The Many Faces of Edge Intelligence

Peltonen, Ella; Ahmad, Ijaz; Aral, Atakan; Capobianco, Michele; Ding, Aaron Yi; Gil-Castineira, Felipe; Gilman, Ekaterina; Harjula, Erkki; Mohan, Nitinder; More Authors

DOI
10.1109/ACCESS.2022.3210584

Publication date
2022

Document Version
Final published version

Published in
IEEE Access

IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received 30 August 2022, accepted 22 September 2022, date of publication 28 September 2022, date of current version 7 October 2022.

Digital Object Identifier 10.1109/ACCESS.2022.3210584

PERSPECTIVE

The Many Faces of Edge Intelligence

ELLA PELTONEN[1], (Member, IEEE), IJAZ AHMAD[2], (Member, IEEE), ATAKAN ARAL[3], (Member, IEEE), MICHELE CAPOBIANCO[4], AARON YI DING[5], (Member, IEEE), FELIPE GIL-CASTIÑEIRA[6], EKATERINA GILMAN[1], (Member, IEEE), ERKKI HARJULA[1], (Member, IEEE), MARKO JURMU[2], TEEMU KARVONEN[1], MARKUS KELANTI[1], TEEMU LEPPÄNEN[7], (Senior Member, IEEE), LAURI LOVÉN[1], (Senior Member, IEEE), TOMMI MIKKONEN[8], NITINDER MOHAN[9], PETTERI NURMI[8], SUSANNA PIRTTIKANGAS[1], (Member, IEEE), PAWEŁ SROKA[10], (Member, IEEE), SASU TARKOMA[1,8], (Senior Member, IEEE), AND TINGTING YANG[11], (Member, IEEE)

[1]Faculty of Information Technology and Electrical Engineering, University of Oulu, 90570 Oulu, Finland
[2]VTT Technical Research Centre of Finland Ltd., 90570 Espoo, Finland
[3]Faculty of Computer Science, University of Vienna, 1010 Wien, Austria
[4]Business Innovation Manager, Pordenone, Italy
[5]Department of Engineering Systems and Services, TU Delft, 2628 Delft, The Netherlands
[6]Enxeñaría telemática, University of Vigo, 36310 Vigo, Spain
[7]Information Technology, Oulu University of Applied Sciences, 90570 Oulu, Finland
[8]Department of Computer Science, University of Helsinki, 00100 Helsinki, Finland
[9]Connected Mobility, Technical University of Munich, 80333 München, Germany
[10]Institute of Radiocommunications, Poznan University of Technology, 60-965 Poznań, Poland
[11]Pengcheng Laboratory, Shenzhen 518066, China

Corresponding author: Ella Peltonen (ella.peltonen@oulu.fi)

This paper has been written by an international expert group, led by the 6G Flagship at the University of Oulu, Finland (AoF grants 318927, 326291, 323630; Infotech Oulu grants B-TEA, TrustedMaaS). This work is partially supported by the European Union's Horizon 2020 research and innovation programme under the grant agreement No. 101021808 and Marie Skłodowska-Curie grant agreement No. 956090, National Science Centre in Poland (grant 2018/29/B/ST7/01241), Austrian Science Fund (FWF grants Y 904-N31, I 5201-N), CHIST- ERA (grant CHIST-ERA-19-CES-005) and the City of Vienna (5G Use Case Challenge InTraSafEd 5G).

ABSTRACT *Edge Intelligence* (EI) is an emerging computing and communication paradigm that enables Artificial Intelligence (AI) functionality at the network edge. In this article, we highlight EI as an emerging and important field of research, discuss the state of research, analyze research gaps and highlight important research challenges with the objective of serving as a catalyst for research and innovation in this emerging area. We take a multidisciplinary view to reflect on the current research in AI, edge computing, and communication technologies, and we analyze how EI reflects on existing research in these fields. We also introduce representative examples of application areas that benefit from, or even demand the use of EI.

INDEX TERMS Edge intelligence, edge computing, 5G, 6G.

I. INTRODUCTION

Edge Intelligence (EI) is an emerging computing paradigm that enables AI functionalities at the network edge to better serve the needs of increasingly intelligent and autonomous *connected objects*, *connected systems*, and *connected services* [1], [2], [3], [4], [5]. EI builds on the development of powerful AI solutions and the emergence of edge computing as a paradigm that augments computer networks by bringing storage, computing, and other functionality close to the devices that need them. The combination of these developments, as sought by EI, is challenging the dominant centralized cloud-based view of AI by allowing intelligence – or at least some parts of it – to be placed close to the services, applications, and data sources that would require or benefit from it, and by overcoming the limitations of the cloud for many critical applications [6].

Besides challenging the current cloud-based view on AI, EI brings additional benefits that enable new types of applications, a new generation of services, and opportunities

The associate editor coordinating the review of this manuscript and approving it for publication was Derek Abbott.

for other innovations [7], [8]. Having intelligence at the edge *minimizes processing latency*, which is critical for applications with short response-time requirements, such as augmented reality [9] or autonomous vehicles [10], and applications that are characterized by high data velocity, such as real-time visual analytics and medical imaging [11], [12], [13]. Bringing intelligence directly on the network edge can enhance *privacy* by limiting the scope of data disclosure, particularly when distributed models such as federated learning are adopted [14]. Finally, edge computing implies the *decoupling and distribution of application state* and application logic across multiple computing resources, marking a significant shift from today's cloud-centric application development. This, in turn, requires reconsidering *software development practises, principles and processes* to deal with new forms of architectures and enhances flexibility [15].

Prior literature has failed to examine interactions between different components of EI and the challenges these interactions pose. Instead, previous research has examined EI solely from a narrow viewpoint where the focus is on specific AI or edge challenges, on challenges emerging from specific application areas [2], [16], on the implementation in embedded devices or edge platforms [17], [18], [19], [20], or on comparing the cloud and the edge for AI applications [21], [22]. Thus, this article argues for a more holistic understanding of EI, highlighting EI as an emerging new field, discussing the state of research, and identifying its key challenges. We examine EI in a holistic light with the aim of serving as a catalyst for research in EI. We take a multidisciplinary view to reflect on the existing research in AI and edge computing, analyze how EI extends on this research, and identify gaps to establish a research roadmap for the path forward. To summarize, the contributions of this paper are:

- **Synthesis of key challenges** to realize EI, including critical reflection on what there is already implemented in the field of edge computing, and intelligent systems, and how EI goes beyond the state-of-the-art.
- **Research roadmap** of EI, including a critical analysis of what the EI should and could really provide to complement the existing systems, and more critically, how it can enable completely novel applications.
- **Practicality of EI** by identifying representative examples of EI verticals that have already been practically demonstrated and implemented beyond those simply existing as visions.

II. MOTIVATION FOR EDGE INTELLIGENCE

Large-scale uptake of EI requires application scenarios that have sufficient business potential to drive deployment while also posing unique scientific challenges to engage the academic community. Thus far AI scenarios have largely been driven by cloud computing scenarios, such as natural language processing, and computer vision, among others [21]. In contrast, edge computing has mostly operated on scenarios that are characterized by large-volume data streams and the need for low latency, such as real-time video analytics

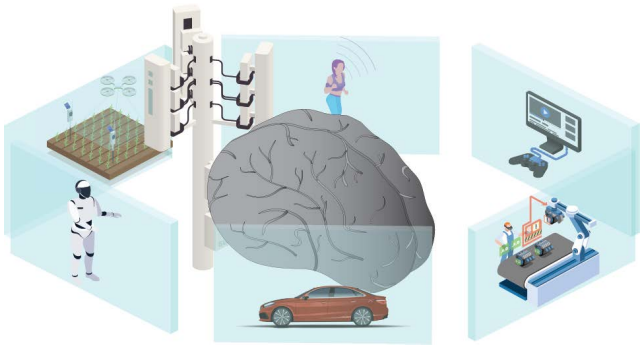


FIGURE 1. Edge Intelligence enables various novel applications.

or cognitive assistance [23], [24]. EI seeks to merge these strands and to harness AI algorithms that are migrated to the edge (instead of the cloud) while offering high bandwidth and low-latency processing and communications [6], [21]. This enables novel applications that involve massive data streams that need to be analyzed and processed in a time-critical, secure, and latency-bounded manner [1], [2], [3], [4]. Representative examples of applications that have already been realized are illustrated in Figure 1 and include managing robotics and vehicles in a spatio-temporally critical environment, distributed manufacturing and logistics, serving users of privacy-critical systems with highly personal data, and so on. Note that these are not intended as an exhaustive list of domains where EI is relevant, rather as diverse examples of applications that have already been deployed in smart factories and in emerging Internet of Things solutions – even if some of the deployments remain highly specific, customized, or rudimentary. Indeed, the use cases for EI stretch beyond these examples, covering societal (e.g., environmental monitoring), commercial (e.g., entertainment, logistics, manufacturing) and governmental use cases (e.g., defense and healthcare) [25]. The envisaged key industry benefit of EI ultimately pertains to all parts of the application chain, covering the algorithms, protocols, enablers, and platform and software engineering methodologies that enable the deployment of data-intensive and low-latency applications across the entire edge-cloud environment. Besides offering increased capabilities for intelligence, EI provides opportunities for innovative applications and services that are impossible to realize without EI. Below we briefly discuss some of these domains, focusing specifically on ones where academic or commercial demonstrations have already been realized. Practical use cases are later presented in Section V.

Manufacturing, smart hospitals, and related data-intensive domains produce large data volumes from a high number of sensors (e.g. manufacturing process monitoring) and data-intensive instrumentation (e.g. PET scanners in hospitals) [26], [27]. The processing of this data requires a high computational capacity and EI can provide the necessary capacity. EI also benefits these domains by driving down hardware costs and the setup complexity.

