Computer Communications 219 (2024) 128–152

Contents lists available at ScienceDirect

Computer Communications

journal homepage: www.elsevier.com/locate/comcom

Review

A survey on integrated computing, caching, and communication in the cloud-to-edge continuum

Adyson Maia [a,*], Akram Boutouchent [a], Youcef Kardjadja [a], Manel Gherari [b], Ece Gelal Soyak [c], Muhammad Saqib [b], Kacem Boussekar [a], Idil Cilbir [d], Sama Habibi [d], Soukaina Ouledsidi Ali [b], Wessam Ajib [b], Halima Elbiaze [b], Ozgur Erçetin [d], Yacine Ghamri-Doudane [a], Roch Glitho [e]

[a] La Rochelle University, La Rochelle, France
[b] Université du Québec à Montréal, Montreal, Canada
[c] Bahcesehir University, Istanbul, Turkey
[d] Sabanci University, Istanbul, Turkey
[e] Concordia University, Montreal, Canada

ARTICLE INFO

Keywords:
Communication
Computing
Caching
Cloud-to-edge continuum
Next-generation network infrastructures

ABSTRACT

Cloud and edge computing have proposed different functionalities to enable multiple applications requiring different communication, computing, and caching (3C) resources. The upcoming futuristic applications (e.g., metaverse, holographic, and haptic communication) impose further stringent requirements (e.g., ultra-low latency, ultra-high reliability) on the infrastructure. These requirements call for a paradigm shift in the infrastructure architecture where all resource components and owners collaborate from the cloud up to the edge, creating a cloud-to-edge continuum of integrated resources. Furthermore, we argue that artificial intelligence (AI) and collaborative-based decisions are promising techniques to efficiently manage the highly complex architecture that jointly leverages 3C in the continuum. This article presents a comprehensive survey of existing research, including AI and collaborative-based studies, targeting the effective and seamless provision of 3C resources and services in the cloud-to-edge continuum. Through an extensive analysis of driving use cases, the synergy between these three main services is scrutinized to highlight its crucial role in the next-generation network infrastructures (NGNI). Finally, a discussion on the opportunities and challenges brought by integrating 3C in NGNI from different perspectives, including architectural design as well as the regulatory and business aspects, are presented.

1. Introduction

Over the last decade, the communication infrastructure has constantly been evolving to accommodate customers' growing demand in terms of network and value-added services. These newly emerging applications, such as *Metaverse* [1], *Holographic Communication* [2], and *Autonomous Vehicles* [3], require considerable improvements in the overall networking infrastructure. Consequently, these infrastructures evolved to integrate with computing and caching services in order to scale and satisfy the low latency and ultra-high reliability requirements. To this end, new computing paradigms emerged. *Fog Computing* [4] moves the computing and caching resources close to the network edge, improving network quality and enabling latency-sensitive applications. Network operators and telecommunication vendors advertise the *Multi-access Edge Computing (MEC)* [5] capabilities they may offer within their 5G Radio-Access and Core Networks. In recent years, GAFAM (Google, Apple, Facebook, Amazon, and Microsoft) also entered the game, offering so-called smart speakers (e.g., Amazon Echo, Apple HomePod, and Google Home) that can serve as Internet-of-Things (IoT) hubs with *Mist/Skin Computing* capabilities.

Nevertheless, these computing paradigms still consider edge, fog, and cloud separate components. Solutions that integrate edge resources alone are unlikely to be sufficient to meet the end-to-end Quality

* Corresponding author.
E-mail addresses: adyson.magalhaes_maia@univ-lr.fr (A. Maia), akram.boutouchent@univ-lr.fr (A. Boutouchent), youcef.kardjadja@univ-lr.fr (Y. Kardjadja), gherari.manel@courrier.uqam.ca (M. Gherari), ece.gelalsoyak@bau.edu.tr (E.G. Soyak), saqib.muhammad@courrier.uqam.ca (M. Saqib), kacem.boussekar@univ-lr.fr (K. Boussekar), idil.cilbir@sabanciuniv.edu (I. Cilbir), samahabibi@sabanciuniv.edu (S. Habibi), ouledsidi_ali.soukaina@courrier.uqam.ca (S.O. Ali), ajib.wessam@uqam.ca (W. Ajib), elbiaze.halima@uqam.ca (H. Elbiaze), oercetin@sabanciuniv.edu (O. Erçetin), yacine.ghamri@univ-lr.fr (Y. Ghamri-Doudane), glitho@ece.concordia.ca (R. Glitho).

https://doi.org/10.1016/j.comcom.2024.03.005
Received 29 June 2023; Received in revised form 28 January 2024; Accepted 4 March 2024
Available online 5 March 2024

of Service (QoS) requirements of these new emerging applications. Especially when considering use cases such as Industry 4.0 which integrates an important number of deployed, computationally limited IoT devices. Applications that rely only on edge resources cannot handle the enormous amount of data that is generated. Additionally, these paradigms are still loosely linked to the communication and data storage infrastructures, even in 5G networks under deployment. To this end, a promising approach is envisioning a network-wide architecture, utilizing a continuum of resources across the network. Combining the capabilities of edge servers, the cloud, and end-devices creates a hierarchical paradigm known as the *Cloud-to-Edge Continuum*. This paradigm enables integrating 3C resources on different network parts, resulting in improved resource utilization and Quality of Experience (QoE). In next-generation network infrastructures, such as beyond 5G and 6G, close coupling of computing platforms with network infrastructure is expected to support services with strong resource demand. Thus, many distributed and heterogeneous devices belonging to different stakeholders will cooperate to perform services or store data in exchange for a reward. Although still in its infancy, this paradigm shift has already prompted several research initiatives aimed at addressing its challenges. For instance, standardization activities such as the Computing in the Network (COIN) research group [6] in IETF are tackling the integration between computing and networking in the cloud-to-edge continuum. COIN naturally fits within the continuum, where expanded resource distribution and tightly integrated computing-networking capabilities are provisioned from the edge to the cloud infrastructure, including points in between.

To this end, fulfilling the rigorous requirements of the emerging applications necessitates an efficient allocation of caching, computing, and communication resources across the Cloud-to-Edge Continuum. Recent advances in Artificial Intelligence (AI) and Machine Learning (ML) hold significant potential in addressing challenges ranging from traffic prediction to service placement and resource allocation [7]. In recent years, several research efforts studied the joint management of 3C resources, mainly at the edge of the continuum. In this study, we consider the entire Cloud-to-Edge Continuum and emphasize that the allocation and management of 3C resources in such a highly distributed, heterogeneous, and multi-tenant environment are pivotal for realizing the envisaged use cases. We highlight the crucial role that AI/ML techniques can play across diverse domains, facilitated by context-aware agents capable of collaborating to guide the network towards the attainment of shared objectives.

1.1. A new era of massively scalable 3C: Cloud-to-edge continuum

The advent of cloud computing has led to a reduction in the need for local storage on end devices, enabling them to offload computationally intensive tasks to remote servers. Although, in order to mitigate latency issues and improve network resilience in the face of heavy traffic, it is increasingly desirable to transfer computational tasks to devices located in proximity to the end users. This has led to an increased interest in edge computing over recent years, as well as the emergence of other distributed computing paradigms such as fog computing and MEC [8]. The proliferation of Internet of Things (IoT) devices and the corresponding increase in data generation have further contributed to this trend. MEC enables the processing of data generated by devices on servers located at the edge of mobile networks, such as 5G and 6G. On the other hand, fog computing introduces an intermediate computing layer between end-user devices and the cloud by distributing nodes with processing, storage, and memory capabilities throughout the network. However, the stringent service requirements of new and innovative applications may necessitate the use of multiple computing paradigms to meet these requirements. This convergence of computing paradigms has resulted in the concept of the cloud-to-edge continuum.

Cloud-to-edge continuum [8,9] is a distributed computing paradigm that seamlessly integrates communication, computing, and caching resources and services at the network endpoints and at any point along the data path between these endpoints. Fig. 1 is a high-level representation of this paradigm which combine multiple technologies in order to empower the integration of 3C all across the network. This paradigm enables the deployment of distributed applications and services by allowing data preprocessing on MEC servers, further processing by fog computing nodes, and, if necessary, transfer to a cloud data center for big data analytics. The cloud-to-edge continuum integrates 3C on the network devices as well, enabled by In-Network Computing (INC), computing and caching resources on these devices are leveraged to participate in the processing of the offloaded requests [10]. The cloud-to-edge continuum is expected to bring several benefits to distributed applications, including low latency, reduced bandwidth usage, and support for mobility. Furthermore, in the continuum, the development and deployment of distributed applications and services can be enhanced by utilizing microservices and serverless computing [11,12]. Microservices, an architectural model in which an application is decomposed into a set of loosely coupled services that can be developed, deployed, and maintained independently. Serverless computing, on the other hand, is a model in which developers provide event-driven functions to cloud or continuum providers, and the provider seamlessly scales function invocations throughout the continuum to meet demands as event triggers occur.

Given the high heterogeneity, multi-tenancy, and geographic distribution of nodes in the cloud-to-edge continuum infrastructure, automated orchestration of applications and end-to-end management of 3C resources and services are necessary. Additionally, to meet the stringent requirements of new applications, efficient integration of 3C functionalities across the continuum is crucial. Likewise, 3C functionalities can also complement and reinforce each other to further improve the overall performance of future networks. For example, computation results can be cached for future use, thus reducing backhaul congestion and response times. On the other hand, cached content can be transcoded to meet specific user demands better, thus economizing storage spaces [13]. However, those 3C integrations increase the complexity of jointly managing and optimizing communication, caching, and computing services in an agile, flexible, and adaptive manner.

Therefore, automatic optimization using AI and Machine Learning (ML) is a promising approach for optimizing the performance of applications and services within the continuum environment. This can aid in determining the best location to run code, store state, start up new execution environments, and keep utilization high while minimizing costs and meeting performance objectives [14].

1.2. 3C key definitions

In the context of the cloud-to-edge continuum, the integration of communication, caching, and computing can be classified into different types. Specifically, *1C* refers to the independent optimization of the allocation and operation of resources for specific management functions, such as communication, computing, or caching [13,15]. With regard to *communication*, it refers to the ability to transmit and receive data through networks, with the goal of achieving a certain QoS assurance for the end-users in terms of, for example, data rate, power, and latency. With regard to *caching*, it pertains to the storage of a certain amount of data at the network infrastructure. Typically, a caching strategy is employed to determine the contents to cache and their location. Metrics such as caching gain, content hit ratio, and caching diversity can be used to measure the performance of caching systems [16,17]. In terms of *computing*, it refers to the capability to perform computational operations on information flows within the network. These operations can range from simple algebraic or logical operations to more complex computations, such as data aggregation or filtering, image compression, and video transcoding.

