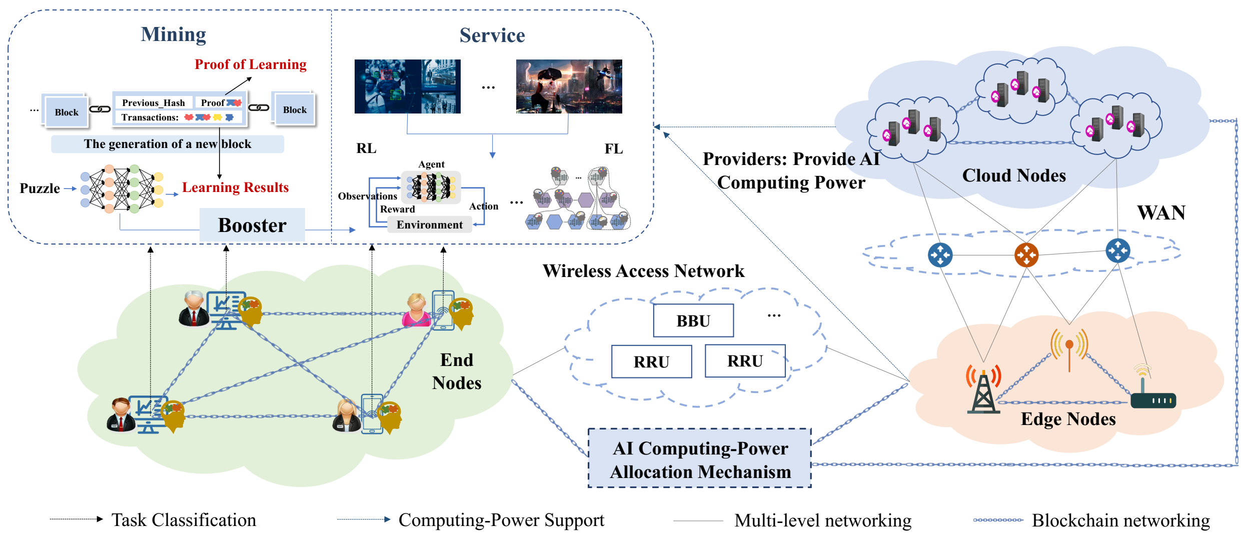


Fig. 3: The task model and its networking supports in the Net-in-AI framework

1) AI Computing Power for Mining: In this paper, we resort to the PoL consensus mechanism [8], rather than the meaningless cryptography puzzle. The PoL consensus mechanism, inspired by machine learning competitions, involves three types of actors: suppliers, trainers and validators. Specifically, suppliers publish machine learning competitions, trainers are in charge of training and submitting models for released tasks, while validators evaluate the models, reach consensus, and propose new blocks to the chain. Actually, the mining process of PoL is to work out a computational NNs training puzzle.

2) AI Computing Power for Service: In addition to the mining task, a mobile device demands intensive computation and high energy consumption to support the intrinsic intelligent services, such as speech recognition, face recognition, natural language processing, and augmented reality, sustained by NNs, RL, federated learning (FL) algorithms, etc. Certainly, these algorithms can take advantage of the emerging training results from the mining task as a booster of the whole AI training process.

From the foregoing, the NNs could be multiplexed. It is worth noting that, the training model of an AI task could be divided and partitioned by some segmentation methods in the horizontal partition or vertical partition manner [3]. With approaches like these, we could regard one partition as the NNs training puzzle and consider others partitions as the latent prop of the deep learning algorithms. That is, PoL not only can be used to reach agreement on a single data block among multiple computing nodes, but accelerates the training of NNs for a better service. Moreover, the computing end nodes would execute aforementioned multifarious tasks while some other computing nodes having excess computing power are idle, which causes computing-power supply contradiction. Therefore, as shown in Fig. 3, it is indispensable to devise a mechanism for AI computing-power dispatch and allocation. This mechanism would, in effect, agilely invoke computing power from the providers, i.e., cloud and edge nodes, to help out with computation-intensive tasks of mobile devices.