AR assistance for the person operating the equipment to complete elaborate tasks is a paradigmatic example of domains that benefit from EI. In this domain, the EI application is capable of receiving a real-time video stream containing the environment (set of pieces and tools, their orientation, state, etc.) and the interactions performed by the operator [28]. The application should understand the actions completed by the operator and next steps, providing a tactile visual guidance in real time. Nowadays, it is not possible to equip operators with the required computing power, but EI can offer the necessary intelligence to support this task without violating response-time requirements.

Future traffic systems and connected vehicles are foreseen to take advantage of EI. Examples include applications of extended sensors and remote or fully autonomous driving, that require highly reliable and low-latency data processing and analysis [29]. While some degree of automation can be achieved with in-vehicle processing, more advanced algorithms require computational power and resources that are not available locally. Sensor data collected from cars and passengers is also an essential element of smart traffic management. Excessive network dynamics, latency and reliability constraints hinder efficient management using a centralized approach, whereas using distributed reasoning with EI can accommodate and adapt to these challenges.

Generative Internet is, in our view, a candidate for being the killer application for EI. Indeed, the main impact of EI results from multi-edge and multi-cloud support. In a *generative Internet*, the application logic is generated and provisioned across the communications, computing, and AI infrastructure. Dynamic self-management of the communication network itself is one of the core examples of such a vision for EI as an automatic and intelligent adaptation of the network is fundamental for developing an intelligent Internet that *integrate AI across the Internet*. However, such a intelligent, self-aware Internet is still far from our current state of art, even if early applications of EI move towards directions of self-aware, dynamic applications as discussed above.

Common to all of the application areas discussed above are certain enablers from the edge computing and artificial intelligence. However, simply applying AI into edge – or edge into AI – is not sufficient enough to harness the full presumed capabilities of EI. Indeed, as will be discussed through several objectives in Section IV, EI is more than the sum of its parts (i.e., the combination of edge and AI.)

III. DEFINITIONS

Before discussing the key research challenges and reflecting on the state of the research with respect to these challenges, we briefly describe what we mean when we talk about edge computing, intelligence, and edge intelligence.

A. EDGE COMPUTING

Networking consortia such as ETSI, OpenEdge, and Industrial Internet Consortium (IIC), view the edge as a way to bring additional capabilities closer to devices with some differences in the definitions related to the clients and the networking infrastructure that is expected to be available. For example, ETSI refers to the availability of cloud computing capability at the Radio Access Network of cellular operators [30] whereas IIC sees the edge as the boundary between digital and physical entities that is delineated by IoT devices. Academic definitions in turn, consider the edge as a generic entity that can be seen as a ubiquitous platform, which is not necessarily restricted to specific resource type capability, deployment location, or other characterizing parameters (such as storage, network, and computing capacity).

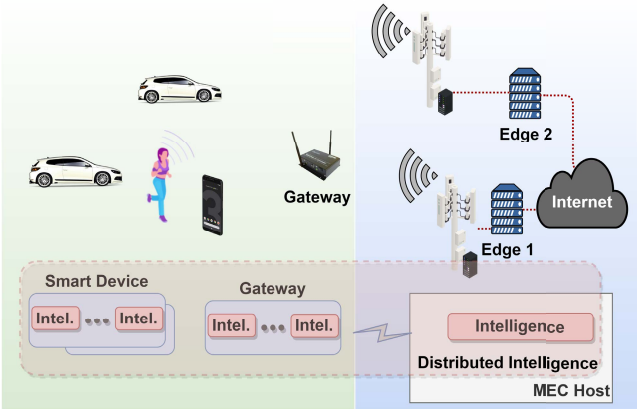


FIGURE 2. EI as a network of intelligent operations and services.

B. INTELLIGENCE

Intelligence is complex to define in a general way and as a result there are hundreds of definitions in the literature. For our purposes, we follow Legg and Hutter [31] who collected different definitions for intelligence and found three common features that characterize intelligence: (i) it belongs to a subject and measures the subject's ability to interact with its environment; (ii) it measures the capability to set and reach objectives; and (iii) it characterizes the ability to adapt behaviour in response to the environment.

C. EDGE INTELLIGENCE

refers to the amalgam of edge computing and intelligence. The definitions for edge computing highlight that EI is supported through a ubiquitous platform that is not restricted by specific resource type constraints while being able to support applications and services, whereas the definitions for intelligence define this platform to be intelligent by being able to optimize its behavior and react to changes in its environment. This definition highlights that EI goes beyond deploying (artificial) intelligence tools on a platform that is used to support applications (intelligence on the edge) and requires that the platform is capable of optimizing behaviour and reacting to changes in its operational environment. It also goes beyond the traditional platforms deployed in predefined locations and evolves towards new distributed architectures

TABLE 1. Key objectives for EI and comparison to edge computing and intelligent systems.

Objective	Currently on Edge	Currently on Intelligence	EI should provide
Never-sleeping systems	Self-recovery of edge servers [32], [33]	Automatic fault-recovery and fault-prediction [34]	Device and network outage resiliency; new KPIs to ensure emerging application QoS and QoE
Latency of experiences	Latency as communication speed [35]	Performance as model-building time [36]	KPIs for full operational loop; low-latency network reconfiguration; latency of milliseconds
Localized intelligence	First steps toward edge-capable AI, federated learning; no generalization over multiple problems [37]	Cloud-centric AI/ML capabilities; lack in real-time demands [1], [38]	Distributed model building, sharing, and cooperation between different application verticals
Where is the edge	Local computational resources; tasks distribution over the network [39]	Cloud computing paradigm [21]	Dynamic collaboration and resource sharing between end-user devices, specific application-domain devices, and cloud services
Ubiquitous resources	Sensing for context-awareness and localized actions [39], [40]	Application composition solutions, ML and reasoning to support the end-users [41], [42]	Enabling self-management properties (e.g. migration, service continuity, application self-healing)
Developer experiences	Virtualization, CI/CD & DevOps, microservices [15], [43], [44]	ML-specific edge frameworks; Cloud-based APIs for ML/AI [40]	Assisting methodologies with analytics and ML/AI for edge development. Dynamic allocation of intelligent components
Integration and interoperability	First attempts in references, data models, protocols and APIs for resource-constrained and distributed architectures [15]	Centralized solutions for management of distributed architectures [21]	Aligning/handling various edge platform solutions to enable platform portability
Privacy, security, and reliability	Decentralized security and failure prevention mechanisms [2], [45], [46]	Decentralized trust management and decision making [45], [47]	Dynamic adaptation into changing situations; intelligent security and privacy prevention; locally adjusted trust management

in which all the involved components collaborate to build ubiquitous intelligence and to provide composed services to users, other devices and applications; see Figure 2.

IV. OBJECTIVES TO REACH EI

Edge intelligence adds *specialized intelligence* and *specialized services* to leverage the current and emerging cloud and local intelligence into a network of intelligent operations and services. In this section, we reflect on prior research to identify challenges, summarized in Table 1, that need to be addressed to fully realize the potential of EI. The analysis was completed by exploring previous publications to identify the challenges networks face in providing new specialized services. Then, we studied how such challenges are being addressed by edge architectures and AI developments. Thus, our vision aims towards generalized dynamic EI solutions over the Internet and in our summary we also incorporate what is already implemented in the edge or AI fields (middle columns in Table 1), and highlight what is further needed to fully realise the potential and capabilities of EI (far right column in Table 1). Naturally, these challenges are not exhaustive and we have prioritized challenges that we have encountered in our development of practical EI solutions and applications.

A. SYSTEMS THAT NEVER SLEEP

Autonomous systems are one of the primary use cases for EI [48], [49]. In many domains, such as smart cities, medical monitoring, industrial control systems, or defense, these systems also cannot be shut down but must operate continuously. Ensuring these systems can operate consistently requires access to sufficient resources, and meeting highly dynamic resource demands. Edge solutions that can intelligently adapt functionality to meet these changing resource requirements, and software solutions that can scale up to ever-increasing amounts of devices are critical to achieve this. It is also important to keep in mind that as the solutions and underlying compute platforms become more sophisticated and widely available, the requirements for quality of service along with expectations for these applications also increase. Within the smart city context, the problem has evolved from a ''city that never sleeps'' paradigm to enabling ''connected megacities'' with an increasingly connected life, with more connected objects and increasingly autonomous systems [50]. The advantage of the new communication technologies, and the intelligent platforms they provide, is that we allow even small cities to access added value services and function exactly like those ''megacities'' with highly scalable infrastructure investments and limited additional costs.

From infrastructure and operator's point-of-view, systems that never sleep require the application of Quality-of-Service (QoS) and Quality-of-Experience (QoE) to become part of the operational and provisioning decisions [51]. A smart city, in particular, can be viewed as a ''multiagent cyber-physical system'' presenting a synergy between human agents and intelligent agents – encompassing infrastructure, transportation systems, waste collection, smart energy systems, surveillance, security, etc. Not only should the human agent seamlessly interact with other cyber-physical agents in the environments, but there is also a tight integration between the different intelligent agents to achieve a holistic operation. An example can be provided as the control function of traffic lights in such a city. For a truly smart operation, the traffic light not only must keep track of vehicle density on road but also presence of pedestrians, mobility patterns derived by external factors such as school or office hours, and weather changes. A smart city is a formidable example of a situation where the high level of interaction and ever-growing

requirements offer new opportunities for a high level of interaction and more challenging requirements. Moreover, a sophisticated operation of smart city applications envision some level of service sentience, i.e. the applications should operate as per predefined service level agreements (SLAs) regardless of the time of day. From a practical standpoint, ''never sleeping'' refers to ensuring high SLA requirements for different operations of the city, such as energy, surveillance, and security.