In the same context, the joint optimization of the allocation and operation of resources for two functions within the same context is

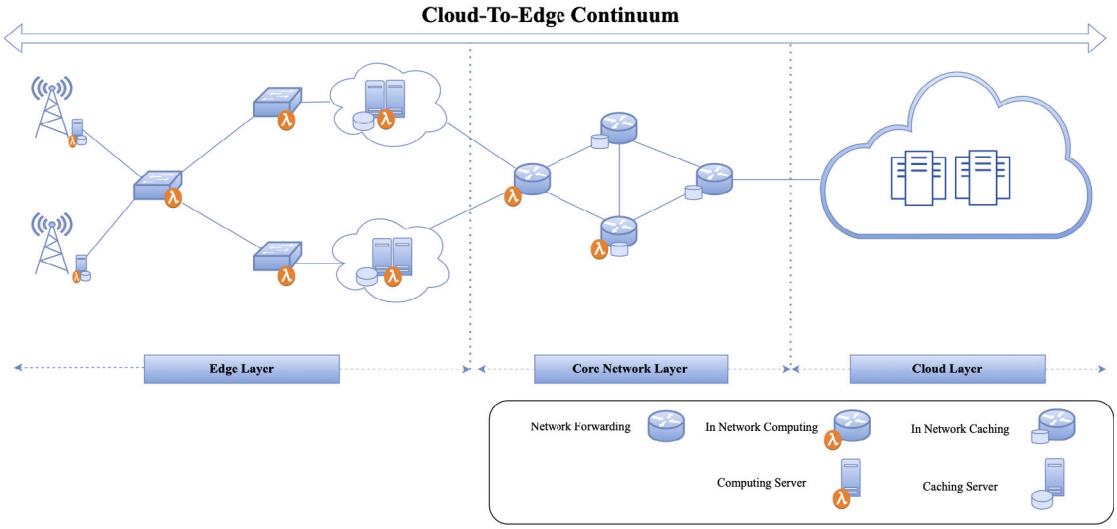

Fig. 1. Cloud-to-edge continuum is a distributed paradigm that promotes the integration of computing, caching, and communication at any point of the network. The cloud layer integrates the cloud computing paradigm with computing, caching, and communication resources. The core network layer is enabled by in-network computing and in-network caching. The edge layer is enabled by edge/fog computing to integrate 3C on the edge of the network.

known as *2C*. Wherein the optimization is aimed at achieving a synergistic effect by considering the interdependencies between functions. An example of this would be optimizing the caching function to operate in conjunction with the computation function, with the goal of maintaining proximity between the content and data and the computation service, such as an application that performs computation and stores its output data locally.

Alternatively, the simultaneous optimization of resource allocation and operation for the three key functions of computing, caching, and communication within a unified framework is considered a *3C integrated framework*. Such integration is very efficient to meet diverse user requirements such as high data rate, low latency, and efficient computation. This may involve strategically positioning computing resources in proximity to user applications while ensuring that caching functions are appropriately located to meet storage needs concerning computation. Additionally, communication functions such as Virtual Network Functions (VNFs) are deployed to minimize traffic congestion and decrease latency between various network components.

1.3. Related surveys

Several surveys related to cloud, fog, and edge computing exist in the literature. We focus on survey papers that deal with 3C integration. Wang et al. [13] address the integration of 3C resources to meet the diversity and intensiveness of user requirements on mobile services. In [18], the authors survey research activities regarding the mobile edge network paradigm integrating computing, caching, and communication resources. The authors focus on edge computing and caching architectures, including the ETSI MEC, fog computing, cloudlet, and edge caching. Mehrabi et al. [19] present a comprehensive survey on device-enhanced MEC, especially mechanisms that jointly utilize the resources of the community of end devices and the installed MEC to provide services to end devices. Mach et al. [20] assess user-oriented use cases in the MEC, considering decisions on computation offloading; allocation of computing resources within the MEC; and mobility management. In [21], the authors conduct a review of resource scheduling issues in edge computing, including computation offloading, 3C resource allocation, and resource provisioning. Research efforts on edge-intelligence reciprocal advantage are surveyed in [22]. Bittencourt et al. [23] provide a literature review on infrastructure, management, and applications aspects of IoT–fog–cloud ecosystems. Ren et al. [24] introduces a comprehensive survey of emerging computing paradigms from the perspective of end-edge–cloud orchestration and presents state-of-the-art research in terms of computation offloading, caching, security, and privacy. The work in [25] presents a review and taxonomy of research studies on low latency service delivery in NGNI. Duan et al. [26] address the field of distributed machine learning in the cloud-to-edge continuum. The authors provide a comprehensive survey on the distributed artificial intelligence empowered by end-user devices, edge computing, and cloud computing. In [27], the authors present a survey on the applications of Deep Reinforcement Learning (DRL) to solve communication and networking issues, such as network access, wireless caching, and data offloading. Zhao et al. [28] study the infrastructure and application issues of edge computing and networking to provide an overview of how edge can be integrated with cloud computing and how edge computing can benefit applications.

It is clear from the discussion mentioned above and Table 1 that most of the related papers that address the 3C paradigm only focus on the edge. Few papers have addressed the 3C integration in the cloud-to-edge continuum without considering the computing dimension of network devices. To the best of our knowledge, this is the first survey article that conducts a systematic overview of 3C integration in the cloud-to-edge continuum. This article revisits several aspects of the state-of-the-art computing paradigm. We lay the foundation for the entire survey by describing the fundamental concepts of 3C and elaborating on how AI and collaboration enable efficient 3C resource integration in NGNI, which includes but is not limited to beyond 5G and 6G networks.

The contributions of the article can be summarized as follows.

- A summary of existing research on 3C integration in the cloud-to-edge continuum, highlighting the role of AI and collaboration in shaping NGNI.
- An extensive analysis through several use cases to highlight the necessity of integrating 3C services to fulfill the stringent requirements of emerging applications.
- A review of challenges brought by leveraging 3C integration in NGNI from different perspectives, including architectural design as well as regulatory and business aspects.
- A summary of opportunities and future research directions such as autonomous networking, AI within the network and adding the control dimension to 3C (4C).

Table 1
Related work classification.

Papers	Year	Context coverage	Location		Role of AI	Contribution
			Edge	Across continuum		
[13]	2018	• Computing • Caching	✓			Integration of Networking, Caching, and Computing in Wireless Systems: A Survey, Some Research Issues, and Challenges
[18]	2017	• Computing • Caching	✓			A Survey on Mobile Edge Networks: Convergence of Computing, Caching, and Communications
[19]	2019	• Computing • Caching	✓			Device-Enhanced MEC: Multi-Access Edge Computing (MEC) Aided by End Device Computation and Caching: A Survey
[20]	2017	• Computing • Caching	✓			Mobile Edge Computing: A Survey on Architecture and Computation Offloading
[22]	2020	• Computing • Caching	✓		✓	Convergence of Edge Computing and Deep Learning: A Comprehensive Survey
[24]	2019	• Caching • Computing		✓		A Survey on End-Edge–Cloud Orchestrated Network Computing Paradigms: Transparent Computing, Mobile Edge Computing, Fog Computing, and Cloudlet
[28]	2019	• Caching	✓			Edge Computing and Networking: A Survey on Infrastructures and Applications
[27]	2019	• Communication • Caching • Computing		✓	✓	Applications of Deep Reinforcement Learning in Communications and Networking: A Survey
[21]	2021	• Communication • Caching • Computing	✓		✓	Resource Scheduling in Edge Computing: A Survey
[23]	2018	• Communication		✓		The Internet of Things, Fog and Cloud continuum: Integration and challenge
[25]	2021	• Caching		✓	✓	Towards Low-Latency Service Delivery in a Continuum of Virtual Resources: State-of-the-Art and Research Directions
[26]	2022	• Caching • Computing	✓		✓	Distributed Artificial Intelligence Empowered by End-Edge-Cloud Computing: A Survey
Our survey	2023	• Communication • Caching • Computing	✓	✓	✓	Towards an Integrated Computing, Caching, and Communication in the Cloud-to-Edge Continuum: State-of-the-Art and Research Challenges

1.4. Survey organization

The rest of the survey is organized based on the rationale outlining of the rich state-of-the-art research contributions that tackle the integration of 3C on the cloud-to-edge continuum. Given the comprehensive scope covered in this survey, Fig. 2 provides an overview of its structural organization. This serves as a helpful tool for readers to navigate efficiently and access sections and content aligned with their interests. Regarding the organization of the survey, we adopted a top-down approach, starting in Section 2 with the presentation of potential use cases and applications facilitated by the integration of computing, caching, and communication on the cloud-to-edge continuum. In Section 3 we detail the enabling technologies for such a paradigm shift mentioning MEC, network softwarization and virtualization. Following, in Section 4, we present a literature review covering research contributions that address resource allocation problems within the context of 3C integration. These contributions are classified based on the specific 3C integration perspective considered in each work, including computation offloading assisted by caching and communication, content caching assisted by computation and communication, and other forms of 3C integration. In contrast, Section 5 presents a literature review of AI and collaborative-based works on 3C integration in the continuum. Notably, this section focuses on the role of AI/ML and collaboration techniques in the integration of 3C across the cloud-to-edge continuum. Section 6 discusses the set of challenges that still need to be addressed by the research community and proposes some research directions to be investigated. Finally, the key lessons learned from the critical appraisal of the literature and our conclusions are offered in Section 7.

2. Selected use cases for integrated 3C

The next-generation networking infrastructures (NGNI) are expected to accompany the new access technologies offered by 5G and 6G. NGNI is envisioned to enable new use cases through the integration of communication, computation, and caching (3C) across the cloud-to-edge continuum. These new use cases include applications with stringent requirements such as remote surgery, autonomous vehicles, holographic telepresence, and are expected to be supported on post-5G mobile networks by seamlessly integrating 3C technologies [29–32].

In this section, we describe several use cases involving selected applications, emphasizing the constraints these applications face on current network infrastructures. For each use case, we elaborate on how the integration of 3C can help to meet the application requirements.

2.1. Multisensory AR/VR/XR applications

The concept of the Metaverse envisions a system where various data sources in the physical world continuously generate input for analytics and visual rendering services, to provide real-time, immersive multimedia experiences [33]. Towards this vision, AR and VR applications are evolving into multi-sensory eXtended Reality (XR) experiences, where the audio and video data is augmented by the multi-modal human sensory data including touch, smell, taste, and even emotion [29,34]. Applications leveraging this multisensory capability particularly utilize haptic interactions, which employ devices like sensors and actuators, enabling the users to feel, touch, and manipulate virtual and remote objects. For instance, remote repair applications could allow technicians

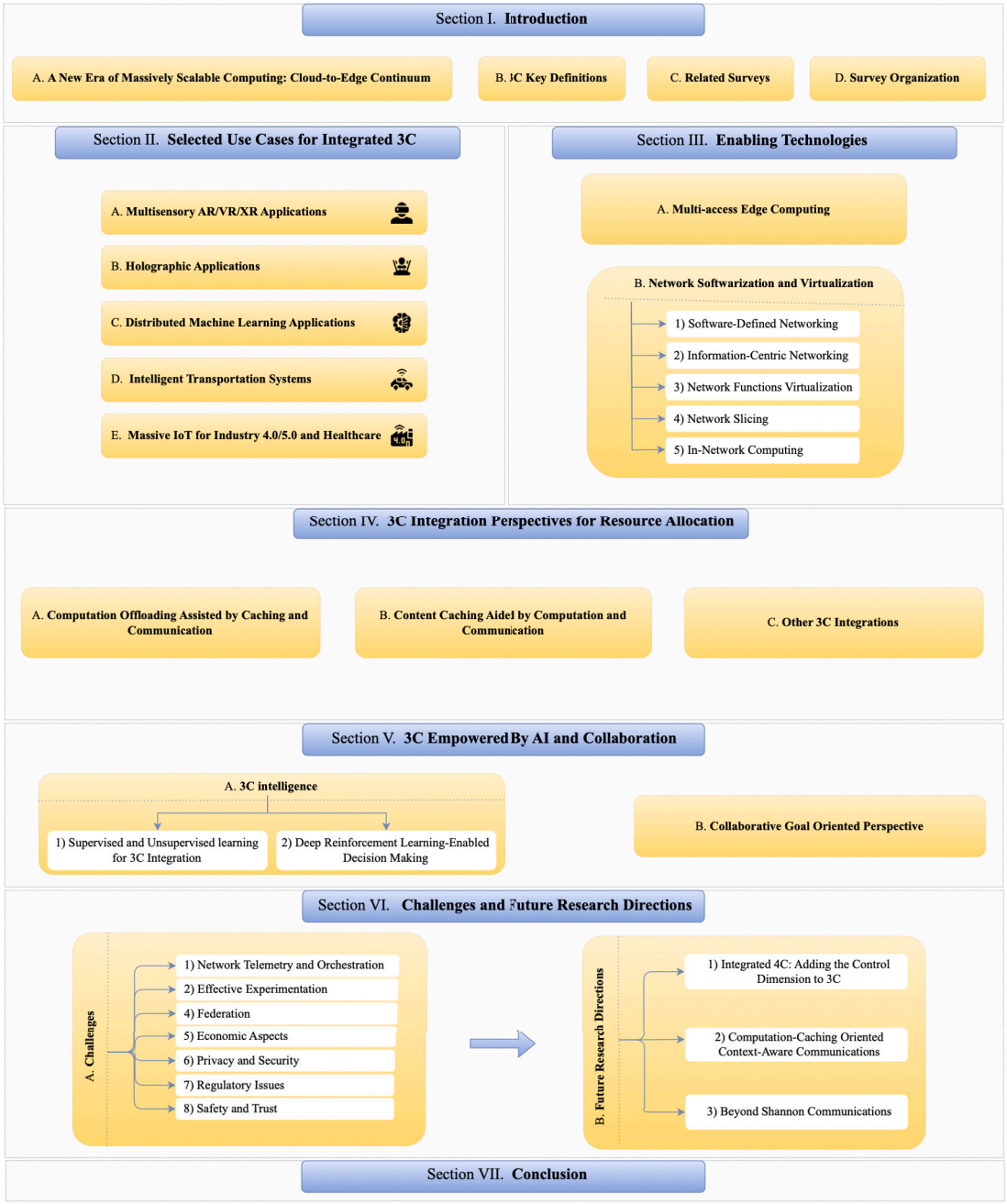

Fig. 2. Structural Overview of the Survey Organization. The primary sections, depicted in blue boxes, encompass various subsections delineated within yellow boxes. Additionally, each sub-section can further be broken down into a series of sub-subsections outlined in white boxes.