C. AI Computing-Power Allocation Mechanism

The intrinsic nature of non-shared computing power at the traditional network creates a competitive and enclosed environment for the mobile users. Hence, we design an *AI Computing-Power Allocation Mechanism* for transparency, invigorative and shared resource allocation in the Net-in-AI. This mechanism would optionally allocate the computing power and adaptively define the computing resource unit price according to distinct computing capabilities, customized AI service demands, and networking status. More concretely, the *AI Computing-Power Allocation Mechanism* is shown as Fig. 4, where AI computing power is sold by the providers and users submit price-bids to the AI computing-power distributor for buying a certain amount of computing power. In the distributor's executive process, it collects a group of AI requests from the mobile users, including submitted bids and customized demands. Furthermore, it would allocate befitting computing power to the solicited users and proclaim the corresponding prices that they need to pay.

In this allocation mechanism, mobile users compete for AI computing resources from providers to support these multiple business requirements. Moreover, the networking should not only render the required communications between users and providers, but cooperatively provision computing power in the Net-in-AI by elastic deployment. To manage the fluctuations of user requirements in Net-in-AI while taking into consideration of the available resources in the computing nodes and networking management, the allocation mechanism should satisfy some diversified demands from the following three perspectives as shown in Fig. 5.

*1) **The Adaptability from User Perspective**:* Firstly, this allocation mechanism should adaptively satisfy the various quality-of-service (QoS) requirements from the user perspective. The QoS requirements include delay requirement, ultra-reliable transmission, reward-preferred, safety autonomy, etc. As some computing nodes are located in close physical proximity to some users, the performance of Net-in-AI is highly affected by the variability of user's requirements. Moreover, it