The digital transformation processes for worldwide cities are accelerating, with a focus on sustainability and improving citizen's quality of life [52]. Such processes are progressing together with a constantly growing introduction of sensors and connected objects, and together with an explosion of data production. D'Amico *et al.* [33] explore the main challenges related to sensors in cities, emphasizing the opportunities and critical issues of this growing digitalization of urban context. A city that never sleeps needs to communicate mores, at more levels, and more frequently. However, a city that communicates more, at more levels, and more frequently becomes progressively a smarter city that never sleeps [34]. With more strict requirements in terms of service levels of the basic connectivity functions and of the basic functions that are expected from the communication networks.

Serving such networks is not only a question of low latency and high bandwidth but also presents several other challenges. Just from software engineering perspective, such systems must endure and remain resilient towards disconnections and outages of individual connected devices. Therefore, edge intelligence must seamlessly support condition monitoring, fault detection, network reliability, and essential resilience functions within its control decisions to not just be reliable to state changes, but also be reactive.

B. LATENCY OF EXPERIENCES

The efficacy of Edge Intelligence is significantly driven by the context and requirement of the application that incorporates it. While edge computing, by nature, can help application developers leverage resources closer to the users, the needs and demands for optimal user experience can differ significantly for different applications. Mohan *et al.* [53] find that three strict human vestibular thresholds guide the latency requirements of edge-driven applications. Immersive applications, such as AR/VR, must abide by motion-to-photon (MTP) latency of ≈20 ms - which requires the sensory input and interactions to be completely synchronized [54], [55]. Interactive applications, such as gaming or video streaming must operate within perceivable latency (PL) of around 100 ms for optimal QoE [56]. Finally, applications that require active user inputs and engagements, e.g. teleoperated surgery, are highly dependent on the human reaction time (HRT) threshold of 250 ms. Vital applications within the smart healthcare and smart city domain, like remote surgery, also fall in this category [23].

The latencies for optimal quality of experience of edge applications described above do not just include network or processing latency but latency for the entire end-to-end process. For example, out of the 20 ms latency quota of AR/VR applications, ≈13 ms is reserved for display technology [54] due to refresh rate, pixel switching, and other functionality. Therefore, the application processing pipeline only has the remaining 7 ms to accommodate all communication, processing, modeling, and output formulation. Similarly, a typical perceived maximum communication and processing latency for autonomous vehicles is estimated to be below 10 ms and for remote surgery to be below 150 ms. This period does not include requirements to perform the data fusion, processing, and ML necessary for guiding the efficacy of the application.

The integration of AI with edge can introduce or exacerbate the existing latency challenges in many use-cases of EI. For example, industrial control systems harnessing EI have extremely strict latency requirements [35]. Arjevani and Shamir [36] conclude that many communication rounds will be required in AI processing in the edge and still provide the worst-case optimum in minimum assumption situations. The raw data acquisition, data analysis and training, and the continuous feedback loop in ML will introduce much higher delay simply considering the increased number of communication rounds.

The dynamic nature of many IoT environments and an increasing number of connected devices make flexibility and self-organization are among the most important capabilities that EI must offer. The main challenge is to design the optimized EI pipelines, faster and reliable on-the-air communications, and transparent symbiosis between the edge and end-user devices. The EI infrastructure should be scalable and capable of maintaining latency constraints, including the time required for data processing, model building, and AI/ML. Network virtualization likely becomes a key element where dedicated software can be distributed among nodes to make the best use of the available resources. Sharing fractions of available data or trained AI/ML models can significantly reduce the latency for highly dynamic situations and allow rapid reconfiguration without dropping users' QoE.

In smart cities, latency requirements are again strongly tied to application characteristics. While some applications, such as smart parking or air quality management, can endure an increased level of latency, there also exist safety-critical applications, such as smart traffic management, that cannot accommodate higher latency. Considering the brake reaction time of drivers, pedestrian monitoring and driver notifications systems have to complete their execution under 100 ms, which includes the acquisition of video streams from the pedestrian crossing, analyzing them to detect potential accidents, and transmitting a warning signal to the mobile devices of the affected drivers. A proof-of-concept for such a system has been successfully deployed on the smart traffic lights in Vienna urban area and shown to satisfy the latency quota mentioned above provided that pre-trained lightweight models are used, and 5G connectivity is available [57]. However, it is an open question whether more dynamic (and consequently computationally complex) AI models

(e.g., online, active, or transfer learning) can achieve similar results.

C. WHERE IS THE EDGE ACTUALLY

The edge's size and boundaries are essentially dependant on the used definitions and on the application domain. In a simplistic view, we could consider the edge to be exactly where the devices are connected. However, dealing with EI, this view might be different for the model training and inference phases and can vary in time and space in relation to what is connected to the network and the exact operational conditions. In practice, the network edge consists of a broad range of devices, including base stations, servers, IoT sensors and actuators and personal devices. Their computing capacity, memory, and storage are limited to some extent, in contrast to the cloud, where multiple services can operate in tandem. The connectivity and communication among edge devices are mainly enabled via the underlying wireless networks that are highly dynamic and diverse due to their inherited mobility and spatiotemporal characteristics. In addition, edge devices utilize several different software stacks ranging from almost bare metal to sophisticated container systems [15] with varying adaptation capacity.

In wireless networks, the geographical locations of the devices and their surroundings affect the communication reliability, latency, and capacity, and the resources available for a particular application may not be easily predictable. In this view, the immediate adoption of training and inference architectures from cloud to edge can be highly inefficient, neglecting on-device and communication constraints as well as the dynamics of the operating environment. Hence, optimizing distributed AI/ML algorithms and developing new mechanisms accounting for channel dynamics and communication overhead is of paramount importance. These challenges include the generalization to unmodeled phenomena under limited heterogeneous local data and straggling devices in the training process. The scale of the edge can vary from a few to thousands depending on time, users, and service providers. A small indoor environment can e.g. consist of a few cooperative devices with low-to-no dynamics within the duration of hours, in which enabling intelligence can be based on conventional architectures and algorithms. In contrast, automated vehicles in an intelligent transport system are highly susceptible to network dynamics, heterogeneity and spatiotemporal availability of resources, calling for novel AI/ML designs.

Thus, edge architectures and platforms must satisfy a challenging combination of scenarios and applications with potentially conflicting requirements. On the one hand, from an economic point of view, operators will be interested in minimizing the number of edge locations. On the other hand, edge and fog applications may require many edge instances or even to distribute the edge among a large set of devices (including embedded devices towards what is called mist computing). According to Lan *et al.* [58] applications can be classified according to their requirements:

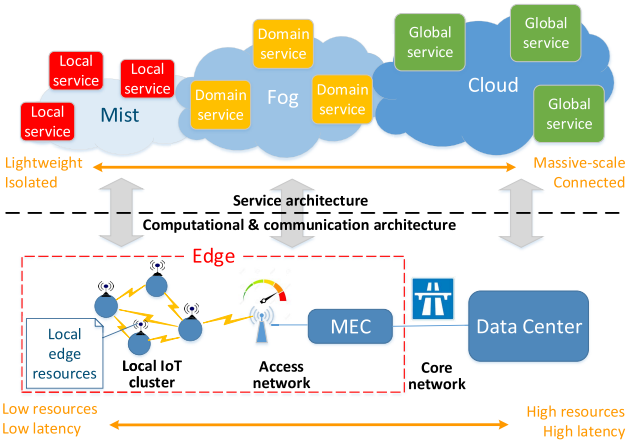


FIGURE 3. Where is the Edge.

- Latency-sensitive applications: Applications with strict latency requirements can only be achieved by executing the services at edge locations physically near the source of the data.
- Autonomous applications: Some applications are deployed in areas with poor connectivity, so they can not take advantage of the cloud paradigm.
- Privacy and security applications: Some applications have to address privacy concerns (e.g. if they manage health-related information), and they have to store and manage the data locally.
- Context-awareness applications: Distributed applications that have to use information such as the location or other local information related to each IoT device. Data processing and computation are conducted on small datasets that can be processed locally to avoid overloading the network.

In this scenario, the edge can not be static. Its functionality has to be distributed among instances in different locations and even taking advantage of the computing capabilities of all the devices participating in the network.

Thus, in edge-assisted cloud computing, applications can take advantage of all the available infrastructure. Cloud systems can better serve applications requiring low latency while saving computational and networking resources at core networks and data centers. The parts of services that require low latency or provide functions for reducing data, such as filtering, fusion or other processing, are beneficial to deploy at MEC hosts residing at access networks near base stations. The main benefits are the low end-to-end latency between the local node and MEC node and the reduced amount of data that needs to be delivered to data centers. As can be seen, IoT and smart environments can significantly benefit from MEC residing at the mobile access networks.

However, the current model where MEC hosts are deployed at servers located within or near the access network base stations also has its limitations [39]. In many smart space and IoT applications, to deal with possible connectivity

problems and limit the propagation of sensitive data outside the domain, at least some degree of processing of the sensor data and the decision-making/control logic is beneficial to be managed locally on-site [18], [21]. Therefore, in many scenarios it is beneficial to bring EC capacity within local IoT clusters, as illustrated in Figure 3. Since it cannot be expected that local IoT/IoE clusters include devices with sufficient stability and hardware capacity to accommodate full-functional MEC host, alternative decentralized solutions fitting better to the IoT/IoE environments need to be studied. A vital thrust towards utilizing the full potential of the cloud-edge continuum is the three-tier edge architecture as proposed in the present literature [39].