to visualize information about the remote machine, such as schematics, diagnostic data, or step-by-step repair instructions, overlaid onto the physical machine [35], doctors could perform surgery on remote patients [30], or therapists can treat phobic patients by systematically exposing them to their feared objects (e.g., snakes or spiders) via a VR environment without any danger [36].

The effective transmission and processing of such multimodal data streams poses a significant challenge, especially due to the dynamic nature of users and network contexts. These applications impose stringent technical requirements, *i.e.,* high data rates (in the order of terabits per second), ultra-low latency (0.1–1 ms), and high reliability (99.999–99.99999%) due to the large amount of data generated and the sensitivity of human perception [34,37]. While current mobile networks are not designed to meet these requirements, NGNI has the potential to address this gap by incorporating 3C management across the cloud-to-edge continuum. This architecture reduces latency and network load by allowing certain data analysis tasks such as aggregation, filtering, and preprocessing, to be performed in-network as data passes through the core. At the network's endpoints, edge nodes can deliver content and offload computation-intensive tasks, such as AR/VR video coding, decoding, or content rendering, thereby further improving latency and throughput.

Fig. 3 illustrates an example procedure for delivering content to an XR application using the integrated 3C architecture. In this scenario, the end-user initiates a request to the edge server for a specific content via their AR/VR headset. If the requested content is present in the cache and is suitable for the user's device specification and preference, it is sent to the user. However, if the content requires adaptation (e.g., reducing the video quality, resizing, transcoding) to meet the

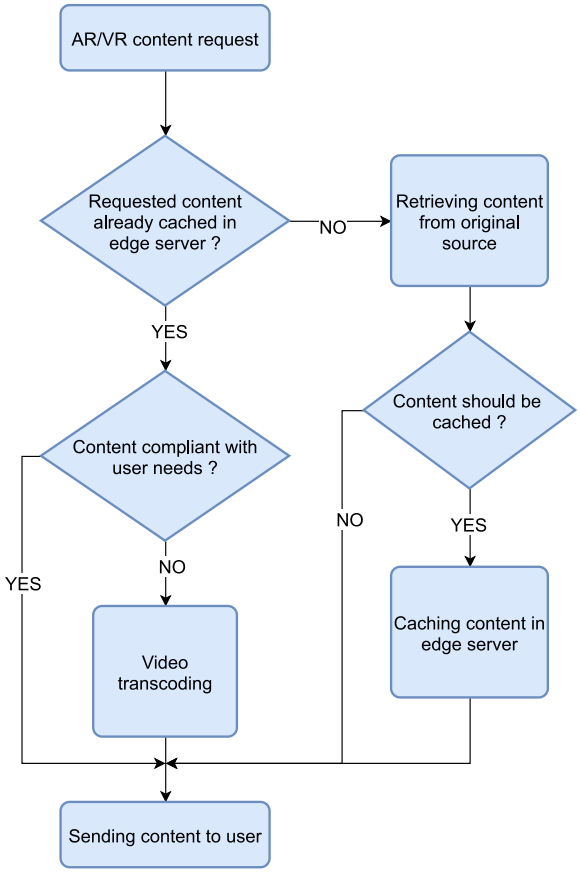

Fig. 3. AR/VR content retrieval procedure.

user's needs, the server performs the necessary computations before transmitting the video. This approach minimizes communication and caching overhead by transmitting only the essential information and not having to store all versions of the video content.

2.2. Holographic applications

Holographic communication entails the real-time transmission and rendering of three-dimensional images to remote locations where they are projected onto the physical world using 3D holographic displays [31]. Applications leveraging holographic communications include immersive education, virtual tours, product prototyping, and gaming.

In holographic telepresence, the goal is to make it feel as if the remote person or object is physically present in the same space as the viewer. Achieving this level of realism requires capturing the scene from various viewpoints to create a three-dimensional representation that can be viewed from different angles. Capturing and synchronizing such a large number of view angles requires an array of cameras or sensors to capture the scene from multiple points, and then advanced imaging technologies are used to synchronize these perspectives for a coherent holographic display.

To enable real-time rendering of holographic content, the network needs to support the *transfer* of large amounts of data quickly and efficiently. Specifically, a colored hologram at 30 fps, without any compression, requires approximately 4 Tb/s [38]. Hence, the data rates required for 3D holographic display lie in the range of hundreds of gigabits per second up to terabits per seconds. At the same time, the latency requirement is in the sub-millisecond range. In contrast to a few view angles with AR/VR, holographic telepresence requires the synchronized view angles. The 3C integration within the cloud-to-edge continuum supports holographic computational tasks, as well as content to be placed close to the end users, reducing the latency and increasing end-to-end data rates.

2.3. Distributed machine learning applications

The integration of AI/ML technologies in everyday systems has become indispensable, enhancing efficiency, personalization, and decision-making across various aspects of daily life. In the process, traditional data-driven systems that rely on centralized cloud servers to collect and train data have raised concerns surrounding user privacy, application latency, and network congestion. To mitigate these issues, distributed and collaborative learning schemes such as federated learning have gained attention. These schemes involve sharing locally trained machine learning models as opposed to transferring raw data and thus, provide benefits in terms of storage, acquisition, and privacy.

Despite the reduced communication overhead with distributed learning schemes, the network remains a significant performance bottleneck for large-scale systems [39,40]; this is mainly due to the large size of exchanged model updates and the potential of straggling workers to slow the training process [40]. To address these challenges, the integration of 3C in the cloud-to-edge continuum can be promising. NGNI can offer a hierarchical architecture for large-scale distributed learning, as shown in Fig. 4. While the computation is performed on the edge nodes close to data sources, aggregation can be performed in intermediate network nodes as data flows toward a centralized server. Additionally, in-network caching can quickly deliver the updated global model to local workers, thus reducing latency and speeding up the training process [39–41]. Indeed, in-network neural networks (i.e., running neural networks directly on the programmable forwarding plane) is another possibility thanks to 3C in the cloud-to-edge continuum [42].

2.4. Intelligent Transportation Systems (ITS)

The spread of urbanization and industrialization has led to an increased interest in Intelligent Transportation Systems (ITS) within both industry and academia. ITS plays a critical role in providing innovative services that assist traffic authorities in efficiently managing traffic flow, making roads safer, less congested, and reducing emissions. Connected and Autonomous Vehicles (CAVs) are expected to be a key component of future smart cities [29,34], as they enable a wide range of applications with their various sensors, communication modules, and on-board units with computing and storage capabilities.

ITS systems are highly dynamic and data-intensive, with stringent delay and reliability requirements. Specifically, reliability greater than 99.999%, latency less than 1 ms, and throughput greater than 100 Gbps are necessary for the safe functioning of these systems. The integration of CAV resources with those of NGNI through the Vehicle-to-Vehicle (V2V) and Vehicle-to-Infrastructure (V2I) collaboration can provide high reliability, low latency, and high throughput for various services. As an example, information that is collected from a vehicle's on-board sensors as well as from the roadside infrastructure can be processed by computing devices deployed at roadside units or mobile base stations; thus, vehicle safety driving may be assisted, or vehicles with potential safety hazards can be quickly detected and a warning signal may be quickly issued. In-vehicle infotainment, such as traffic and weather reports, high-definition maps, and entertainment videos, is another ITS service that can be enhanced by caching content and offloading related processing tasks within the network [43].

2.5. Massive IoT for industry 4.0/5.0 and healthcare

The integration of advanced technologies such as the Internet of Things (IoT), cyber–physical systems, mobile networks, and cloud computing into manufacturing and supply chains is known as Industry 4.0. This digital transformation aims to improve productivity, flexibility, and efficiency through the automation of processes, devices, and systems through a network of sensors and actuators.

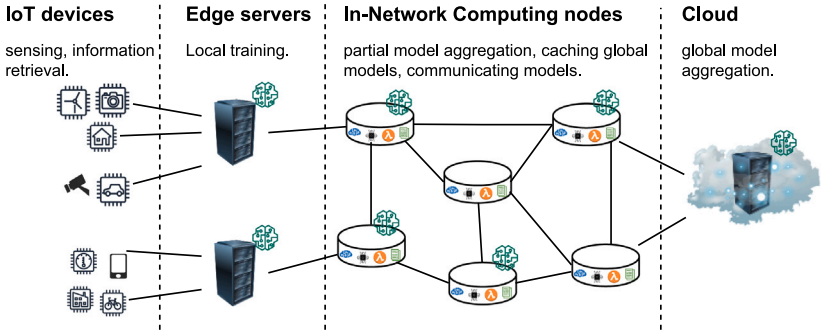

Fig. 4. Distributed Machine Learning in the Cloud-to-Edge Continuum.

Table 2
Summary of use case requirements and the motivation for 3C integration in cloud-to-edge continuum.

Use case	User experience requirements	Service quality requirements	NGNI premise
Multisensory XR applications	• Immersiveness • Alignment of multi-sensory signals	• End-to-End Latency (0.1–1 ms) • Data rate (Tbps)	In-network filtering, aggregation, preprocessing Video coding and content rendering at the edge
Holographic applications	• High-quality visuals • Synchronized viewpoints	• Data rate (Tbps) • Low-latency (0.1 ms) • Low jitter	Holographic computation throughout the network; Content placed close to end users
Distributed ML	• Privacy • Accuracy	• Low-latency (10–100 μs) • Data rate (Gbps)	Hierarchical architecture possible Processing throughout the network In-network caching speeding up training
Intelligent Transportation	• Safety • Responsiveness • Infotainment	• Reliability (> 99.999%) • Low-latency (< 1 ms) • High throughput (> 100 Gbps)	Processing and caching at the edge and within the network
IoT for Industry 4.0 and Healthcare	• Responsiveness	• Latency (< 1 ms) • Data rate (Gbps) • Reliability (> 99.999%)	In-network processing for control decisions In-network caching of health information Offloading tasks to programmable switches and SmartNICs

The automation for Industry 4.0 and beyond requires processing large volumes of data generated by millions of IoT sensors in real-time. This task projects stringent requirements for latency and reliability, which are currently not fully met by existing networks [29,30]. NGNI has the potential to provide the necessary balance of rate, reliability, and latency for Industry 4.0 through the use of the cloud-to-edge continuum. With this approach, data-related operations can be performed as the data flows through the network, reducing network traffic and increasing throughput. Control decisions can be made in-network or at the edge nodes closer to the factory equipment, reducing both the path length and processing latencies [41]. Additionally, by supporting collaboration among the network nodes, distributed control techniques can be used to improve reliability and latency performances.