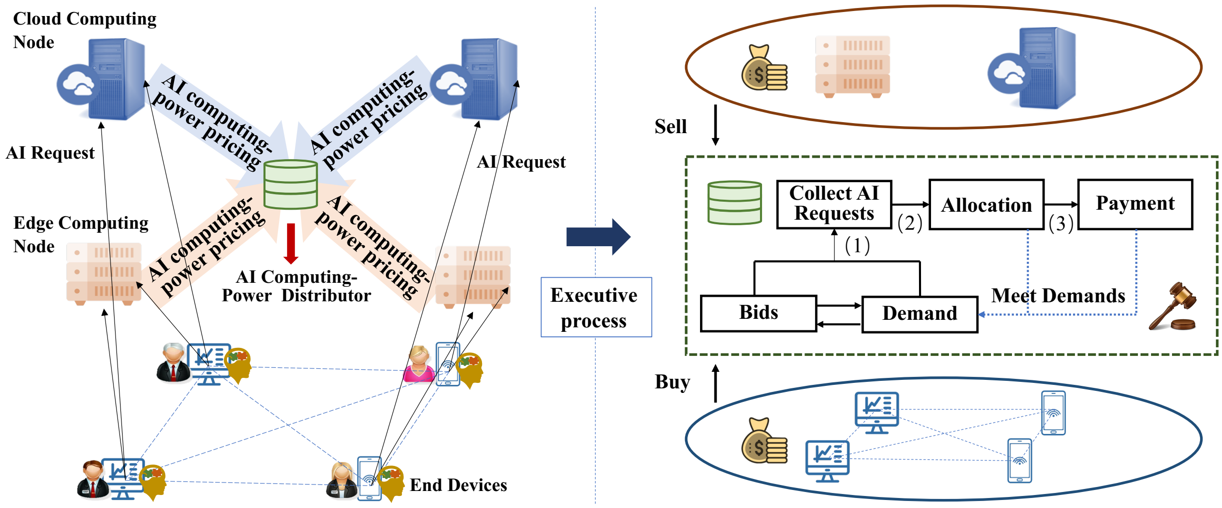


Fig. 4: The *AI Computing-Power Allocation Mechanism*

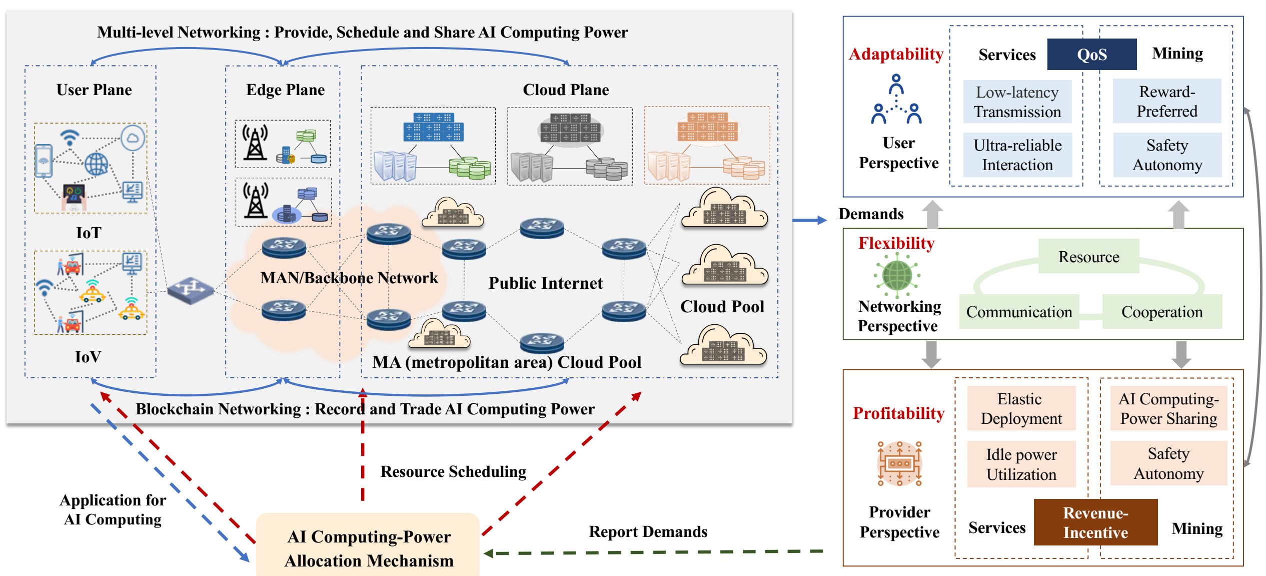


Fig. 5: The demand of *AI Computing-Power Allocation Mechanism*

can be more sensitive to service migration and mining, which may cause wastage of resources and information leakage. As such, for the users, they want to acquire greater utility while satisfying the diversified QoS demands through the adaptability of computing services. In the computing-power allocation process, the device would consume computation time, and the cost that the user has to pay to the provider. Meanwhile, in the blockchain environment, the first miner who successfully obtains the solution of the PoL and reaches the consensus would receive the mining reward. We assume that users have quasi-linear utilities, while the user utility gained in the Net-in-AI is given by the reward due to mining minus the intrinsic computation time and payment for competing resources from the providers. In this setting, the mechanism decides how to allocate computing power to users to maximize their average utility, while adaptively satisfying the multifold computing services, which can be formulated as:

$$
\begin{aligned}
& P1: \max \quad \textbf{\textit{Average Utility}} \\
\textbf{s.t.}\ & C_u^1: Time_l \le Deadline, \ \forall\, l \\
& C_u^2: \delta \cdot \sum_{l \in K} Security_{l,j} \le Security_j, \ \forall\, j \\
& C_u^3: \sum_{j \in M} Bid_{i,j} \le Budget_i, \ \forall\, i
\end{aligned} \tag{1}
$$