D. UBIQUITOUS USE OF EDGE RESOURCES

Ubiquitous computing and IoT introduce physical environments as opportunistic playgrounds for distributed applications. In such environments, EI has a crucial role in providing context-aware services and maintaining QoE for users. Key factors for orchestration of the service deployment and access include user location, computational and communication resources, and application data. In essence, edge resources must be placed [59], [60], [61] and their resources allocated [62] in a way that considers such factors and their trade-offs.

Edge intelligence can provide tools for such orchestration, considering and predicting user activities and the resulting fluctuating requirements in terms of multi-tenant resources: locations, migrating application contexts, providing connectivity, redirecting network traffic and maintenance. With intelligence, as self-capabilities, the applications become aware of the edge environment and can continuously negotiate their reliable and robust execution with the help of system services. Such developments lead to distributed EI where user, application, and system components intelligently adapt, offload, relocate, negotiate, and collaborate without a central authority to become the de-facto architectural model.

Nevertheless, orchestration is a system-wide collaborative effort [41], where resource management and control functionality is often separated from application functionality, e.g., data flows, at architectural level [42]. It is clear that the resource management functionality relies on AI solutions at large. The AI is to be distributed across the systems, where we believe EI, in particular, has a role in reducing the gap between separated functionalities. Therefore, an extensive set of edge services would be introduced [40], such as service discovery, on-demand logical topology and support for self-configuration, self-optimization, self-healing, and self-protection. Interoperability issues, such as shared functionality, interaction protocols, and portability should be supported on a technical level. The resulting distributed operation across the edge platform calls for distributed lightweight service provisioning and control mechanisms among the applications and systems components [39]. Taking a step further, virtual resource pools, including micro-operator resources, autonomous vehicles, network infrastructure components, mobile user devices, and everyday appliances, call for intelligent resource sharing solutions as exemplified by 3C-L and Tactile internet.

The discussed services require standardization, such as reference architectures and APIs, edge-specific software, and service modelling practises that support immersed intelligence. The starting point here would be the ETSI MEC reference architecture [63], currently under standardization. The ETSI standards provide the overall edge system architecture, required system components and their outlined functionality, set of APIs for system operation, information dissemination and third-party integration, guidelines and best practises, and set of Proof-of-concept (PoC) applications currently under consideration. Such efforts provide a solid base to realized edge systems, where some of the MEC system components (e.g., MEC orchestrator) and PoC applications largely are seen as relying on AI. However, the realization of AI functionalities is open, and EI has not yet been considered as a built-in capability of the edge system.

Moreover, Frameworks such as Fog05 [64], MobileFog [65], Distributed data flow (DDF) [66], or FogFlow [67] are being developed to manage the resources and to simplify the programming. Ubiquitous applications can take advantage of these frameworks to handle the different parts of their life-cycle (such as the development, deployment, execution, and management) transparently – to be deployed in a distributed edge architecture without requiring operators and third party developers to worry about managing the reservation and orchestration of computing, networking or storage resources, while satisfying the application requirements in terms of latency, mobility, heterogeneity, scalability or quality of service.

E. HIGHLY LOCALIZED INTELLIGENCE

The characteristics of EI solutions depend on a number of factors. First, the quality of data, in terms of volume, velocity, and variety; and the availability and location of processing, communication and storage resources restrict the potential functionality of the EI solution. Further, the functions, devices, and users to be served by the EI solution at a particular location set their requirements on, for example, the degree of autonomy needed. The key questions are: *where should intelligence be deployed*, *how does the deployed intelligence adapt to the local environment*, and *how do the localized intelligence interact*.

For example, a smart city needs to consider phenomena such as weather, air quality, and traffic. The corresponding data generating processes contain prominent spatio-temporal dependencies, which are reflected in the collected data. In some cases, the structure emerging from such spatial dependencies may be significant enough to affect the resulting model. On the positive side, such dependency structures can offer a way to distribute the model. For example, Lovén *et al.* [37] propose a distributed interpolation method that takes advantage of spatio-temporal dependencies and

partitions data for local model learning along boundaries projected on the spatial dimension.

Localization introduces several challenges. Massive-scale data analysis requires time for data delivery and processing, let alone building complex ML/AI models. Further, geographically distributed data, even if extensive, does not always improve the accuracy of the learned models, especially if local models learn only from local data. Such local models may be easy and lightweight to implement whenever they fit the application profile. Still, they can suffer in quality and generality due to a lack of variety in the local data sets. To overcome this problem, more delicate model communication standards and protocols need to be studied. Local models should exchange information and learn from each other to improve model quality and generality. It is imperative to determine what models can effectively be distributed and trained with highly localized data and how the life-cycle management of such distributed models can be arranged. Current promising approaches aim to identify the most significant updates and minimize communication while maximizing the knowledge and experience transfer [38]. EI solutions based on federated learning, such as In-Edge AI [1], solve the problem by periodically replacing local models with a global one, but consequently lose out any localized characteristics.

F. EDGE DEVELOPER EXPERIENCE

The ETSI MEC Application Development Community and PoCs already provide demonstrations in the realization of edge benefits and practical development aspects, such as feasibility, interoperability and testing, through a set of use cases. In addition, available tools for edge software development are well-established, including DevOps and MLOps practises with automatized continuous integration and delivery (CI/CD) on top of virtualization technologies and based on microservices and serverless computing paradigms. Platforms for managing and deploying ML/AI solutions on edge have been proposed (e.g. KubeFlow and MLFlow). Also, more fine-grained platforms (e.g. Function-as-a-Service, FaaS) and runtimes for elastic on-demand serverless computing have appeared, omitting the need to also focus on infrastructure/platform by application/service developers [43].

However, already IoT software engineering (SE) as such is complicated because tools, techniques, and skills in nearly all areas of modern software development are needed for developing end-to-end (E2E) systems [15]. A new layer of complexity appears with EI [21], which influences the design methodologies, architectures, tools, best practices, and the overall software life-cycle. The platforms supporting EI should provide elastic host service support and tools and means for straightforward deployment of on-demand software components that can exploit reliable, near real-time runtimes and execution environments with fast access to data and computing resources. EI can also mean another architectural layer to introduce complexity in system maintenance and resource sharing across the device-edge-cloud continuum.

Hence, the SE discipline must increasingly address EI as a building block towards autonomous, adaptive, and intelligent applications in an opportunistic and elastic online environment. Such SE discipline could be located in the intersection of APIs, distributed heterogeneous execution environments, distributed computing platforms, and AI. Here, the MLOps practises facilitating automatization of the management, operation and life-cycle of all types of ML/AI models for the edge applications atop the edge infrastructure.

As these somewhat different fields require different competencies, a risk is that the developer's role will become more complex, as has happened in the context of full-stack web development [44]. Still, systematic and consolidated methodologies for software development are needed with EI integrated from distinct perspectives. A novel EI software development process integrates applied ML/AI techniques and software modeling practises in the initial design phases [40]. ML/AI helps identify and assess the inherent opportunistic elements in parallel with domain expertise and provide feedback during the initial stages of the process. At the system development and deployment stages, EI is already aware of these aspects and could address them in operation.

G. INTEGRATION AND INTEROPERABILITY EFFORTS

In contrast to the centralized data center setting, EI needs to address challenges originating from hardware heterogeneity and resource management in a highly dynamic and decentralized environment. Therefore, interoperability becomes a key issue for portability, communication protocols, and data models. Portability enables EI applications and services to be deployed to different vendor solutions. Heterogeneous hardware and low-level communication should be beneath the provided standardized abstraction levels, e.g. with infrastructures and APIs suggested for the application and service developers. Requirements assessment, authentication, resource discovery, system configuration and deployment, and life-cycle management provided by the IE platform should have unified interfaces that enable more rapid adoption of IE technologies by vendors and industry. Here, the ETSI standardization provides well-defined architecture with a set of functionalities, APIs and development practices as the background to realize edge systems. Currently, it is unclear how the EI capabilities can be built into the edge systems, requiring specifications for further APIs, different software constructs and system services for integration, with support from the underlying edge infrastructure. Commonly accepted practices should be established, considering the existing standardization and frameworks.

At the extreme, this may lead us to isomorphic IoT architectures in which the devices, gateways, and cloud are able to run the same applications and services, allowing flexible migration of code between any element in the overall system [15]. Instead of learning different incompatible software development methods, one base technology will suffice and cover all aspects of E2E development. Although fully

isomorphic IoT systems are still years away, their arrival may ultimately dilute or even dissolve the boundaries between the cloud and edge. Isomorphic systems will allow computations to be transferred dynamically and performed on any level of the cloud-edge architecture that provides the optimal performance, storage, latency, and energy-efficiency characteristics.

In parallel with the edge system standardization, data specifications, models, interfaces, and representations should be agreed to ensure that concepts and their relationships are interpreted in the same way in the edge platforms, services, and applications. While this is impossible to achieve generally, the data integration and management tasks between the system, applications and application-specific services can be standardized, providing common mechanisms for data integration, discovery and management.

H. SECURE, PRIVATE, RELIABLE, AND RESILIENT EDGE

Edge intelligence inherits the existing security, reliability, and privacy challenges of edge computing. In addition, massively interconnected and high-speed communication networks introduce amplified security and privacy problems. Increasingly autonomous systems can attract large-scale attacks and introduce a wealth of vulnerabilities at different parts of the interconnected systems. A new category of risks emerges from possible malicious intelligence. Edge servers can be considered as aggregating points for all sensors in a local, possibly unprotected area, providing a single entry point for malicious entities that can access feeds from multiple sensors but target attention towards a single server responsible for handling the operation. For mature and trusted services, EI needs to be self-learning in all the layers of communication-related to end-to-end security [2], [45]. This calls for the careful design of centralization versus decentralized security protocols.