NGNI also has the potential to revolutionize the healthcare sector by facilitating remote healthcare services such as remote monitoring, diagnosis, prevention, treatment, and even remote surgery. The use of advanced technologies, such as wireless body area networks, edge and cloud computing, in NGNI can enable the quick and reliable transportation and processing of large volumes of medical data from sensor devices placed close to, attached on, or implanted inside the human body. Additionally, the use of artificial intelligence, in-network computing, and caching can enable the real-time provision of health information.

We summarize the requirements of these use cases, as well as the premise with 3C integration for better user experience in that use case, in Table 2.

3. Enabling technologies

The emergence of next-generation networks was the logical response to the growing expectations of end-users and service providers. A variety of new technologies have been explored to deliver infrastructures with greater scalability, control, performance, and isolation for different use cases. Technologies such as multi-access edge computing, network softwarization and virtualization, played a key role in integrating computing, caching, and communication functions across the cloud-to-edge continuum. We explore these technologies in this section while illustrating the importance of each in the convergence of the 3C.

3.1. Multi-access edge computing

With the rise of virtualization techniques such as Virtual Machines (VMs) and containers, the placement of compute and caching resources jointly with networking functions has become more flexible. MEC has become an important component of application-centric networks to support compute-intensive and latency-sensitive applications. The key feature of MEC is to move computing and caching resources as close to the end user as possible, which can lead to a significant decrease in power consumption and latency of mobile applications, and thus remove the main barriers to achieving the NGNI vision [44]. Despite the benefits of edge computing, a MEC server remains less powerful compared to a resourceful cloud data center. Consequently, each MEC server cannot meet all computational and big data demands from user devices when operating independently. A collaboration between distributed MEC servers is then needed to offer a seamless pool of 3C resources at the edge [45]. In this context, several frameworks [45–53] have been proposed to support the integration of computing, caching, and networking functions at the edge of the continuum.

Fig. 5 presents how the integration between computing, caching, and communication can be enabled in the MEC domain. Virtualization of the three functions enables their joint placement as well as their coordination. Such architectural concept has been included in the work of [45], which proposed a framework to jointly optimize computing, caching, and communication as well as control functions. In [45], the proposed system is composed of MEC servers attached to base stations, forming collaboration spaces with near servers, and a cloud data center

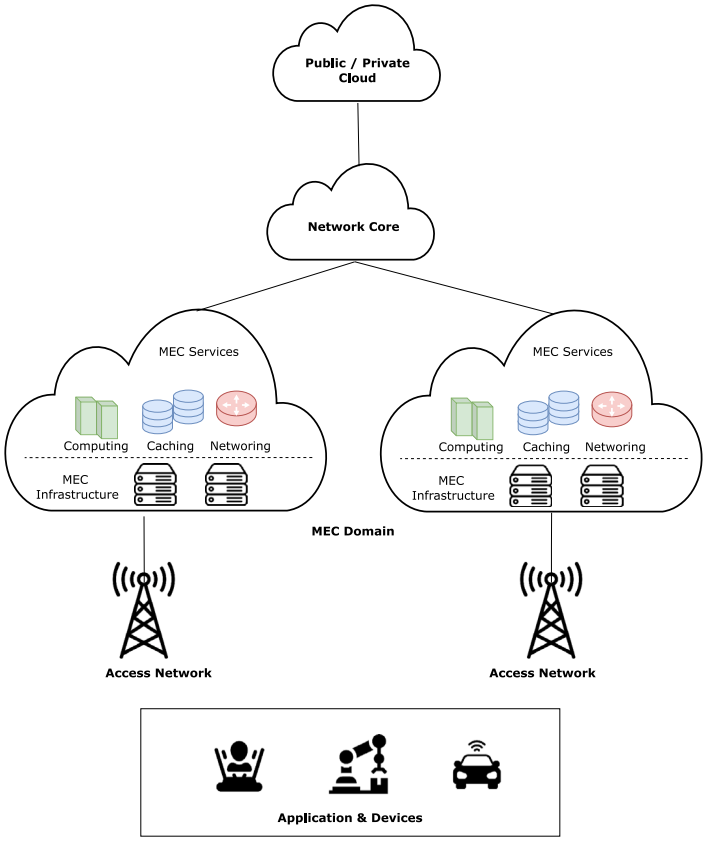

Fig. 5. Edge Computing, Caching and Communication Architecture enabled by MEC.

connected to the servers with fiber links. Moreover, [45] proposed an optimization approach to optimize computation tasks offloading in the context of a big data use case to minimize the bandwidth consumption as well as network latency. In another context, authors in [50] proposed a framework that combines MEC with information-centric networking in a heterogeneous network context. The proposed framework aims at enabling content caching and computing in virtualized networks. In [50], the architecture is based on base stations that provide computing and caching resources, and the objective is to allocate virtualized resources accordingly.

In [46] authors proposed a framework named "Edge CoCaCo" to jointly optimize the management of computing, caching, and communication resources in a MEC-enabled network. A network architecture where base stations attached to edge clouds was designed, and a new concept of computation task caching was presented where the computation result of a task is cached instead of arbitrary content. The proposed architecture enabled task offloading from user devices to the nearest edge cloud, where the edge cloud either makes use of caching resources in case of a hit or computing resources to realize the task. Furthermore, the allocation of base station network resources was optimized in order to minimize the delay. In their following work [47], the framework has been enhanced with a Smartness layer where they consider the use of AI techniques to take smart decisions about tasks offloading. Another layer is considered which is the cloud layer, where end-user tasks can also be offloaded to the cloud data centers.

In another use case, IoT networks have been targeted in the context of integrating 3C in fog-enabled IoT networks. Authors in [48] tackled the joint issues of content caching, computation offloading, and radio resource allocation in a fog-enabled IoT network. The network architecture in [48] consists of a cloud layer with intensive computing and caching resources, a fog layer composed of an edge router linked to a set of fog nodes as well as compute and caching servers. The user equipment's layer includes a set of UEs that may request either a content or a computation task offloading. The objective is to optimize the end-to-end service latency. Task offloading and caching have also been studied by [53], where an architecture has been proposed composed of base stations joined with MEC servers. The computation offloading decision, resource allocation, and content caching strategy are considered in their optimization problem with the objective of maximizing the network's total revenue.

Mist computing is an extension of fog computing, which enables end-user devices to participate in the computing and caching continuum by providing their computing and caching resources. In work [49], authors consider a mobile virtual reality use case, where they proposed a framework that includes VR devices and fog access points both equipped with certain caching and computing capabilities, as well as a central cloud with unlimited resources. The proposed framework targeted the joint optimization of the 3C resources both on VR devices as well as in fog access points, demonstrating the tradeoffs between the 3C functions. In the same context, [52] proposed a framework to optimize VR/AR videos based on an architecture composed of a set of base stations that provides computing and caching resources to be used for deciding on the optimal strategy for streaming/rendering/caching. Where in their latest work, authors in [51] tackled the context of enabling metaverse applications, where they proposed a framework that combines the orchestration of compute, caching, and communication resources in a distributed cloud network. The authors focused on designing a policy in order to optimize the joint decision of routing paths, cache selections, and database placement in order to maximize network performance and enable metaverse applications.

3.2. Network softwarization and virtualization

The integration of various key principles in software and in virtual network architectures have been made possible through the following

technologies: (i) Software-Defined Networking (SDN), (ii) Information Centric Networking (ICN), (iii) Network Function Virtualization (NFV), (iv) network slicing, and (v) In-Network Computing (INC).

3.2.1. Software-defined networking

SDN simplifies traffic engineering by separating the control plane from the data plane, enabling control through a software controller. Specifically, the network's intelligence and decision-making processes are moved to a logically centralized controller, which communicates with the switches and routers in the data plane using, generally, the OpenFlow protocol [54]. This separation of control and data planes enhances network programmability, making it easier to adapt to changing requirements and traffic patterns as well as to develop new protocols and applications. Thus, it is particularly beneficial to the cloud-to-edge continuum, as it facilitates managing and distributing 3C resources.

In [55], the authors proposed a framework called Software Defined Networking, Caching and Computing (SD-NCC), that combines networking, caching, and computing to meet the different application demands efficiently and to enhance the end-to-end system performance; the work also considered the energy consumption during resource placement. In order to address the complexity of resource orchestration in dynamic environments with device heterogeneity and changing resource requirements such as Industrial Internet Of Things (IIoT), an SDN-enabled Resource Management scheme (SDRM) has been proposed [56]. The suggested method computes optimal resource allocation across different IIoT network models and dynamically adjusts resources without violating Service Level Agreements (SLAs). Using a software-defined approach, SDRM manages resources (e.g. memory, power, bandwidth) across the edge–cloud continuum, and thus improves resource orchestration efficiency through dynamic workload balancing and edge–cloud resource utilization.SDN provides a flexible and efficient way to manage and orchestrate network resources, essential for implementing and optimizing Multicast Service Function Chain (SFC) orchestration. Through SDN, the routing of multicast traffic can be dynamically adjusted to accommodate changing network conditions and user fluidity. A comprehensive study on multicast SFC orchestration in SDN/NFV-enabled networks, addressing the challenges of multicast routing and user fluidity is presented in [57].

3.2.2. Information centric networking

To better meet the growing demand for data caching and computing services, some intermediate nodes in the network should be equipped with storage and computing capabilities. ICN is an Internet architecture that prioritizes information by placing it at the core. In-network caching [58], which implements caching at the network layer, is a fundamental feature of ICN. Unlike in IP-based networks, in ICN, the content is assigned a name and can be retrieved without knowing its precise location [59]. Under the aegis of the ICN paradigm, numerous architectures have been put forth [59]. Name Data Networking (NDN) [60] is regarded as the most widely used ICN architecture. NDN uses two types of packets for communication, namely Interest, and Data, and adheres to a pull-based communication model. The consumers send Interest packets with the requested content's name in them. Upon receiving an Interest message, the provider or any other node may respond with content, and all nodes along the path may cache the content per the caching policies.

The consolidation of ICN with the Multi-access Edge Computing (MEC) technology offers new opportunities for realizing the vision of cloud-to-edge continuum [61]. ICN's core functionalities such as in-network caching and content naming can enable high-speed content and service migration between MEC nodes and central cloud systems. Additionally, the pervasive in-network caching reduces the load in the network and increases service quality; and signaling traffic for session mobility management in MEC systems offers reduction in bandwidth consumption.

Aside from these use cases or architectures, combining the ICN and SDN technologies, e.g., utilizing both the benefits of holistic resource management and content-centric communication, can enhance the integration of 3C in the cloud-to-edge continuum. In this regard, [62–65] propose network architectures that integrate 3C based on ICN and SDN principles. Fig. 6 depicts the three planes in the software-defined 3C architecture, namely the data plane, control plane, and application plane.

The data plane is composed of devices (e.g., switches, routers) that are responsible for networking, caching, and computing operations according to the flow tables installed by the control plane. In [62,64], the data plane is based on ICN/NDN, i.e., the forwarding, caching, and processing operations are based on names in packets instead of addresses. On the other hand, [63,65] extend OpenFlow protocol to include storage and compute actions.

The control plane collects network status, including information about the 3C resources, of the data plane to offer a global view of the network to the application plane [62–65]. Moreover, it is also responsible for attaining the network policies defined in the application plane and installing them as control commands on the underlying data plane devices. Aside from the logically centralized controllers proposed earlier [62,64,65], more recently, a distributed control plane for battlefield networking composed of subnets with their own domain controller has been proposed [63].