where C_u^1 considers the task delay as one of QoS constraints. It means that each accommodated task must be completed within the specified deadline determined by its AI application. The completion time $Time_l$ consists of transmission time between computing nodes, the execution time of task and the queuing delay before processing. Except for the delay constraint, another QoS requirements of tasks is the security requirement C_u^2, i.e., the computing node that performs the assignments must be sufficiently secure and reliable, so as to meet the specified security requirement of the allocated tasks within the security strength of the node [13]. The binary variable δ represents whether the task is selected to be distributed to the computing node. Beyond that, constraint C_u^3 ensures that the total bids submitted by the user i to the provider j could not exceed the available budget of the user. Therefore, the proposed mechanism can provide users with the adaptability in computing services, and further receive fairly good utility.

*2) **The Flexibility from Networking Perspective**:* Infrastructures of AI networks are going through a radical shift from the network infrastructure with information transmission to an intelligent infrastructure integrating perception, transmission, storage, computation and processing, which brings forward greater flexibility requirements for networking management [14]. More specifically, the Net-in-AI framework renders networking of diverse services with the flexibility to capture the business demands of users, and set up dynamically on-demand connections across the network between data and services. The networking in Net-in-AI also refers the tasks could be flexibly offloaded to the most appropriate computing nodes, so as to achieve rapid responses. Although the proposed framework enables convenient access to the ample heterogeneous pools of computing resources, migrating the computation intensive tasks from end devices to the computing nodes could induce network congestion and irrational use of the computing sources. It may further incur long network delay and low resource utilization. We then introduce the network congestion index that is defined as the ratio between network waiting time and resource utilization. Therefore, from the networking perspective, we aim to minimize the network congestion index to relieve pressure on networking and offer flexibility in networking service, which can be described by:

$$\begin{aligned} &P2: \min \quad \textbf{\textit{Network Congestion Index}} \\ \textbf{s.t. } &C_n^1: Time_l \leq Deadline, \ \forall\, l \\ &C_n^2: Average\ Utility_i \geq 0, \ \forall\, i \\ &C_n^3: Welfare_j \geq 0, \ \forall\, j \end{aligned} \tag{2}$$

where the constraints include the delay requirement C_n^1 and the quantitative index restraints from the user and provider, e.g. C_n^2, C_n^3. Hence, from the perspective of networking, the computing-power allocation mechanism could lessen the amount of information that flows in the network, while flexibly providing native physical and virtual mobility for supporting mobile AI applications in a cost-efficient manner.

*3) **The Profitability from Provider Perspective**:* An effective incentive mechanism based on profitability should enable the provider to obtain the maximum benefit, while stimulate more providers to participate in the Net-in-AI. Actually, the mobile devices will compete for computing power derived from different types of computing-power providers in the resource-constrained mobile environment. Specifically, user i requests a bundle of AI computing-power units and submits a bid for mining or service. The requests collected from the users are submitted to the computing-power distributor that determines the allocation of computing power to devices and the prices the devices have to pay the providers. From the provider's perspective, the service computing nodes that provide the computing-power units consist of a set of computing and communications facilities that consume electric power to perform the tasks. Meanwhile, they would gain the revenue due to serving the user demands. Assume the provider who furnishes computing power to users does not attend the mining task. Then, the provider welfare gained in the computing-power networking is given by the payment from users minus the electricity cost of executing the tasks. In this setting, the allocation mechanism decides how to allocate the AI computing power to users to maximize the computing-power provider welfare, which can be calculated by:

$$\begin{aligned} &P3: \max \quad \textbf{\textit{Welfare}} \\ \textbf{s.t. } &C_p^1: \sum_{i \in N} Capacity_{i,j} \leq Capacity_j, \ \forall\, j \\ &C_p^2: Payment_{i,j} \leq Bid_{i,j}, \ \forall\, i, j \end{aligned} \tag{3}$$

where the constraint C_p^1 ensures the total allocated capacity, e.g., computing power, to the user i from the provider j could not exceed the available capacity of the provider. The second constraint ensures that the submitted bids from the user i to the provider j is no less than the corresponding payment. Thereby, from the perspective of provider, this computing-power allocation mechanism could motivate more providers participate in the Net-in-AI and maximize the profitability of the computing-networking system.