The use of AI also introduces security vulnerabilities [68] which, if exposed or exploited, can have severe consequences for EI and connected and derived functions. Computing and storage resources in the edge will be limited, and deploying complex AI procedures require higher resources, which can cause resource exhaustion attacks easier, as discussed in the case of IoT in [35]. Furthermore, mixing data from diverse sources can lead to unpredictable entanglements and hidden feedback loops [68]. Therefore, security validation of AI procedures and techniques and consequential analysis of the deployment of AI techniques in edge platforms must be carried out before enabling the automated EI infrastructure. Modular and hierarchical distribution of AI tasks can also minimize security risks.

Trust mechanisms are needed to guarantee the validity and trustworthiness of the EI devices and data providers. Distributed Ledger Technologies (DLT), such as Blockchain, has emerged as a potential solution to provide distributed and decentralized trust through mutual consensus mechanism among various actors [45]. Safeguards on user privacy are presently governed by the General Data Protection Regulation (GDPR) and by the Cybersecurity Regulation in the EU, and the restrictions of the data usage under the directives will affect the AI/ML paradigms in EI [2]. These represent fundamental legal milestones ensuring that privacy and security are reinforced.

Failure proneness of edge servers is another important issue that might endanger the overall reliability. Being deployed in exposed locations without data center-level advanced support systems increases the potential impact of hardware failure at the edge, which may risk the EI integrity. Existing reliability mechanisms such as re-execution or check-pointing might be infeasible, particularly for real-time EI [46]. While EI promises more resilience by edge flexibility within critical and transient failures due to network fluctuations, transparent control and reconfiguration mechanisms must be designed and implemented. Currently, there exist fault tolerance solutions for edge computing infrastructure [46], and neural network architecture [47]; however, joint consideration of the two aspects is missing.

The European Union Agency for Cybersecurity ENISA, has recently defined key research directions and innovation topics in cybersecurity. Indeed, it is clear that the more connectivity and the more intelligence is available at the edge, the more the balance between security and utility, privacy enhancement, and failure proneness need to be studied and adequately considered. Solutions will emerge from more intelligent security threat prevention, dynamically changeable privacy prevention acts, and locally adjusted trust management. The question about privacy that will remain open in the future, for instance in 6G, will be that: how personal can the information be in the time of shared storage, processing and data economy [69]?

V. EARLY CASE STUDIES

A fundamental enabler for EI is the underlying edge infrastructure based on medium- or small-scale edge servers. Due to the data and computation requirements for AI/ML and required system services in orchestration and sharing resources with real-time responsiveness, the placement of edge components is a crucial concern. However, the resulting architectures are typically fixed based on mobile networks and capabilities dictated by many factors, such as core network topology, capacity and traffic, operator policies, and user mobility. Therefore, intelligent approaches are needed for scalable edge infrastructure placement in different scenarios. To maintain QoE, extensive sets of real-world parameters need to be considered for online intelligence. Furthermore, architecture that dynamically supports such placement is an important requirement for the hosting software infrastructure. Below we discuss some application use cases for such deployments. As noted in Section II, EI is relevant for a wide range of domains and the application scenarios are not meant as exhaustive list of all possible solutions, rather as examples that demonstrate the practical benefits and have potential for uptake. Specifically, we require the application scenarios to meet two criteria. First, we require the application use cases

(a) Edge node deployment. (b) Smart traffic light integration.

FIGURE 4. Edge nodes setup on Vienna's chosen intersection and the integration into the traffic-signal chambers [57].

to be novel representative and concrete examples of already implemented use cases rather than application domains that have thus far only been envisioned. Second, the case studies were chosen to cover three promising application areas for EI as has been identified in prior surveys [2], [16]: Smart Cities, Industrial IoT, and Environmental Monitoring.

Intelligent traffic light solution utilizing EI is implemented in Vienna, Austria, in order to improve traffic safety [57]. The road sections around dangerous intersections with low visibility and their surroundings are continuously monitored with video cameras deployed at traffic lights, where the streaming video is processed locally. Relevant events such as pedestrians or cyclists entering the road segments are detected in real-time and nearby drivers are alerted via a mobile application. The local processing of data not only reduces response time by avoiding long-distance transfer of big streaming data but also enables continuous delivery of the service even if the remote infrastructure is not accessible. Moreover, the privacy of pedestrians is preserved since the video recordings never leave the traffic light.

In this case study, single-board Raspberry Pi edge devices extended with Google's Coral Edge TPU accelerators have been integrated into traffic signal chambers (Figure 4). These nodes run pre-trained TensorFlow Lite models to detect pedestrians or cyclists and send alerts to nearby drivers' mobile devices via MQTT protocol over 5G. Real-world evaluation shows that affected drivers can be notified in around 100 ms, 18 ms of which is the processing time of a frame and the rest is for communication with the guaranteed delivery of the alert [57].

Automotive EI application for the automotive scenario is studied at Poznan, Poland, where dynamic management of autonomous car platooning is supported using rich context information stored in databases. The study focuses on improving the reliability of intra-platoon wireless communications that suffer from channel congestion in the 5.9 GHz frequency band. The utilization of alternative frequency bands is proposed, such as the TV white spaces or mmWave, that are dynamically selected based on the additional information from databases (e.g. the observed TV signal power at a specific location). A hierarchical structure of edge intelligence support is considered, where, depending on the origin of information and its scope, it can be stored in regional or local databases or even in distributed form. The initial findings indicate that it is possible to improve communication reliability for platooning with EI significantly [70], [71].

Environmental sensing The EDISON project, studied in Oulu, Finland, proposes an edge-native method and architecture for distributed interpolation [37]. EDISON assumes a large fleet of mobile sensors, collecting environmental data. The mobile nodes are calibrated upon rendezvous with sparse high-quality fixed sensors, transmitting their data for learning and inferring with a distributed interpolation model running at edge nodes. Early simulation studies promise an improvement over baseline distributed methods as well as a global interpolation model, assuming data is generated by relatively independent, spatially distributed processes.

Urban-scale air quality sensing is an example of city-scale application domains that can benefit from edge intelligence. The MegaSense programme at the University of Helsinki explores how to extend the scale of air quality monitoring to support dense and high-resolution information [72]. Air quality is traditionally monitored using professional-grade measurement stations that are highly expensive to both deploy and operate. Increasing the monitoring scale requires integrating sensors of different types, such as low-cost sensors carried by citizens to industrial-grade sensors located at industrial sites and in the urban infrastructure with the professional-grade monitoring stations. Low-cost sensors tend to suffer from lower accuracy, which can be mitigated using machine learning-based calibration [73]. The idea in calibration is to learn a model that can compensate for the errors in the low-cost sensors. Air quality information tends to have strong spatial correlations, and thus the calibration models of sensors in the same spatial area can share information instead of learning a separate model for each sensor. Edge deployments are essential for ensuring the calibration can operate efficiently, e.g., recent work at the University of Helsinki has demonstrated how deploying the calibration on edge can reduce latency and minimize overall communication bandwidth in city-scale deployments [74].

VI. CONCLUDING REMARKS

Cloudification has so far helped to promote the adoption of AI/ML methods and develop intelligent applications and services. Local use of these techniques on edge is now progressively growing. In the near future, we envisage AI being pervasive and supporting many aspects of the operations of future communication networks and their edge. We took a broad view on emerging EI solutions, discussing the motivations and applications that will not simply benefit from EI but will be enabled by EI, and presented some early case studies in emerging application areas. We also identified aspects that we consider priorities for the likely R&D and innovation activities.

Recent large-scale cyber-security incidents and attacks rely on a traditional communication network and relatively

non-autonomous interconnected systems and devices. What could happen with an increasingly intelligent and autonomous network if not properly instructed to support privacy, security, reliability, and resilience by design. Considering the experience of 5G and 5G acceptance, a proper consideration of these aspects since the early phases of any beyond 5G development is an absolute necessity. As a regulatory example, under the EU Commission's new digital strategy, additional regulatory actions have been planned, including creating a specific AI framework addressing safety and ethical challenges, and the adaptation of existing safety and liability frameworks to possible new technologies. A dedicated extension to future intelligent communication networks is essential, with certification schemes for privacy, security, reliability, and resilience to boost the development of secure and robust networking environments, while ensuring that the relevant legislation, initiatives, and policies are fully respected.

ACKNOWLEDGMENT

This paper has been written by an international expert group, led by the 6G Flagship at the University of Oulu, Finland (AoF grants 318927, 326291, 323630; Infotech Oulu grants B-TEA, TrustedMaaS).

REFERENCES

[1] X. Wang, Y. Han, C. Wang, Q. Zhao, X. Chen, and M. Chen, "In-edge AI: Intelligentizing mobile edge computing, caching and communication by federated learning," *IEEE Netw.*, vol. 33, no. 5, pp. 156–165, Sep. 2019.

[2] 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," *Proc. IEEE*, vol. 107, no. 8, pp. 1738–1762, Aug. 2019.

[3] A. H. Sodhro, S. Pirbhulal, and V. H. C. de Albuquerque, "Artificial intelligence-driven mechanism for edge computing-based industrial applications," *IEEE Trans. Ind. Informat.*, vol. 15, no. 7, pp. 4235–4243, Jul. 2019, doi: 10.1109/TII.2019.2902878.

[4] Y. Liu, C. Yang, L. Jiang, S. Xie, and Y. Zhang, "Intelligent edge computing for IoT-based energy management in smart cities," *IEEE Netw.*, vol. 33, no. 2, pp. 111–117, Mar./Apr. 2019, doi: 10.1109/MNET.2019.1800254.