The application or management plane in [62–65] runs atop the control plane to define network policies. As advocated in [62], it includes not only traditional network applications (e.g., routing, intrusion detection, and load balancing) but also new applications such as content distribution, big data analytics, and distributed computing. As such, the application plane enables flexible and intelligent management at a high level of 3C resources.

3.3. Network functions virtualization

NFV involves virtualizing and decoupling network functions from dedicated hardware devices, transforming them into software-based services. Then, NFV utilizes virtualization environments to run these functions on standard IT hardware, such as servers and switches. This concept is integral to a broader vision aiming for more efficient 3C resource utilization and simplified network management within modern network architecture by offering unparalleled flexibility, scalability [66–68], and cost efficiency [69,70]. For instance, by setting up numerous VMs for running various network functions or performing multiple tasks simultaneously, NFV explicitly enables a single-edge device to provide computing services to countless mobile devices [71].

In another context, microservices and cloud-native technologies are getting a lot of attention from industry leaders and the research community. As a result, network functions are now designed to be deployed as microservices in conjunction with IT and exchange services on the same underlying infrastructure without any constraints. An important illustration of this technology is the new 5G service-based architecture, which includes a set of network functions designed as microservices [72]. This step has several advantages in terms of reduced operating costs and improved quality of service.

3.3.1. Network slicing

The concept of network slicing emerged as a result of the recent advances in cloud computing and NFV. Network slicing is the process of decoupling the physical network infrastructure resources into dedicated logical networks, which facilitates the vertical segmentation of networks, services and applications [73]. These virtual networks are known as "slices" and can be tailored to meet the specific needs of different applications or services. For instance, targeting communication requirements, a slice with high bandwidth and low latency can be dedicated to online gaming and virtual reality applications such as metaverse. Moreover, a network slice can operate as an independent

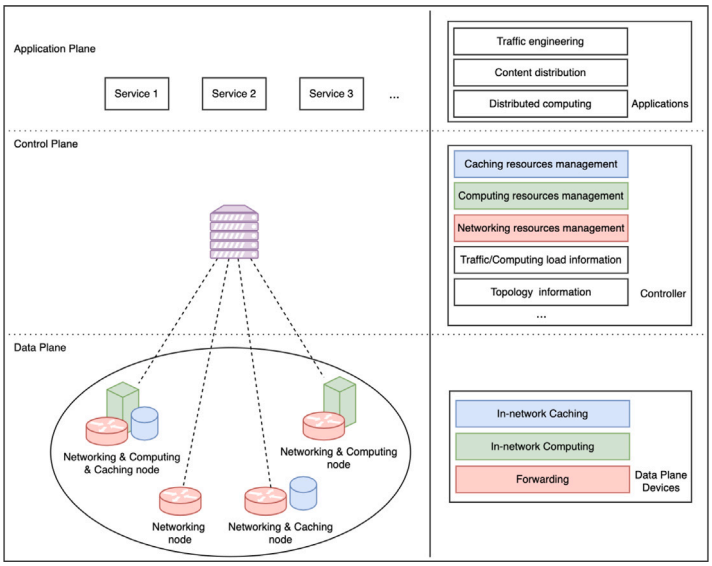

Fig. 6. Software-defined 3C architecture.
Source: Adapted from [62].

end-to-end virtual network within the cloud-to-edge continuum with its own dedicated 3C resources, connectivity, and management policies.

Typically, network slices can span multiple technological and administrative domains by integrating virtualized 3C resources into the cloud-to-edge continuum. In their work [74], the authors studied the allocation of 3C resources in a MEC-enabled network slicing scenario. The infrastructure is decoupled into a set of virtualized network slices that are optimized for different service providers. The authors relied on the use of ML techniques to solve the allocation problem and thus optimize the utility. Emphasizing allocating caching resources, content delivery networks could benefit from network slicing by having a dedicated slice optimized for data throughput and low latency, which are essential for effective caching strategies that ensure faster access to frequently requested data [75,76].

In [77], authors propose a multi-tier metaverse architecture; the metaverse infrastructure provider introduces metaverse applications known as *Metaslices*. These Metaslices consist of a collection of functions called *MetInstances*, each placed based on its specific requirements, with a preference for locating latency-sensitive functions on the first tier, closer to end-users. An essential component of this structural setup is the *Admission Controller*, which utilizes a deep reinforcement learning algorithm to decide which Metaslice requests to accept and how to allocate essential 3C resources. The goal is to optimize resource efficiency and improve QoS.

3.4. In-network computing

SDN has achieved remarkable flexibility through the use of the SDN Controller, which communicates with the data plane using the OpenFlow protocol. However, OpenFlow had some initial limitations, such as varying traffic processing components and data path hardware designs that were dependent on switch vendors. To overcome these limitations, P4 was introduced. A P4 program allows a packet header to be classified and an action to be executed on an incoming packet, which realizes the concept of In-Network Computing. INC integrates computing and caching in the network, where network elements like switches and routers provide connectivity between end devices and the edge infrastructure and between the edge and cloud infrastructure.[10]. The INC technology blurs the lines between cloud, edge, and end-user, creating an infrastructure that facilitates smooth communication, computing, and caching integration. In-network computing can be jointly used with MEC resources in order to provide computing resources across the continuum. In [78], authors proposed a framework in the context of industrial IoT, composed of both INC nodes and MEC nodes, where they leverage a task offloading strategy of sending critical tasks to INC nodes and other tasks to MEC nodes. The authors demonstrated that INC nodes provide more latency performances than typical MEC nodes but, on the other hand, offer less computing capacity. Therefore, they showed that a collaboration between these nodes could be very important.

In a network computing model based on SDN, fragments of application logic can be executed using P4. Central to this model is the SDN controller, which is instrumental in determining the most effective locations for processing and scheduling tasks. This concept is supported by various research studies that explore strategies for optimal processing placement, taking advantage of the centralized control plane of SDN. One such approach is the In-Network Computing Architecture (INCA) [79], which strategically positions requested functions within the network to satisfy QoE requirements. INCA's controller monitors network devices' processing load and the optimal pathways connecting them. Another significant study focuses on the application of SDN and P4 in industrial control systems [80]. This research suggests offloading time-sensitive control tasks from cloud-based systems to local network elements. By doing so, these tasks can be processed with a guaranteed maximum delay, enhancing efficiency and responsiveness in industrial environments. In computer networking, data aggregation and in-network caching are two of the most significant techniques used to implement INC services through SDN. The controller plays a crucial role in this process, as it defines the aggregation tree and sends the corresponding processing rules to the switch in the form of table entries [81]. When data packets arrive, the switch parses and aggregates them based on flow rules defined by the controller. In the paper [82], a framework is proposed to enhance the network throughput and reduce latency by integrating a packet load analysis module and data aggregation processing engine into the switch. In addition to data aggregation, the controller manages state consistency by keeping an updated registration table on the switches in the data center network [41,83].

Table 3 summarizes the enabling technologies and their importance in leveraging 3C in the cloud-to-edge continuum. Note that these technologies do not compete; rather, they complement each other to seamlessly integrate 3C into the continuum.

Table 3
Summary of enabling technologies for 3C in cloud-to-edge continuum.

Enabling technology	3C Coverage	Key features	Importance to the continuum
MEC	• Computing • Caching	• distributed computing and caching at edge servers	• service's QoS improvement • efficient resource usage
SDN	• Communication	• data and control plane separation • logically centralized controller	• holistic resource management • programmable network infrastructure
ICN/NDN	• Communication • Caching	• content-centric communication • in-network caching	• efficient content discovery and delivery • efficient resource usage
NFV	• Communication • Computing	• hardware and network function decoupling • softwarized and virtualized network function	• programmable network infrastructure • simplified resource management
Network Slicing	• Communication	• end-to-end network virtualization	• customized service provisioning • security and isolation enhancement
INC	• Communication • Computing • Caching	• distributed computing and caching at network devices	• service's QoS improvement • efficient resource usage

4. 3C integration perspectives for resource allocation

In most of today's network architectures, intermediate nodes are solicited only for communications. Computing services and cached data remain mainly at the endpoints of the network, where they are often managed separately from each other [13,15]. However, the increasing demand for storage and computing resources has made it imperative to redesign the network architecture and leverage the possible network components in order to realize content- and compute-hungry services. In addition, towards global network optimization, the three main functionalities that cover the service provisioning process, namely, computing, caching, and communication, must be tightly coupled. One of the main challenges in the cloud-to-edge continuum is the joint allocation of computing, caching, and communication resources to establish a network that faces today's blossoming service demands. To this end, different approaches for 3C integration, such as computing assisted by caching and communication and caching aided by computation and communication, can be performed to meet the heterogeneous service demands. Therefore, in this section, we present research studies that facilitate resource allocation across various parts of the cloud-to-edge continuum according to different perspectives of 3C integration.

4.1. Computation offloading assisted by caching and communication

By incorporating computing and caching capabilities into base stations, wireless gateways, and other network devices within the access network, MEC is one of the main enablers of 3C convergence in wireless and mobile systems. 3C integration in these systems contribute in alleviating many challenges, among which is the optimal compute offloading. Offloading tasks from mobile users to external platforms, such as edge and cloud servers, necessitates an optimal management of radio, compute, as well as caching resources for an efficient and effective task execution. For example, given that multiple user devices can repeatedly call tasks over time, storage on MEC servers can be used for service caching, e.g., pre-storing programs, libraries, and databases required for task execution. This can reduce delays caused by task initialization or remote compute service migration for the required programs. However, the limited resource capacity of edge servers requires a careful design of 3C integration strategies. An edge server generally does not have enough storage space for caching all required software for each task. In this regard, the authors in [84] jointly optimized computation offloading, service caching, and bandwidth allocation in MEC to minimize the mean latency experienced by mobile devices by considering the 3C resource capacity of servers and expiration times of cached files. In order to minimize task execution time, [85] investigated task offloading, service caching, and transmit power allocation in a multi-tier computing system composed of relay access points with 3C capabilities and a cloud center. In another work [86], the authors jointly addressed service caching, computation offloading, and 3C resource allocation to minimize computational and latency costs.

Moreover, energy consumption of mobile devices can be reduced by offloading computation-intensive tasks. However, energy usage for data transmission is not negligible. Towards this, a proactive caching of executive codes of tasks at MEC servers was proposed in [87], in order to reduce the weighted sum of task execution delay and users' energy consumption. Similarly, considering the limited caching space at resource-constrained edge servers, the authors in [88] jointly studied computation offloading, service caching, and transmit power allocation to reduce computation delay and users' energy consumption.

In addition to reducing energy usage, some works also impose a deadline constraint for task execution. Liu et al. [89] followed this strategy to improve energy efficiency by jointly optimizing service caching and resource allocation of computation and communication. In another work [90], the authors proposed a joint software caching, computation, and communication mechanism involving multicast software delivery to minimize energy consumption under deadline constraints. Fan et al. [91] aimed to optimize energy efficiency while guaranteeing the task processing delay of mobile devices covered by a heterogeneous network consisting of WiFi access points and base stations with 3C capabilities to process offloaded tasks and cache services. Under radio-resource and computation-latency constraints, [92] proposed a proportional-fair scheduling solution that jointly adjusts 3C resources for tasks offloading and service caching in a MEC system based on Time-Division Multiple Access (TDMA).