Responding to the ever-increasing computing-power demands from AI applications, we establish the computing sharing problems of three perspectives from user, networking and provider for the blockchain-assisted Net-in-AI framework. These three problems are closely related with and interacted on each other. In our model, the *Average Utility* and *Welfare* are negatively correlated with the *Network Congestion Index*, which makes it possible for one of adaptability or profitability and flexibility to be simultaneously satisfied. Furthermore, the optimization problems between user perspective and provider perspective could be regarded as a two stages Stackelberg game, and the Nash equilibrium exists in this strategic game [15]. As such, the computing-power networking can synchronously possess the adaptability, flexibility and profitability and meet the needs of the network members. By addressing the above problems, the computing-networking system enables the computing functions into a novel networking system, which mainly assists in the deployment and management of computing resources. Additionally, it can also empower users to nearby access the network and realize the load balancing of services, while helping massive applications and massive computational resources form an open ecosystem. Furthermore, in this considered framework, the potential incentive mechanism would motivate more members to join the computing-power networking to further acquire the interests, thereby achieving multi-win.

IV. Simulation Results and Discussions

In this section, we present an extensive experimental analysis to evaluate the performance of the Net-in-AI framework with respect to the three metrics defined in Section III.

A. Experiment Settings

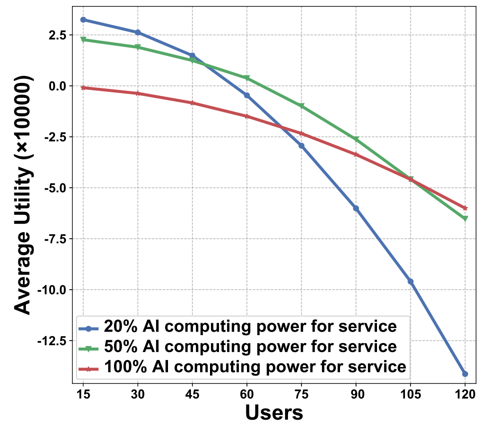
In our example scenario, disparate AI computing resources come from the end devices that possess low computing capabilities, and eight geographically distributed edge nodes and one cloud having abundant AI computing-power units. To investigate the capabilities of the Net-in-AI framework for computation power allocation, we suppose the WLAN bandwidth $w_1 = 200$ Mbits/s, WAN bandwidth $w_2 = 30$ Mbits/s. The computing power of the cloud node is set as $f_c = 2 \times 10^6$ MIPS, while the computing power of the edge node is set as $f_f = 8 \times 10^5$ MIPS and the computing power of the mobile device is $f_d = 2 \times 10^4$ MIPS. This paper aims at making computing-power networking achieve the adaptability, flexibility and profitability objectives simultaneously. Particularly, we employ the Greedy algorithm to address the optimization problems based on P1 and P3 and leverage the simulated annealing (SA) algorithm to solve the optimization problem P2.

B. Performance Analysis

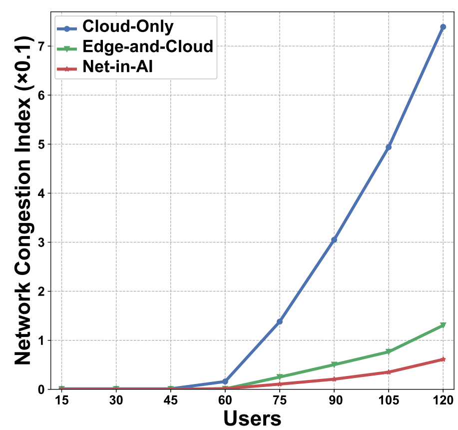
To elucidate the performance of the proposed framework, experiments on computing-power allocation mechanism are carried out under the following three perspectives.