[5] J. Mendez, K. Bierzynski, M. P. Cuéllar, and D. P. Morales, "Edge intelligence: Concepts, architectures, applications and future directions," *ACM Trans. Embedded Comput. Syst.*, to be published, doi: 10.1145/3486674.

[6] E. Cavalieri d'Oro, S. Colombo, M. Gribaudo, M. Iacono, D. Manca, and P. Piazzolla, "Modeling and evaluating a complex edge computing based systems: An emergency management support system case study," *Internet Things*, vol. 6, Jun. 2019, Art. no. 100054, doi: 10.1016/j.iot.2019.100054.

[7] G. Plastiras, M. Terzi, C. Kyrkou, and T. Theocharidcs, "Edge intelligence: Challenges and opportunities of near-sensor machine learning applications," in *Proc. IEEE 29th Int. Conf. Appl.-Specific Syst., Archit. Processors (ASAP)*, Jul. 2018, pp. 1–7, doi: 10.1109/ASAP.2018.8445118.

[8] S. Deng, H. Zhao, W. Fang, J. Yin, S. Dustdar, and A. Y. Zomaya, "Edge intelligence: The confluence of edge computing and artificial intelligence," *IEEE Internet Things J.*, vol. 7, no. 8, pp. 7457–7469, Aug. 2020, doi: 10.1109/JIOT.2020.2984887.

[9] A. Al-Shuwaili and O. Simeone, "Energy-efficient resource allocation for mobile edge computing-based augmented reality applications," *IEEE Wireless Commun. Lett.*, vol. 6, no. 3, pp. 398–401, Jun. 2017.

[10] F. Dressler, H. Hartenstein, O. Altintas, and O. Tonguz, "Inter-vehicle communication: Quo vadis," *IEEE Commun. Mag.*, vol. 52, no. 6, pp. 170–177, Jun. 2014.

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

[12] E. Lagerspetz, J. Hamberg, X. Li, H. Flores, P. Nurmi, N. Davies, and S. Helal, "Pervasive data science on the edge," *IEEE Pervasive Comput.*, vol. 18, no. 3, pp. 40–49, Jul. 2019.

[13] *Medical Imaging's Next Frontier: AI and the Edge*. Accessed: Jun. 11, 2021. [Online]. Available: https://www.forbes.com/sites/insights-inteliot/2020/12/09/medical-imagings-next-frontier-ai-and-the-edge

[14] S. Wang, T. Tuor, T. Salonidis, K. K. Leung, C. Makaya, T. He, and K. Chan, "Adaptive federated learning in resource constrained edge computing systems," *IEEE J. Sel. Areas Commun.*, vol. 37, no. 3, pp. 1205–1221, Jun. 2019.

[15] A. Taivalsaari and T. Mikkonen, "A taxonomy of IoT client architectures," *IEEE Softw.*, vol. 35, no. 3, pp. 83–88, May 2018.

[16] J. Chen and X. Ran, "Deep learning with edge computing: A review," *Proc. IEEE*, vol. 107, no. 8, pp. 1655–1674, Aug. 2019.

[17] B. Cao, L. Zhang, Y. Li, D. Feng, and W. Cao, "Intelligent offloading in multi-access edge computing: A state-of-the-art review and framework," *IEEE Commun. Mag.*, vol. 57, no. 3, pp. 56–62, Mar. 2019, doi: 10.1109/MCOM.2019.1800608.

[18] M. Merenda, C. Porcaro, and D. Iero, "Edge machine learning for AI-enabled IoT devices: A review," *Sensors*, vol. 20, no. 9, p. 2533, Apr. 2020, doi: 10.3390/s20092533. [Online]. Available: https://www.mdpi.com/1424-8220/20/9/2533

[19] T. Sipola, J. Alatalo, T. Kokkonen, and M. Rantonen, "Artificial intelligence in the IoT era: A review of edge AI hardware and software," in *Proc. 31st Conf. Open Innov. Assoc. (FRUCT)*, Apr. 2022, pp. 320–331, doi: 10.23919/FRUCT54823.2022.9770931.

[20] W. Lin, A. Adetomi, and T. Arslan, "Low-power ultra-small edge AI accelerators for image recognition with convolution neural networks: Analysis and future directions," *Electronics*, vol. 10, no. 17, p. 2048, Aug. 2021, doi: 10.3390/electronics10172048.

[21] J. Yao, S. Zhang, Y. Yao, F. Wang, J. Ma, J. Zhang, Y. Chu, L. Ji, K. Jia, T. Shen, A. Wu, F. Zhang, Z. Tan, K. Kuang, C. Wu, F. Wu, J. Zhou, and H. Yang, "Edge-cloud polarization and collaboration: A comprehensive survey for AI," *IEEE Trans. Knowl. Data Eng.*, early access, May 27, 2022, doi: 10.1109/TKDE.2022.3178211.

[22] L. Sun, X. Jiang, H. Ren, and Y. Guo, "Edge-cloud computing and artificial intelligence in Internet of Medical Things: Architecture, technology and application," *IEEE Access*, vol. 8, pp. 101079–101092, 2020, doi: 10.1109/ACCESS.2020.2997831.

[23] G. Ananthanarayanan, P. Bahl, P. Bodík, K. Chintalapudi, M. Philipose, L. Ravindranath, and S. Sinha, "Real-time video analytics: The killer app for Edge computing," *Computer*, vol. 50, no. 10, pp. 58–67, 2017.

[24] K. Ha, Z. Chen, W. Hu, W. Richter, P. Pillai, and M. Satyanarayanan, "Towards wearable cognitive assistance," in *Proc. 12th Annu. Int. Conf. Mobile Syst., Appl., Services*, Jun. 2014, pp. 68–81.

[25] S. Xu, Y. Qian, and R. Q. Hu, "Edge intelligence assisted gateway defense in cyber security," *IEEE Netw.*, vol. 34, no. 4, pp. 14–19, Jul. 2020.

[26] A. Holzinger, C. Röcker, and M. Ziefle, *From Smart Health to Smart Hospitals*. Cham: Springer International Publishing, 2015, doi: 10.1007/978-3-319-16226-3_1.

[27] A. Kusiak, "Smart manufacturing," *Int. J. Prod. Res.*, vol. 56, nos. 1–2, pp. 508–517, 2017.

[28] Z. Chen, W. Hu, J. Wang, S. Zhao, B. Amos, G. Wu, K. Ha, K. Elgazzar, P. Pillai, R. Klatzky, D. Siewiorek, and M. Satyanarayanan, "An empirical study of latency in an emerging class of edge computing applications for wearable cognitive assistance," in *Proc. 2nd ACM/IEEE Symp. Edge Comput.*, Oct. 2017, pp. 1–14, doi: 10.1145/3132211.3134458.

[29] S. Pan, P. Li, C. Yi, D. Zeng, Y.-C. Liang, and G. Hu, "Edge intelligence empowered urban traffic monitoring: A network tomography perspective," *IEEE Trans. Intell. Transp. Syst.*, vol. 22, no. 4, pp. 2198–2211, Apr. 2021.

[30] F. Giust, G. Verin, K. Antevski, J. Chou, Y. Fang, W. Featherstone, F. Fontes, D. Frydman, A. Li, and A. Manzalini, "MEC deployments in 4G and evolution towards 5G," *ETSI White Paper*, vol. 24, p. 24, Feb. 2018.

[31] S. Legg and M. Hutter, "A collection of definitions of intelligence," *Frontiers Artif. Intell. Appl.*, vol. 157, p. 17, Jun. 2007.

[32] Y. Zhao, W. Wang, Y. Li, C. C. Meixner, M. Tornatore, and J. Zhang, "Edge computing and networking: A survey on infrastructures and applications," *IEEE Access*, vol. 7, pp. 101213–101230, 2019, doi: 10.1109/ACCESS.2019.2927538.

[33] G. D'Amico, P. L'Abbate, W. Liao, T. Yigitcanlar, and G. Ioppolo, "Understanding sensor cities: Insights from technology giant company driven smart urbanism practices," *Sensors*, vol. 20, no. 16, p. 4391, Aug. 2020.

[34] T. Yigitcanlar, K. C. Desouza, L. Butler, and F. Roozkhosh, "Contributions and risks of artificial intelligence (AI) in building smarter cities: Insights from a systematic review of the literature," *Energies*, vol. 13, no. 6, p. 1473, Mar. 2020.

[35] I. Ahmad, S. Shahabuddin, T. Sauter, E. Harjula, T. Kumar, M. Meisel, M. Juntti, and M. Ylianttila, "The challenges of artificial intelligence in wireless networks for the Internet of Things: Exploring opportunities for growth," *IEEE Ind. Electron. Mag.*, vol. 15, no. 1, pp. 16–29, Mar. 2021.

[36] Y. Arjevani and O. Shamir, "Communication complexity of distributed convex learning and optimization," in *Proc. 28th Int. Conf. Neural Inf. Process. Syst. (NIPS)*, vol. 1. Cambridge, MA, USA: MIT Press, 2015, pp. 1756–1764, doi: 10.5555/2969239.2969435.

[37] L. Lovén, T. Lähderanta, L. Ruha, E. Peltonen, I. Launonen, M. J. Sillanpää, J. Riekki, and S. Pirttikangas, "EDISON: An edge-native method and architecture for distributed interpolation," *Sensors*, vol. 21, no. 7, p. 2279, Mar. 2021.

[38] A. Aral, M. Erol-Kantarci, and I. Brandić, "Staleness control for edge data analytics," *ACM SIGMETRICS Perform. Eval. Rev.*, vol. 48, no. 1, pp. 19–20, Jul. 2020.