Besides service caching, data input of some offloaded tasks can also be cached to avoid duplicate transmission of large input data shared among multiple users, reducing delay and bandwidth usage. In [45], the authors proposed a distributed optimization control algorithm for joint computing, caching, communication, and control in order for data to be offloaded, processed, and cached at MEC servers. Moreover, the goal of the proposed control is to minimize network latency and backhaul bandwidth consumption. Inspired by online gaming applications, [93] considered a bidirectional computation model. In this model, the input data of a task is composed of two parts, one generated by mobile users (e.g., player's location in a game) and the other originating from the Internet (e.g., map information) that can be shared among users. Then, [93] formulated an offloading mechanism that caches remote input data to decrease bandwidth consumption based on Lagrangian relaxation. The authors in [94] also followed the bidirectional data input approach to minimize latency under energy consumption restrictions by optimizing the offloading decision, caching decision, computational resource, and transmit power allocation jointly. In a scenario of low-latency demands of industrial IoT applications, [95] assumed the bidirectional model to formulate a joint problem of user association, computation offloading, and 3C resource allocation for wireless fog networks enabled by caching. Then, an outer approximation algorithm is proposed to solve the formulated problem.

The data result of computation tasks is another aspect that can benefit from caching. By caching the output data of duplicate tasks, it is possible to avoid repeated allocation of 3C resources and skip the computing process for tasks frequently called, reducing resource usage and response time. In [96], the authors studied a multi-objective joint optimization of computation offloading, task results caching, and bandwidth resource allocation to minimize delay and energy consumption. A genetic algorithm is then used to solve the formulated problem. In [97], a joint power control, 3C resource allocation, computation offloading, and data output caching problem is proposed to minimize energy consumption and computing resource usage. The work [50] analyzed virtual resource allocation for 3C in information-centric virtual heterogeneous networks with in-network caching and MEC functionalities. In this work, the allocation strategy includes matching physical and virtual resources and deciding whether or not to cache the contents before or after the computation to maximize revenue. Furthermore, the authors in [50] proposed a distributed algorithm where each base station only solves its subproblem.

Besides the 3C integration in the end-points of the network (edge and cloud), the cloud-to-edge continuum enables 3C integration across the network by extending the functionality of the network to perform additional computation and cache operations. Thus, data can be processed and cached as it progresses through the network to improve system performance. In this regard, the authors in [98] proposed a framework for joint computing, caching, and request forwarding in a data-centric computing network with arbitrary topology and 3C resources available at each node in the network. In the proposed framework, users send requests to process tasks on a required piece of data. A computation request is forwarded through the network until a node decides to perform the task locally. However, this node can only process the task when it has the required data object, either from its own cache or by fetching it from other nodes by sending a data request. In order to decide the computing, caching, and request forwarding actions, the authors designed a distributed algorithm that considers the dynamic demand for both computation and data objects to optimize throughput.

4.2. Content caching aided by computation and communication

In the context of content delivery, a content may have different versions (e.g., format and quality). Caching all variants of a content may overwhelm the storage capacity at edge and network nodes. Hence, the computation capability in the cloud-to-edge continuum can be used on-demand to transform or transcode a cached variant of the required content to a specific requested version. However, this transformation process is generally a computation-intensive task. Therefore, it is necessary to design caching schemes that efficiently utilize 3C resources. In this regard, [99] intended to minimize energy consumption by jointly studying bandwidth provisioning and caching with content transformation tasks (e.g., transcoding, compression, and coding) in software-defined mobile networks with edge computing and caching. Then, a decentralized scheme based on the alternating direction method of multipliers is designed to solve the formulated problem.

A caching strategy with video transcoding is proposed in [100]. It aims to minimize the delay of adaptive bitrate video streaming applications while jointly considering access control, resource allocation, and user preferences. In order to increase system revenue, [101] investigated the optimization of content caching as well as radio and computation resource allocation for video services. As radio resource allocation and caching decisions may happen in different time scales, [101] proposed a two-stage optimization algorithm to accommodate the different optimization cycles. The two different time scales is also taken into account by [102] in the proposition of a greedy-based approach to allocate radio resources, set transmit power, and distribute adaptive bitrate video chunks among users. The proposed approach is designed to maximize users' QoE and reduce backhaul traffic by utilizing information regarding traffic patterns and desired video quality.

Likewise, [103] formulated a two-subproblem decomposition, (*i*) power allocation subproblem and (*ii*) joint caching and computing subproblem, to minimize delivery delay and energy consumption under different video sizes and playback rate requirements of adaptive video streaming. A minimum transmission rate is also a requirement in the 3C scheme designed in [104], to optimize the energy efficiency and delay of edge caching-based video streaming. Continuing in the video streaming context, [105] addressed the QoE improvement by jointly considering network-assisted video rate adaptation and caching, bandwidth provisioning in traffic engineering, and computing resource scheduling. In another work [106], the authors studied a joint optimization of video segment caching and transcoding in MEC servers as well as wireless resource allocation to improve the QoE, system throughput, and backhaul traffic.

In [107], the authors investigated cost-efficient video distribution over next-generation networking infrastructure based on NFV and SDN. Specifically, virtualized 3C resources can be rented to deploy video caching and transcoding services into switch nodes based on the NFV concept. Moreover, with the SDN technology, a centralized routing strategy can be used to control the entire network. Therefore, in order to orchestrate routing, caching, and transcoding functions, [107] proposed a two-step iterative approach. The first phase of the proposed approach maximizes cache hits by optimally allocating 3C resources for a giving routing policy. Then, for a given resource allocation decision, the second phase minimizes networking costs by optimally configuring the routing matrix.

By integrating video transcoding to named function networking, an extension of ICN to perform in-network computation besides content retrieval, [108] formulated an optimization problem for content management (i.e., caching and transcoding) and routing control of ICN routers. Then, a two-step interactive algorithm is developed to improve the cache hit ratio and service delay of adaptive video streaming applications. First, given a routing scheme, the proposed algorithm finds the corresponding optimal configuration of caching and transcoding. Second, based on the previous resource configuration, the algorithm searches the associated optimal routing scheme. Then, these two-step procedures are repeated until convergence.

Delivery of VR content is another application case that can take advantage of 3C integration in edge systems to improve the sense of immersion and interactivity. However, transmitting the whole stereoscopic panoramic VR video leads to unnecessary delay and bandwidth usage. Instead, as a user normally requires a portion of the 360° VR video, only the requested viewpoints need to be transmitted, reducing bandwidth usage and delay. The delivery performance can be further improved by using edge servers to cache VR videos, which can be in a 3D (three-dimensional stereoscopic) or 2D (two-dimensional monocular) video format. A 3D video can be directly rendered by a VR device but requires at least twice larger caching and bandwidth resources than a 2D format. On the other hand, a 2D video must be projected to 3D to be rendered, which can be a computationally-intensive task. Considering this 3C tradeoff, the authors in [49,109] addressed 3C resource allocation to minimize the average latency of VR video delivery while guaranteeing a minimum transmission rate in fog radio access networks. In another work [110], a collaboration of VR devices, MEC servers, and the cloud center is explored to obtain lower end-to-end latency of VR delivery by jointly formulating a caching, power allocation, and video projection offloading problem. Then, a discrete branch-reduce-and-bound algorithm is proposed to solve the formulated problem. A QoE-aware caching, computing and transmission scheme is proposed in [111] to reduce the delivery delay of VR content under the promise of QoE requirements and 3C capabilities of MEC servers. The proposed scheme then uses a dual-problem decomposition to solve the studied problem.

4.3. Other 3C integrations

Some other works investigate 3C resource allocation with the co-existence of computing and caching services, but without considering the direct support of these services with each other to improve performance. In this regard, the work in [112,113] investigated 3C resource allocation, user association, and power control for two kinds of heterogeneous services: (*i*) high-data-rate and (*ii*) computation-sensitive services. Content caching and wireless communication are jointly considered to provide high-data-rate service without the cost of backhaul resources. On the other hand, computation-sensitive services are offloaded to MEC servers to guarantee the delay requirement for users. A distributed algorithm based on the alternating direction method of multipliers is then adopted to obtain optimal solutions. Similarly, Tan et al. [53,114] presented a distributed algorithm for computation offloading, resource allocation, and content caching optimization in a MEC-enabled wireless cellular network in order to maximize revenue. In this network, MEC offers two kinds of services to mobile users: (*i*) execution of offloaded tasks and (*ii*) caching of content from the Internet.

The authors in [115] proposed a distributed control scheme based on game theory by integrating bandwidth allocation, caching splitting, and computation offloading. A base station as leader of a Stackelberg game splits its cache capacity for offloading data and content caching and decides the price for communication and computation services. As followers, user devices monitor their leader's decisions and select their best strategies in a distributed fashion.

By harvesting and collaborating various fog devices via cellular and Device-to-Device (D2D) connections for 3C resource sharing, [116] proposed a fog-enabled 3C resource-sharing framework for energy-efficient IoT data stream processing by solving an energy cost minimization problem under 3C constraints. As the data processing task could involve 3C resource sharing across multiple fog devices, [116] designed an iterative fog clustering formation mechanism and a minimum cost flow transformation to solve the 3C resource sharing problem.

In [117], the authors investigated the placement of applications that process continuous data streams at the edge and across the network, with support for heterogeneous hardware. A stream processing application consists of a set of dependent tasks representing different types of operations on a data stream (e.g., filtering, pre-processing, aggregation, and caching). Moreover, a task can have alternative software and hardware accelerator implementations, and it requires the allocation of multi-dimensional resources (e.g., bandwidth, computing, and storage resources) to be hosted in a network node. Then, [117] analyzed different placement strategies in terms of multiple metrics; latency, throughput, bandwidth, energy, and financial cost.

Table 4 gives a comparison between resource allocation papers discussed in this section according to their 3C integration perspective, optimization objective, and location of the integration. In this table, we can observe that most works jointly address 3C resource allocation on the end-points of the network, indicating that the resource allocation issue on the cloud-to-edge continuum needs to be further studied in order to fully leverage and benefit from integrated 3C services in NGNI.

5. 3C empowered by AI and collaboration

The integration of computing, caching, and communication can be significantly complex to achieve in highly distributed and heterogeneous next-generation networking infrastructures. Given the recent advances in AI/ML algorithms, these AI-based techniques can be considered very promising to use, control, and manage 3C functions. Moreover, collaborative schemes among distributed entities (e.g., computing and network nodes) can further improve the 3C resource management in the edge-to-cloud continuum. Therefore, this section presents a literature review of AI and collaborative-based works on 3C integration in the continuum.

5.1. 3C intelligence

The evolution of network infrastructures has created opportunities to integrate computing, caching, and communication to improve overall network performance. In addition to the complexity of 3C integration, the heterogeneity of network devices and the increasing demand for their capabilities have created the need for better management of multidimensional network resources. Considerable research has been devoted to proposing new algorithms to effectively control and manage 3C resources in a joint manner. Although, traditional techniques and optimization algorithms can be very limited in effectively managing these complex infrastructures [119]. Artificial intelligence, on the other hand, proved to be very efficient in tackling several problems ranging from prediction, classification, and decision-making in networking [7]. Several research works have therefore been proposed to leverage the use of such algorithms in order to reduce the complexity of the 3C integration. Fig. 7 represents a classification of the main branches of machine learning and their possible application based on the reviewed works. In what follows, we investigate the set of AI-enabled works used to efficiently control the integration of computing, caching, and communication resources across the cloud-to-edge continuum.

5.1.1. Supervised and unsupervised learning for 3C integration

The rapid development of AI algorithms has led to the widespread use of machine learning models for solving various problems in different fields. One branch of machine learning, supervised learning, can be used for prediction and classification problems. Another branch, unsupervised learning, is used for solving clustering and feature learning problems. These techniques have been applied in many works to reduce the complexity of integrating 3C resources across the cloud-to-edge continuum.

Regarding resource allocation in network slicing, the authors in [119] propose a network-slicing architecture that employs an artificial neural network-based resource allocation scheme to manage 3C resources. Their approach advocates the use of existing resources and infrastructure to provide personalized services, avoiding the costly and inefficient use of more resources. The neural network used in their solution is trained to classify service types based on the corresponding channel status and service demands. Based on this classification, the 3C resources are effectively allocated into a network slice, improving both network performance and overall user experience.