1) User Perspective: In our proposal, the tasks of computing nodes include service and mining. Indeed, it is important to consider the proportion of computing power for service on users' utility. As shown in Fig. 6(a), for the user $i \in [15 : 120]$, the utility of user is negative without mining. Simultaneously, the expected utility for most users who performed mining has the potential for the greatest volatility. For the users $i \in [15 : 45]$, the average utility would be better when they resolve to employ 20% computing power for service due to the fact that mining might fetch more rewards. Nevertheless, the average utility may result in poor performance for large-scale users in the same ratio. That is because concentrating most of the resources on mining results in a high risk of revenue loss. When more users join computing-power networking to compete for computing resources, the probability that a user successfully solves the PoL puzzle and reaps the reward would be decreased. The successful miners would get more rewards, while the aborted miners would suffer worse returns than users dedicated to service on account of the waste of computing power in mining. Therefore, in the case of a large user population, the more computing power is used for service, the better the average utility. Meanwhile, with the increase of the mobile consumers, the users' average utility is reduced owing to the lower computing power of these users.

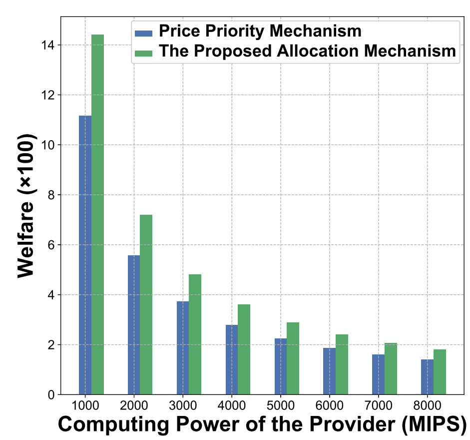
2) Network Perspective: Fig. 6(b) compares the average congestion index in different computing-power scheduling schemes. From Fig. 6(b), we can see that there is a modest gap between the three schemes when the user number is between 15 and 60. That, the result goes, is because the small-scale data transmission task will not give rise to serious network congestion. With an increasing number of users, the average congestion index of each of the different schemes shows an upward tendency. It is clear that our proposed framework yields the better performance compared to the



(a) Comparisons on average utility with different weight coefficients for service



(b) The network congestion index in the different architectures



(c) The welfare of the provider in different price mechanisms

Fig. 6: Performance evaluation of the Net-in-AI

other schemes, especially with a large number of users. This because the networking scheme prefers to perform tasks in a cooperative and flexible manner, and the Net-in-AI could schedule the computing power of the computing nodes optimally. This causes the average waiting delays to diminish while the resource utilization increases on account of the computing convenience of devices, resulting in a lower average congestion index.

3) Provider Perspective: Aforementioned experiments analyzes performance metrics from the perspective of users and networking. Here, we examine the impact of distinct price mechanisms on the welfare of the providers. In Fig. 6(c), the *Price Priority Mechanism* signifies that each user bids some amount of money for the AI computing power, with a higher each user bid yielding computing power. In our allocation mechanism, the computing-power distributor declares a computing-power unit price to each user, then the users decide the bids in line with their demands. From Fig. 6(c), it can be observed that the provider welfare obtained by adopting the proposed mechanism is always higher than that under the *Price Priority Mechanism*. Moreover, the provider welfare degrades when it possesses abundant computing power. This is due to the fact that the submitted prices, i.e., the bids of users, are inversely proportional to the computing capacity of providers. Specifically, as computing power grows, the intense competition for it will ease, which engenders the unit price of computing power falls off, leading to further decrease in the welfare of the provider.

V. Conclusions and Future Work

In this article, we have presented the Net-in-AI framework of computing-power networking for ubiquitous AI, aided by blockchain. We have shown the potential benefits of this framework in adaptability for computing-power users, the flexibility for networking, and the profitability for computing-power providers. We have formulated the computing-networking allocation problems, with the joint consideration of these three aspects. Experimental results have confirmed the effectiveness of Net-in-AI.