[39] E. Harjula, P. Karhula, J. Islam, T. Leppänen, A. Manzoor, M. Liyanage, J. Chauhan, T. Kumar, I. Ahmad, and M. Ylianttila, "Decentralized IoT edge nanoservice architecture for future gadget-free computing," *IEEE Access*, vol. 7, pp. 119856–119872, 2019.

[40] T. Leppänen, C. Savaglio, and G. Fortino, "Service modeling for opportunistic edge computing systems with feature engineering," *Comput. Commun.*, vol. 157, pp. 308–319, May 2020.

[41] Z. Wen, R. Yang, P. Garraghan, T. Lin, J. Xu, and M. Rovatsos, "Fog orchestration for Internet of Things services," *IEEE Internet Comput.*, vol. 21, no. 2, pp. 16–24, Feb. 2017.

[42] C.-H. Hong and B. Varghese, "Resource management in fog/edge computing: A survey on architectures, infrastructure, and algorithms," *ACM Comput. Surv.*, vol. 52, no. 5, pp. 1–37, Sep. 2020.

[43] I. Baldini, P. Castro, K. Chang, P. Cheng, S. Fink, V. Ishakian, N. Mitchell, V. Muthusamy, R. Rabbah, and A. Slominski, "Serverless computing: Current trends and open problems," in *Research Advances in Cloud Computing*. Cham, Switzerland: Springer, 2017, doi: 10.1007/978-981-10-5026-8_1.

[44] A. Taivalsaari, T. Mikkonen, C. Pautasso, and K. Systä, "Full stack is not what it used to be," in *Proc. Int. Conf. Web Eng.* Cham, Switzerland: Springer, 2021, pp. 363–371.

[45] L. Liu, F. R. Yu, X. Li, H. Ji, and V. C. Leung, "Blockchain and machine learning for communications and networking systems," *IEEE Commun. Surveys Tuts.*, vol. 22, no. 2, pp. 1392–1431, 2nd Quart., 2020.

[46] A. Aral and I. Brandic, "Learning spatiotemporal failure dependencies for resilient edge computing services," *IEEE Trans. Parallel Distrib. Syst.*, vol. 32, no. 7, pp. 1578–1590, Jul. 2021.

[47] T. Liu, W. Wen, L. Jiang, Y. Wang, C. Yang, and G. Quan, "A fault-tolerant neural network architecture," in *Proc. 56th ACM/IEEE Design Autom. Conf. (DAC)*, Jun. 2019, pp. 1–6, doi: 0.1145/3316781.3316819.

[48] P. Sroka and A. Kliks, "Towards edge intelligence in the automotive scenario: A discourse on architecture for database-supported autonomous platooning," *J. Commun. Netw.*, vol. 24, no. 2, pp. 192–208, Apr. 2022, doi: 10.23919/JCN.2022.000005.

[49] B. Yang, X. Cao, K. Xiong, C. Yuen, Y. L. Guan, S. Leng, L. Qian, and Z. Han, "Edge intelligence for autonomous driving in 6G wireless system: Design challenges and solutions," *IEEE Wireless Commun.*, vol. 28, no. 2, pp. 40–47, Apr. 2021.

[50] Y. Mehmood, F. Ahmad, I. Yaqoob, A. Adnane, M. Imran, and S. Guizani, "Internet-of-Things-based smart cities: Recent advances and challenges," *IEEE Commun. Mag.*, vol. 55, no. 9, pp. 16–24, Sep. 2017.

[51] J. Jin, J. Gubbi, T. Luo, and M. Palaniswami, "Network architecture and QoS issues in the Internet of Things for a smart city," in *Proc. Int. Symp. Commun. Inf. Technol. (ISCIT)*, Oct. 2012, pp. 956–961, doi: 10.1109/ISCIT.2012.6381043.

[52] L. Diez, J. Choque, L. Sanchez, and L. Munoz, "Fostering IoT service replicability in interoperable urban ecosystems," *IEEE Access*, vol. 8, pp. 228480–228495, 2020.

[53] N. Mohan, L. Corneo, A. Zavodovski, S. Bayhan, W. Wong, and J. Kangasharju, "Pruning edge research with latency shears," in *Proc. 19th ACM Workshop Hot Topics Netw.*, Nov. 2020, pp. 182–189.

[54] R. E. Bailey, J. J. Arthur III, and S. P. Williams, "Latency requirements for head-worn display S/EVS applications," *Proc. SPIE*, vol. 5424, pp. 98–109, Aug. 2004, doi: 10.1117/12.554462.

[55] K. Mania, B. D. Adelstein, S. R. Ellis, and M. I. Hill, "Perceptual sensitivity to head tracking latency in virtual environments with varying degrees of scene complexity," in *Proc. 1st Symp. Appl. Perception Graph. Visualizat. (APGV)*, 2004, pp. 39–47, doi: 10.1145/1012551.1012559.

[56] T. Kämäräinen, M. Siekkinen, A. Ylä-Jääski, W. Zhang, and P. Hui, "A measurement study on achieving imperceptible latency in mobile cloud gaming," in *Proc. 8th ACM Multimedia Syst. Conf.*, Jun. 2017, pp. 88–99, doi: 10.1145/3083187.3083191.

[57] I. Lujic, V. D. Maio, K. Pollhammer, I. Bodrozic, J. Lasic, and I. Brandic, "Increasing traffic safety with real-time edge analytics and 5G," in *Proc. 4th Int. Workshop Edge Syst., Anal. Netw.*, 2021, pp. 19–24, doi: 10.1145/3434770.3459732.

[58] D. Lan, A. Taherkordi, F. Eliassen, and G. Horn, "A survey on fog programming: Concepts, state-of-the-art, and research challenges," in *Proc. 2nd Int. Workshop Distrib. Fog Services Design (DFSD)*, 2019, pp. 1–6, doi: 10.1145/3366613.3368120.

[59] L. Loven, T. Lahderanta, L. Ruha, T. Leppanen, E. Peltonen, J. Riekki, and M. J. Sillanpaa, "Scaling up an edge server deployment," in *Proc. IEEE Int. Conf. Pervasive Comput. Commun. Workshops (PerCom Workshops)*, Mar. 2020, pp. 1–7, doi: 10.1109/PerComWorkshops48775.2020.9156204.

[60] T. Lähderanta, T. Leppänen, L. Ruha, L. Lovén, E. Harjula, M. Ylianttila, J. Riekki, and M. J. Sillanpää, "Edge computing server placement with capacitated location allocation," *J. Parallel Distrib. Comput.*, vol. 153, pp. 130–149, Jul. 2021, doi: 10.1016/j.jpdc.2021.03.007.

[61] L. Ruha, T. Lähderanta, L. Lovén, T. Leppänen, J. Riekki, and M. J. Sillanpää, "Capacitated spatial clustering with multiple constraints and attributes," 2021, *arXiv:2010.06333*.

[62] L. Loven, E. Peltonen, E. Harjula, and S. Pirttikangas, "Weathering the reallocation storm: Large-scale analysis of edge server workload," in *Proc. Joint Eur. Conf. Netw. Commun. 6G Summit (EuCNC/6G Summit)*, Jun. 2021, pp. 317–322, doi: 10.1109/EUCNC/6GSUMMIT51104.2021.9482593.

[63] *Multi-Access Edge Computing (MEC): Framework and Reference Architecture*, document ETSI GS MEC 003, ETSI, 2019.

[64] A. Corsaro and G. Baldoni, "FogØ5: Unifying the computing, networking and storage fabrics end-to-end," in *Proc. 3rd Cloudification Internet Things (CIoT)*, Jul. 2018, pp. 1–8, doi: 10.1109/CIOT.2018.8627124.

[65] K. Hong, D. Lillethun, U. Ramachandran, B. Ottenwälder, and B. Koldehofe, "Mobile fog: A programming model for large-scale applications on the Internet of Things," in *Proc. 2nd ACM SIGCOMM Workshop Mobile Cloud Comput. (MCC)*, 2013, pp. 15–20, doi: 10.1145/2491266.2491270.

[66] N. K. Giang, M. Blackstock, R. Lea, and V. C. M. Leung, "Distributed data flow: A programming model for the crowdsourced Internet of Things," in *Proc. Doctoral Symp. 16th Int. Middleware Conf.*, Dec. 2015, pp. 1–4, doi: 10.1145/2843966.2843970.

[67] B. Cheng, G. Solmaz, F. Cirillo, E. Kovacs, K. Terasawa, and A. Kitazawa, "FogFlow: Easy programming of IoT services over cloud and edges for smart cities," *IEEE Internet Things J.*, vol. 5, no. 2, pp. 696–707, Apr. 2018.

[68] J. Suomalainen, A. Juhola, S. Shahabuddin, A. Mammela, and I. Ahmad, "Machine learning threatens 5G security," *IEEE Access*, vol. 8, pp. 190822–190842, 2020.

[69] M. Ylianttila, R. Kantola, A. Gurtov, L. Mucchi, I. Oppermann, and I. Ahmad, "6G white paper: Research challenges for trust, security and privacy," 2020, *arXiv:2004.11665*.

[70] M. Sybis, P. Sroka, A. Kliks, and P. Kryszkiewicz, "V2X communications for platooning: Impact of sensor inaccuracy," in *Proc. Int. Conf. Image Process. Commun.* Cham, Switzerland: Springer, 2019, pp. 318–325, doi: 10.1007/978-3-030-31254-1_38.

[71] P. Kryszkiewicz, A. Kliks, P. Sroka, and M. Sybis, "The impact of blocking cars on pathloss within a platoon: Measurements for 26 GHz band," in *Proc. Int. Conf. Softw., Telecommun. Comput. Netw. (SoftCOM)*, Sep. 2021, pp. 1–6, doi: 10.23919/SoftCOM52868.2021.9559096.