Another type of supervised learning algorithm has been investigated in [120] to enhance network performance. The authors developed an edge intelligence-empowered Internet of Vehicles (IoV) framework to address the issues of Road Side Unit (RSU) peer offloading, vehicle-to-RSU offloading, and content caching. To solve this multi-objective problem quickly and almost optimally, they utilized an imitation learning-enabled Branch and Bound algorithm. The combination of AI with traditional optimization objective algorithms enabled them to make better decisions regarding computation offloading and content caching. This reduces the total network delay under long-term energy constraints.

In [45], the authors addressed the challenge of 3C resource allocation in a MEC-enabled big data context. The goal was to minimize communication delays among MEC servers while facilitating collaboration between these servers. Thus, the authors proposed an algorithm inspired by an unsupervised machine learning algorithm called the Overlapping k-Means Method (OKM). Such a clustering algorithm enables the creation of collaboration spaces that MEC servers can utilize to efficiently allocate computing, caching, and communication resources.

In recent years, deep learning models emerged in the field of machine learning. Characterized by a high density of neural network layers, these models can outperform typical supervised and unsupervised learning models in performing prediction, classification, and

Table 4
Comparison of resource allocation papers for 3C integration.

Papers	3C Integration			Optimization objective						Location		Proposed approach
	Assisted compute	Assisted caching	Non-assisted	Delay	Energy	Revenue/ cost	Resource usage	QoE	Other	Edge/ End-points	Across continuum	
[84]	✓			✓						✓		Inner convex approximation (NOVA)
[85]	✓			✓						✓		Alternating optimization
[86]	✓			✓		✓				✓		Semidefinite relaxation (SDR) + Alternating optimization
[87]	✓			✓	✓					✓		Alternating optimization
[88]	✓			✓	✓					✓		Mixed integer non-linear programming (MINLP)
[89]	✓				✓					✓		Iterative optimization
[90]	✓				✓					✓		Mixed integer non-linear programming (MINLP)
[91]	✓				✓					✓		Iterative optimization
[92]	✓								fair schedule	✓		Iterative optimization
[45]	✓			✓			✓			✓		Block successive upper bound minimization (BSUM)
[93]	✓						✓			✓		Lagrangian relaxation (LR) + Concave–Convex Procedure (CCCP)
[94]	✓			✓						✓		Block successive upper bound minimization (BSUM)
[95]	✓			✓						✓		Outer approximation algorithm (OAA)
[96]	✓			✓	✓					✓		Non-dominated sorting genetic algorithm II (NSGA-II)
[97]	✓				✓		✓			✓		Successive convex approximation (SCA)
[50]	✓					✓				✓		Alternating direction method of multipliers (ADMM)
[99]		✓			✓					✓		Dual-decomposition + Alternating direction method of multipliers (ADMM)
[100]		✓		✓						✓		Non-convex optimization
[101]		✓				✓				✓		Two-stage optimization
[102]		✓					✓	✓		✓		Mixed integer non-linear programming (MINLP)
[103]		✓		✓	✓					✓		Mixed integer non-linear programming (MINLP)
[104]		✓		✓	✓					✓		Alternating optimization
[105]		✓						✓		✓		Alternating direction method of multipliers (ADMM)
[106]		✓					✓	✓	through-put	✓		Heuristic
[49,109]		✓		✓						✓		Lagrangian dual decomposition + Heuristic
[110]		✓		✓						✓		Discrete branch-reduce-and-bound (DBRB)
[111]		✓		✓						✓		Heuristic + Graph coloring
[112,113]			✓			✓				✓		Alternating direction method of multipliers (ADMM)
[53,114]			✓			✓				✓		Alternating direction method of multipliers (ADMM)
[115]			✓						game payoff	✓		Game theory
[116]			✓		✓					✓		Iterative optimization
[117]			✓	✓	✓	✓	✓		through-put		✓	NA
[107]		✓				✓			cache hit		✓	Iterative optimization

(*continued on next page*)

Table 4 (*continued*).

[108]		✓		✓			cache hit	✓	Iterative optimization
[118]			✓		✓	✓		✓	NA
[98]	✓						through-put	✓	NA

clustering. In the context of realizing intelligent 3C resource control over the cloud-to-edge continuum, deep learning can effectively solve several tasks related to network traffic classification, prediction of user mobility, user traffic, and task popularity. In their works [121, 122], the authors tackled a 3C integration problem in the context of Non-Orthogonal Multiple Access (NOMA) enabled networks. A task offloading problem is considered, the system considers that numerous users can offload their tasks in parallel to close servers. The result of the tasks is cached based on the popularity of the task. The problem is therefore decoupled into task popularity prediction and resource allocation subproblems. To predict task popularity, a deep learning Long-Short Term Memory (LSTM) model is used, which predicts the popularity of the offloaded task. In another work [123], the authors tackled a MEC-assisted caching challenge in which they proposed a cognitive agent that performs caching decisions based on each user's behavioral data. This work aims to reduce stress on MEC servers by improving cache hits and thus user throughput. The authors proposed to use LSTM to predict the trajectory of the user's equipment and the types of services requested. The prediction result is used to generate caching strategy and cache activity in order to reduce the task execution delay. Vehicular networks have also attracted the interest of research community. In [124], the authors studied the context of self-driving cars. They studied a vehicular network with a MEC-enabled RSU, where both the autonomous car and the RSU are equipped with computing and caching capabilities. They proposed two deep learning approaches, first a Convolutional Neural Network (CNN) to predict passenger features from facial images. Additionally, they proposed a Multilayer Perceptron (MLP) model to predict the probability of content being requested on certain servers. Based on these outputs, an optimization model is formulated to efficiently allocate computing, caching, and communication resources. Alternatively, authors in [125] consider a location-aware multi-user task offloading problem, where each user has its own service preferences. A deep learning approach is therefore proposed in order to generate caching replacement decisions. Based on these decisions an optimization model is designed to make resource allocation actions that minimize the total sum of energy.

5.1.2. Deep reinforcement learning-enabled decision making

Although deep learning models can be efficient in prediction and classification problems, they are unsuitable for decision-making problems. To enable joint integration of computing, caching, and communication, efficient decision-making models must be leveraged. In this context, Reinforcement Learning (RL) algorithms are widely used to tackle these challenges. RL models sense their environment to generate appropriate decision-making based on a reward system that improves the actions. These models have been further enhanced by integrating a deep learning model, which predicts the action that will result in the best reward. Deep reinforcement learning has been proven to be efficient in dealing with complex decision-making challenges in the context of resource allocation and task offloading [126].

In the context of vehicular networks, deep reinforcement learning has been used in order to efficiently manage computing, caching, and communication resources. In their works [127,128], the authors tackled the resource allocation problem while considering a set of multiple virtual networks that could span across different domains. These virtual networks include a set of virtual 3C resources related to the base station, RSU, and access point. The challenge addressed involves base station assignment to users, as well as the allocation of caching and computing resources with the goal of minimizing the operator's revenue. To tackle this problem and find a suitable resource allocation policy, the authors employed Deep Q-learning. The authors of [129] considered a 3C resource allocation in a Digital Twin (DT) [130] enabled vehicular network. They consider a DT that receives data from different nodes of the network to construct a virtual view. Their system model is composed of a physical view and a digital twin view, where the entities, their connection, and the data are virtualized in the DT view. A DDPG algorithm is proposed in order to solve the resource allocation problem of vehicle task offloading with the objective of minimizing the network delay. In another work [131], the authors explored a vehicular network enabled by edge computing servers. They constructed a multi-objective optimization problem with the goal of minimizing task execution delay and bandwidth costs. To solve this problem, they proposed a DRL solution based on Deep Deterministic Policy Gradient (DDPG) algorithm. The agent senses the network state and takes actions in the form of offloading decisions, content updating decisions, and RSU bandwidth allocation. In [132], the authors discuss a vehicular social network that is divided into areas covered by different RSUs equipped with caching and computing resources. Vehicles can consider multiple content offloading options, including local and V2V content computing and caching local computing and RSU caching, and RSU caching and computing. To maximize the number of vehicles that receive contents, the resource allocation problem is solved using DDPG. In [133], the resource allocation problem takes into account vehicle mobility. The proposed multi-timescale framework considers a small time-scale and large time-scale environment with a massive number of vehicles and RSUs. The objective is to allocate computing, caching, and communication resources efficiently for vehicle task execution while minimizing resource usage costs. Similarly, the work of [134] considers the mobility of vehicles in an Unmanned Aerial Vehicle (UAV) enabled vehicular network scenario. The resource allocation challenge of vehicle task offloading is proposed to maximize the number of successfully accomplished tasks. The formulated problem is then solved using DDPG, with actions of allocating spectrum, caching resources, and computing resources. In the context of fog-enabled vehicular networks, the authors of [135] addressed the problem of task offloading and service caching. Where service caching involves storing the service code in the fog cache. The main goal of their work is, therefore, to minimize resource energy usage. They designed a solution based on DDPG, which executes actions such as offloading decisions, service caching decisions, and resource allocation to solve the problem.

During the last decade, there has been a significant rise in IoT devices. To support the massive amount of devices, the integration of computing, caching, and communication across the cloud-to-edge continuum is necessary. Intelligent algorithms, such as DRL, have been used to efficiently manage the 3C resources and achieve this objective. In this context, [48] proposed a fog-enabled IoT network composed of a set of fog nodes and an edge router. The edge router makes decisions about the scheduling of user computation offloading and content caching. The authors formulated an optimization problem to minimize the total delay and solved it using an Actor–Critic DRL with a natural policy gradient. They also leveraged the use of an experience replay buffer with target and actor networks to improve the policy. In another work, the authors in [136] consider an information-centric intelligent IoT system. The system includes a data layer composed of a set of IoT devices that use machine learning models to perform their tasks. These devices interact with an ICN layer that performs caching of popular content, as well as edge gateways to train these models. An optimization problem is then modeled to decide on the caching policy

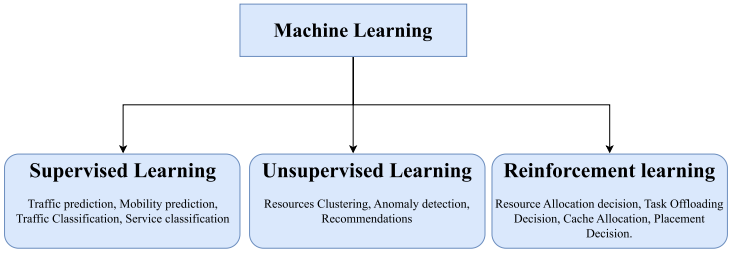

Fig. 7. Classification of Machine Learning Techniques for 3C Integration.

and the training strategy for the models, with the aim of maximizing the system revenue. To solve this problem, a Deep Q Learning (DQN) approach has been used, with a fixed target network and an experience replay memory. In another work, the authors of [137] consider a network model consisting of a set of macro base stations and micro base stations with low coverage. The stations, equipped with computing capacity, can receive tasks offloaded by IoT devices. The challenge is to decide on the offloading type for IoT devices in order to minimize energy consumption. To solve this challenge, an Asynchronous Advantage Actor–Critic (A3C) approach has been leveraged, which involves a set of workers at each micro base station and a global network centralized in the macro base station.

In another context, D2D-enabled network has received interest from researchers for applying DRL to reduce the complexity of 3C integration. The work of [138] considers a video service use case in which the content can be provided by either a base station equipped with caching and computing capabilities or by another device in the same network. Therefore, a DQN method has been used to solve the resource allocation problem. The DRL agent decides on the assignment of video providers, whether to perform video transcoding and the content caching strategy. Similarly, in [139], a D2D-enabled Cloud RAN (C-RAN) network architecture with a fog layer was considered. The authors proposed a resource allocation problem to minimize the total energy consumption of the system. To achieve this, they proposed a deep Q-learning algorithm to allocate the processors of the fog and alternate the communication mode of the user equipment between using D2D and using the Fog RAN.