Research on integrated computing-networking system is still in its infancy, and there are still some unexplored problems. For instance, a dynamically changing networking environment seriously affects the quantitative indexes of members in the computing-networking system. Therefore, a more general quantitative indexes of members need to be further defined. Moreover, the security issues of computing-networking system may exist. Tight interoperation between blockchain and computing-power networking is an interesting topic as well.

VI. Acknowledgment

This work was supported in part by the National Key Research and Development Program of China under Grant 2018YFC0809803 and 2019YFB2101901; the National Science Foundation of China under Grant 62072332, the China Postdoctoral Science Foundation under Grant 2020M670654; the National Natural Science Youth Foundation of China under Grant 62002260. This work was also partially supported by the European Unions Horizon 2020 Research and Innovation Program through the MonB5G Project under Grant No. 871780; the Academy of Finland 6Genesis project under Grant No. 318927 and the Academy of Finland CSN project under Grant No. 311654.

References

[1] Y. Dai, D. Xu, S. Maharjan, Z. Chen, Q. He, and Y. Zhang, "Blockchain and deep reinforcement learning empowered intelligent 5G beyond," *IEEE Net.*, vol. 33, no. 3, pp. 10–17, 2019.

[2] W. Saad, M. Bennis, and M. Chen, "A vision of 6G wireless systems: applications, trends, technologies, and open research problems," *IEEE Net.*, pp. 1–9, 2019.

[3] Z. Zhou, X. Chen, E. Li, L. Zeng, K. Luo, and J. Zhang, "Edge intelligence: Paving the last mile of artificial intelligence with edge computing," *Proceedings of the IEEE*, vol. 107, no. 8, pp. 1738–1762, 2019.

[4] J. Ren, G. Yu, Y. He, and G. Y. Li, "Collaborative cloud and edge computing for latency minimization," *IEEE Trans. on Veh. Tech.*, vol. 68, no. 5, pp. 5031–5044, 2019.

[5] K. Zhang, Y. Zhu, S. Maharjan, and Y. Zhang, "Edge intelligence and blockchain empowered 5g beyond for the industrial internet of things," *IEEE Network*, vol. 33, no. 5, pp. 12–19, 2019.

[6] M. Conti, E. Sandeep Kumar, C. Lal, and S. Ruj, "A survey on security and privacy issues of bitcoin," *IEEE Commu. Surveys Tutorials*, vol. 20, no. 4, pp. 3416–3452, 2018.

[7] S. Guo, X. Hu, S. Guo, X. Qiu, and F. Qi, "Blockchain meets edge computing: A distributed and trusted authentication system," *IEEE Trans. on Industrial Infor.*, vol. 16, no. 3, pp. 1972–1983, 2020.

[8] F. Bravo-Marquez, S. Reeves, and M. Ugarte, "Proof-of-learning: A blockchain consensus mechanism based on machine learning competitions," in *Proc. Conf. on Decentralized Applications and Infrastructures (DAPPCON)*, 2019, pp. 119–124.

[9] M. Król, S. Mastorakis, D. Oran, and D. Kutscher, "Compute first networking: Distributed computing meets ICN," in *Proc. Conf. on Infor. Centric Net., ACM*, 2019, pp. 67–77.

[10] D. Grewe, N. Nayak, U. Ambalavanan, and S. Schildt, "Towards in-network computing infrastructures for connected vehicles," 2020.

[11] C. Qiu, F. R. Yu, H. Yao, C. Jiang, F. Xu, and C. Zhao, "Blockchain-based software-defined industrial internet of things: A dueling deep Q -learning approach," *IEEE Internet of Things Journal*, vol. 6, no. 3, pp. 4627–4639, Jun. 2019.