[72] N. H. Motlagh, E. Lagerspetz, P. Nurmi, X. Li, S. Varjonen, J. Mineraud, M. Siekkinen, A. Rebeiro-Hargrave, T. Hussein, T. Petaja, M. Kulmala, and S. Tarkoma, "Toward massive scale air quality monitoring," *IEEE Commun. Mag.*, vol. 58, no. 2, pp. 54–59, Feb. 2020.

[73] F. Concas, J. Mineraud, E. Lagerspetz, S. Varjonen, X. Liu, K. Puolamäki, P. Nurmi, and S. Tarkoma, "Low-cost outdoor air quality monitoring and sensor calibration: A survey and critical analysis," *ACM Trans. Sensor Netw.*, vol. 17, no. 2, pp. 1–44, May 2021.

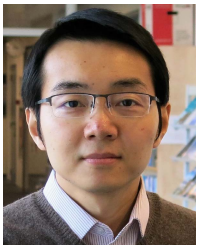
[74] X. Su, X. Liu, N. H. Motlagh, J. Cao, P. Su, P. Pellikka, Y. Liu, T. Petaja, M. Kulmala, P. Hui, and S. Tarkoma, "Intelligent and scalable air quality monitoring with 5G edge," *IEEE Internet Comput.*, vol. 25, no. 2, pp. 35–44, Mar. 2021.

ELLA PELTONEN (Member, IEEE) received the Ph.D. degree from the University of Helsinki, Finland, and the Postdoctoral degree from University College Cork, Ireland. She is currently working as an Assistant Professor (Tenure Track) at the University of Oulu, Finland. Her research interests include pervasive intelligent systems and services in the 5G/6G era.

IJAZ AHMAD (Member, IEEE) received the M.Sc. and Ph.D. degrees in wireless communications from the University of Oulu, Finland, in 2012 and 2018, respectively. He is currently working with the VTT Technical Research Centre of Finland, and an Adjunct Professor at the University of Oulu, Finland. His research interests include cybersecurity, security of 5G/6G, healthcare technologies, and the applications of machine learning in wireless networks.

ATAKAN ARAL (Member, IEEE) received the dual M.Sc. degrees in computer science and engineering from Politecnico di Milano and Istanbul Technical University (ITU), in 2011 and 2012, respectively, and the Ph.D. degree in computer engineering from ITU, in 2016. He is a Project Leader and a Research Fellow at the University of Vienna. His research interests include center around resource management for geo-distributed and virtualized computing systems, such as inter-cloud and edge computing, and edge intelligence.

MICHELE CAPOBIANCO received the M.Sc. degree in electronics engineering from the University of Padova, Italy. He has over 30 years of experience in research and development, product development, and project management for a variety of technology sectors. He is currently working as a consultant in innovation management and technology transfer projects. He is also a Business Innovation Manager at Capobianco, Italy.



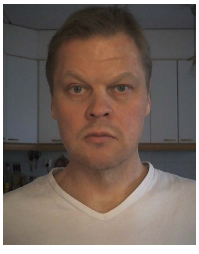
AARON YI DING (Member, IEEE) is a Tenured Associate Professor and leading the Cyber-Physical Intelligence (CPI) Laboratory at TU Delft. He has over 15 years of research and development experience across EU, U.K., and USA. His research interests include edge AI solutions for cyber-physical systems in smart health, mobility, and energy domains. He is an Associate Editor of *ACM Transactions on Internet of Things* (TIOT) and IEEE Open Journal of the Intelligent Transportation Systems.

FELIPE GIL-CASTIÑEIRA received the M.Sc. and Ph.D. degrees in telecommunication engineering from the University of Vigo, in 2002 and 2007, respectively. From 2014 to September 2016, he was the Head of the iNetS area at the Galician Research and Development Center in Advanced Telecommunications. He is currently an Associate Professor with the Department of Telematics Engineering, University of Vigo. His research interests include wireless communication and core network technologies, multimedia communications, embedded systems, ubiquitous computing, and the Internet of Things. He is the Co-Founder of University Spin-Off, Ancora.

EKATERINA GILMAN (Member, IEEE) received the D.Sc. (Tech.) degree. She has worked as a Visiting Researcher at the Northern Finland Biobank Borealis and the Centre for Health and Technology, Oulu, Finland. She is currently a Postdoctoral Researcher supported by the Academy of Finland, Center for Ubiquitous Computing, University of Oulu, Finland. Her current research interests include data analytics, context modeling and reasoning, machine learning in ubiquitous computing, the IoT, and data-intensive systems.

ERKKI HARJULA (Member, IEEE) received the M.Sc. and D.Sc. degrees from the University of Oulu, Finland. He works as an Assistant Professor (tenure track) at CWC-NS Research Group, University of Oulu. He focuses on wireless systems level architectures for future digital healthcare, where his key research topics are wrapped around intelligent trustworthy distributed IoT and edge computing. He has coauthored more than 70 international peer-reviewed articles. He is an Associate Editor of *Wireless Networks* journal (Springer).

MARKO JURMU received the M.Sc. (Hons.) and Dr.Tech. degrees in computer science from the University of Oulu, in 2007 and 2014, respectively. He is a Senior Scientist at Cognitive Production Industry research area, VTT Technical Research Centre Finland Ltd. He has been a Visiting Research Scientist at the University of Maryland, USA; Ludwig-Maximilians University, Germany; and Keio University, Japan. His research interests include data analytics, machine learning, distributed systems, and human–computer interaction.



TEEMU KARVONEN received the Ph.D. degree. He is currently working as a Postdoctoral Researcher at the University of Oulu, in M3S research unit (Information Technology and Electrical Engineering). Before his current position, he worked as a Postdoctoral Research Associate in agile research network at The Open University, Faculty of Science Technology Engineering and Mathematics (U.K., Milton Keynes). His current research interests include software-defined vehicles and vehicular connectivity in 6G. Before his current academic researcher career, he has worked in various industry projects on IP-networks and mobile telecommunications systems and services.

MARKUS KELANTI received the Ph.D. degree from the University of Oulu, Finland. He is currently working as a Postdoctoral Researcher at the University of Oulu. His research interests include software engineering, digital twins, and intelligent systems and services.

TEEMU LEPPÄNEN (Senior Member, IEEE) received the Doctoral degree in computer science and engineering from the University of Oulu, Finland. He is currently serves as a Principal Lecturer at the Oulu University of Applied Sciences. His current research interest includes edge computing for the Internet of Things systems.

LAURI LOVÉN (Senior Member, IEEE) received the M.Sc. degree in statistics and the D.Sc. (Tech.) degree from the University of Oulu, Finland, in 2016 and 2021, respectively. He is currently affiliated as a Faculty Universitätsassistent at the Distributed Systems Group, TU Wien, and a Postdoctoral Researcher at the Center for Ubiquitous Computing, University of Oulu. His research interest includes edge intelligence, in particular, learning and decision making in relation to orchestration in the computing continuum.

TOMMI MIKKONEN received the Doctorate degree from the Tampere University of Technology, Finland. He is a Full Professor of software engineering at the University of Jyväskylä, Finland. His current research interests include the IoT, software engineering, and multi-device programming.

NITINDER MOHAN received the M.Tech. degree (Hons.) from the Indraprastha Institute of Information Technology–Delhi (IIIT-D), India, in 2015, and the Ph.D. (as Marie Curie ITN fellow) degree from the Department of Computer Science, University of Helsinki, Finland, in 2019. He is currently a Senior Researcher at the Chair of Connected Mobility, Technical University of Munich, Germany. His research interests include edge computing, next-generation networked applications, and large-scale internet measurements. He was awarded the "Outstanding Ph.D. Dissertation Award" by IEEE Technical Committee on Scalable Computing (TCSC).

PETTERI NURMI received the Ph.D. degree in computer science from the Department of Computer Science, University of Helsinki, Finland, in 2009. He is an Associate Professor of distributed systems and the Internet of Things, Department of Computer Science, University of Helsinki. His research interests include distributed systems, pervasive data science, and sensing systems.

SUSANNA PIRTTIKANGAS (Member, IEEE) received the M.Sc. degree in mathematics and the D.Sc. (tech.) degree from the University of Oulu, Finland, in 1998 and 2004, respectively. She works as the Deputy Director at the Center for Ubiquitous Computing, University of Oulu, and the Director of the Interactive Edge Research Team, developing adaptive, realiable, and trusted edge computing. She also works as a Freelance Lead AI Scientist at Finnish company Silo.AI.

PAWEŁ SROKA (Member, IEEE) received the M.Sc. and Ph.D. (Hons.) degrees in telecommunications from the Poznan University of Technology (PUT), Poland, in 2004 and 2012, respectively. He is currently employed as an Assistant Professor at the Institute of Radiocommunications, Faculty of Computing and Telecommunications, PUT. His main research interests include 5G systems, radio resource management, vehicular communications (V2X), cross-layer optimization, and MIMO systems.

SASU TARKOMA (Senior Member, IEEE) received the Ph.D. degree in computer science from the University of Helsinki, in 2006. He is a Professor of computer science with the University of Helsinki. He is also a Visiting Professor of the 6G flagship at the University of Oulu. He has authored four textbooks and has published over 250 scientific articles. He holds ten granted U.S. patents. His research interests include internet technology, distributed systems, data analytics, and mobile and ubiquitous computing.

TINGTING YANG (Member, IEEE) received the B.Sc. and Ph.D. degrees from Dalian Maritime University, China, in 2004 and 2010, respectively. She is currently a Professor at the Pengcheng Laboratory, and also with Dalian Maritime University. Her research interests include network AI, edge intelligence, and space-air-ground-sea integrated networks.