[12] Z. Xiong, S. Feng, W. Wang, D. Niyato, P. Wang, and Z. Han, "Cloud/fog computing resource management and pricing for blockchain networks," *IEEE Internet of Things Journal*, vol. 6, no. 3, pp. 4585–4600, 2019.

[13] S. S. Yau, Y. Yin, and H. An, "An adaptive approach to optimizing tradeoff between service performance and security in service-based systems," *International Journal of Web Services Research (IJWSR)*, vol. 8, no. 2, pp. 74–91, 2011.

[14] C. Wang, M. D. Renzo, S. Stanczak, S. Wang, and E. G. Larsson, "Artificial intelligence enabled wireless networking for 5G and beyond: Recent advances and future challenges," *IEEE Wireless Commu.*, vol. 27, no. 1, pp. 16–23, 2020.

[15] H. Yao, T. Mai, J. Wang, Z. Ji, C. Jiang, and Y. Qian, "Resource trading in blockchain-based industrial internet of things," *IEEE Transactions on Industrial Informatics*, vol. 15, no. 6, pp. 3602–3609, 2019.

Xiaofei Wang [S'06, M'13, SM'18] (xiaofeiwang@tju.edu.cn) is currently a professor at Tianjin University, China. He got master and doctor degrees in Seoul National University from 2006 to 2013, and was a Post-Doctoral Fellow with The University of British Columbia from 2014 to 2016. He was a recipient of the National Thousand Talents Plan (Youth) of China. In 2017, he received the "Fred W. Ellersick Prize" from the IEEE Communication Society. His current research interests include social-aware cloud computing, cooperative cell caching, and mobile traffic offloading.

Xiaoxu Ren [S'20] (xiaoxuren@tju.edu.cn) is currently pursuing the Ph.D. degree in College of Intelligence and Computing, Tianjin University, Tianjin, China. She received the B.S. degree in College of Science from Inner Mongolia University of Technology, China in 2016. Her current research interests include machine learning, computing power networking, and blockchain.

Chao Qiu [S'15, M'19] (chao.qiu@tju.edu.cn) is currently a Lecturer in the School of Computer Science and Technology, College of Intelligence and Computing, Tianjin University. She received the B.S. degree from China Agricultural University in 2013 in Communication Engineering and the Ph. D. from Beijing University of Posts and Telecommunications in 2019 in Information and Communication Engineering. From September 2017 to September 2018, she visited Carleton University, Ottawa, ON, Canada, as a Visiting Scholar. Her current research interests include machine learning, computing power networking and blockchain.

Yifan Cao [S'20] (yifancao@tju.edu.cn) is currently pursuing the M.S. degree in College of Intelligence and Computing, Tianjin University, Tianjin, China. He received the B.S. degree in School of Electrical and Information Engineering from Tianjin University, China in 2020. His current research interests include machine learning, edge computing, and blockchain.

Tarik Taleb [S'05, M'05, SM'10] (Tarik.Taleb@aalto.fi) received the B.E. degree (with distinction) in information engineering in 2001, and the M.Sc. and Ph.D. degrees in information sciences from Tohoku University, Sendai, Japan, in 2003 and 2005, respectively. He is currently a professor with the School of Electrical Engineering, Aalto University, Espoo, Finland. He is the founder and Director of the MOSA!C Lab (http://www.mosaic-lab.org/). He is the Guest Editor-in-Chief for the IEEE JSAC Series on Network Softwarization and Enablers.

Victor C. M. Leung [S'75, M'89, SM'97, F'03, LF'20] (vleung@ieee.org) is a Distinguished Professor of Computer Science and Software Engineering at Shenzhen University. He is also an Emeritus Professor of Electrical and Computer Engineering and the Director of the Laboratory for Wireless Networks and Mobile Systems at the University of British Columbia (UBC). His research is in the areas of wireless networks and mobile systems. Dr. Leung is a Fellow of the Royal Society of Canada, a Fellow of the Canadian Academy of Engineering, and a Fellow of the Engineering Institute of Canada.