In a cloud-to-edge continuum, the collaboration between cloud and edge nodes is critical. Allocating 3C resources between the edge and cloud layers is a challenge that was addressed in [140]. The authors considered a network model composed of a set of base stations equipped with limited computing and caching capabilities. They formulated a Mixed Integer Linear Problem (MINLP) to allocate resources to offloaded traffic in a cloud-to-edge cooperation network environment. The authors then proposed a DRL approach to maximize the offloaded traffic. Similarly, the authors in [141] addressed the resource allocation of 3C in a VR-enabled scenario. The VR users request 3D content from base stations equipped with computing and caching resources and linked to a cloud layer. With an objective of minimizing the QoE cost (difference between requested QoE and actual QoE), the authors propose a DRL agent that proactively allocates 3C resources. In another work, the authors of [142] addressed the joint problem of computation offloading, cache decision, transmission power allocation, and CPU frequency allocation for cloud–edge heterogeneous network systems with multiple independent tasks. They leveraged a two-level alternation method framework based on reinforcement learning (RL) and sequential quadratic programming (SQP) with the objective of minimizing energy and execution delay. In the upper level, given the allocated transmission power and CPU frequency, the task offloading decision and cache decision problem are solved using the deep Q-network method. In the lower level, the optimal transmission power and CPU frequency allocation with the offloading decision and cache decision are obtained by using the SQP technique. In a different context, the authors of [121] addressed a problem of task offloading and caching in the context of NOMA-based edge computing and caching. They considered a multi-user cache-aided MEC network with the goal of minimizing the total consumed energy. The authors used LSTM to predict task popularity, and a single-agent Q-learning approach was used to make the offloading decisions. Alternatively, a network slicing resource allocation problem has been investigated in [74]. A system consisting of an infrastructure provider, service provider, and Mobile Virtual Network Operator (MVNO) is considered, where the main objective is to maximize the revenue of the MVNO. To address this problem the authors proposed a twin DDPG approach in order to make slice-level actions and user-level actions. Slice-level actions correspond to MEC resource allocation for different slices, whereas the user-level actions represent the task offloading decisions.

Table 5 gives a comparison between papers discussed in this section according to their 3C integration perspective, objective, and the AI algorithm used.

5.2. Collaborative goal oriented perspective

Independent and isolated decision-making algorithms show poor performance in managing today's network resources. Network nodes may exchange relevant information and collaborate to make better management decisions. To that end, collaborative schemes and techniques are needed to efficiently manage the highly complex and heterogeneous architecture that jointly leverages 3C in the network. Artificial intelligence can play an important role to enable an efficient collaboration to reach a shared and common goal. Following are some of the works that tackled the collaboration aspect of nodes to achieve better network performance.

The authors in [143] acknowledge that, with the complexity of computing, caching, and communication, a single MEC server or node is not enough to manage these multi-dimensional resources while operating independently with no cooperation with its peers. Thus, they highlight collaboration between MEC servers for executing computation tasks and caching data. They propose a collaborative scheme that allocates caching and computational resources to many service requesters based on metrics (i.e., resource demands and payments) to minimize the data exchange between end-users and remote cloud servers and maximize the resource utilization at the edge. Their algorithm groups MEC servers into collaboration spaces of 'F hop count distance' diameter. These collaboration spaces share their resources and regularly exchange information or knowledge on their available resources. Then, they formulated the resource demands generated by service requesters as bids for caching and computing and defined a pricing framework to be used by a single mobile network operator, which allows the service requester to make bids while being constrained by its budget. In another context, the authors of [144] address the problem of request scheduling and service caching in a MEC-enabled IoT network. They model a task offloading system where the service's libraries and databases can be cached in the MEC servers to improve service quality. The authors propose a collaboration scheme between MEC servers and the cloud layer in order to tackle the offloaded requests. The proposed solution is based on DRL which makes resource allocation decisions with the reward of reducing latency. Similarly, the authors of [145] target a task

Table 5
Comparison of AI-driven research works for 3C integration.

Papers	3C Integration			Objective					AI Algorithm		
	Assisted compute	Assisted caching	Non-assisted	Cost/ revenue	Delay	Energy	Resource usage	Other	Supervised learning	Unsupervised learning	RL/ DRL
[120]	✓				✓				Imitation learning		
[45]	✓				✓					Overlapping K-Means for Collaboration Space (OKM-CS)	
[127,128]	✓			✓							Deep Q-learning
[133]	✓			✓			✓				Deep Q-learning
[135]	✓				✓	✓					DDPG
[139]	✓					✓					Deep Q-learning
[142]	✓				✓	✓					Deep Q-learning
[129]	✓				✓						DDPG
[121]		✓				✓			LSTM		Q-learning
[123]		✓			✓				LSTM		Q-learning
[124]		✓			✓				CNN + MLP		
[132]		✓						cache hits			DDPG
[136]		✓		✓							Deep Q-learning
[138]		✓		✓							Deep Q-learning
[140]		✓						through-put			Deep Q-learning
[122]		✓						QoE	LSTM		DDPG
[141]		✓						QoE			DDPG
[125]		✓				✓			Deep NNs		
[119]			✓	✓					Deep NNs		
[131]			✓	✓	✓						DDPG
[134]			✓		✓						DDPG
[48]			✓		✓						Actor–Critic DRL
[137]			✓			✓					Advantage Actor–Critic (A3C)
[74]			✓	✓							Twin-actor DDPG

offloading and resource allocation problem. They introduce the concept of clusters, which refers to a set of geographically close MEC servers that can collaborate to improve network performance. The authors consider tasks that can be divided into a set of small tasks to be realized by the MEC servers, which also decide whether to cache the task results. To tackle this problem, they follow a Meta RL approach that is proved to surpass their initial DRL solution.

In [146], the authors studied the computation offloading problem with caching enhancement for mobile edge cloud networks. To minimize overall execution latency, avoid duplicated computation tasks, and encourage computation result sharing among users, they proposed an optimal offloading with the caching-enhancement scheme (OOCS) for both femto-cloud and MEC scenarios. The authors mainly considered multiple mobile users running the same components of the same application and sharing part of the data (e.g., users running an AR application with the same objective can share some computational tasks, input, and output data). In such a context, the authors based their proposed collaborative offloading and caching scheme on the concept of the joint call graph. Thus, they formulate the offloading and caching relationships between components of an application while allocating users to servers using a coalition formation game. In this way, users connected to the same edge server could cooperate through sharing input/output data and collaboratively executing components of an application.

One of the important challenges of collaborative 3C is privacy and security, which might be addressed with the use of blockchain in future wireless networks [147]. Such technology will allow a secure collaboration between the Cloud-To-Edge Continuum nodes. Such a perspective was addressed by authors in [148,149]. In their work, the authors of [148] address a task offloading and resource allocation problem in the context of vehicular networks. The base stations equipped with edge servers collaborate in a secure way to address the offloaded requests from vehicles. Each base station is considered a blockchain node and participates in the validation of offloaded requests. In this scenario, the authors design an optimization problem that is addressed by the use of the Asynchronous Actor–Critic DRL approach with the objective of reducing energy. Alternatively, the authors of [149] tackle a service caching and task offloading problem in a wireless network. MEC servers close to each other collaborate together to efficiently tackle offloaded requests. Whereas blockchain is integrated into the environment to improve synchronization and evaluate trust between MEC servers. An optimization problem is presented to maximize the service's credibility and reduce latency, which is addressed by a DRL approach that showed important results.

The authors of [150] addressed the collaboration between a set of intelligent agents to effectively allocate 3C resources. They address a task offloading and resource allocation problem in a Fog-RAN (F-RAN) network. The Fog-Access Point (F-AP) receives the offloaded tasks and leverages its equipped computing and caching resources to achieve the tasks. A multi-agent reinforcement learning approach was proposed in order to tackle this problem where the F-AP agents make resource allocation decisions in order to reduce the overall service latency and resource usage.

The authors [151] studied VR video services. As these services necessitate large data transmissions and computing power with stringent delay requirements, it is one of the most suitable scenarios that could

Table 6
Comparison of collaborative-based research works for 3C integration.

Papers	3C Integration			Objective				Proposed approach	Collaborative entities
	Assisted compute	Assisted caching	Non-assisted	Delay	Energy	Resource usage	Other		
[143]			✓			✓		Integer Linear Programming	edge servers
[144]			✓	✓				DRL (PPO)	edge and cloud servers
[145]	✓			✓	✓		QoE	Meta-RL	edge servers
[146]	✓			✓				Game Theory	users
[148]	✓			✓	✓		trust	DRL (A3C)	base stations and edge servers
[149]	✓			✓			trust	DRL	edge servers
[150]	✓			✓				Multi-Agent RL	fog-access points
[151]		✓		✓				Lagrangian Relaxation	VR devices and edge servers
[52]		✓			✓			Dynamic Programming	Small-cell base stations
[152]		✓		✓		✓	QoE	DRL (DDPG)	transcoding-enabled base stations
[153]		✓		✓	✓			Multi-Agent Federated RL	Unmanned Aerial Vehicles

benefit from an efficient 3C integrated framework. For this reason, the authors proposed a user-centric network based on MEC, in which they leverage MEC caching to keep copies of 2D/3D field of view (FOV) files proactively cached. Moreover, they also take advantage of MEC computing capabilities to perform the projection process from 2D FOV to 3D FOV on MEC servers when necessary or locally at the VR device, in a collaborative manner. Then, to achieve a good trade-off of the used multi-dimensional 3C resources of MEC, they proposed a cooperative caching and computation-offloading scheme. Their simulation results showcase the effectiveness of the proposed cooperative approach (i.e., cooperation of caching and computing resources between the VR devices and MEC servers) in minimizing the transmission requirement (i.e., preserving the communication resource) while preserving the required time constraint. In the same context, previous works have also tackled VR video service challenges in many attempts to improve the QoE while jointly considering the 3C capabilities of edge networks. The authors of [52] formulated an optimization framework that allows base stations to cache, render cooperatively, and stream 360° videos while only delivering the remote scene viewpoint needed by the VR user. In another context, the authors of [152] addressed a scenario of 360° video transcoding and caching. The proposed system is composed by a set of base stations that form clusters. These base stations are equipped with computing and caching resources to cache or realize the transcoding of video tiles. They model an optimization problem with the objective of minimizing delay and resource usage, which is addressed with the use of a DDPG approach.Furthermore, the authors of [153] proposed a joint 3C strategy to minimize delay and energy consumption in mobile VR delivery networks. In their system, mMIMO (massive Multiple-Input Multiple-Output)-aided UAV base stations are considered as edge servers serving field of view video segments to users. To satisfy the large number of user requirements, they designed a federated multi-agent deep reinforcement learning (RL) approach. The UAVs, acting as learning agents, collaborate through federated aggregation to improve the overall network's learning performance.

Unlike legacy frameworks of 1C/2C, for which there are many cooperation-based works in the literature [154,155], there is still a lack of initiatives tackling the 3C integrated framework using cooperative solutions between base stations, servers, or even user equipment. Table 6 gives a comparison between papers discussed in this section, according to their 3C integration perspective, objective, proposed approach, and the collaborative entities in their proposal.

6. Challenges and future research directions

Despite the promising opportunities brought by integrating 3C in NGNI, several research challenges remain to be addressed, including architectural design, regulatory, and business aspects. In this section, we present some of these research challenges, followed by a discussion on future research directions.

6.1. Challenges

First, we provide an overview of the challenges in implementing the 3C in the cloud-to-edge continuum, for realizing the use cases described earlier.

6.1.1. Network telemetry and orchestration

The available resources in the cloud-to-edge continuum are diverse and geographically dispersed. These resources can be mobile (e.g., drones) or dynamic (i.e., availability can vary over time [156]). In performing 3C optimization and resource placement over these resources to meet the performance quality requirements of different applications for different users in the network, ubiquitous telemetry mechanisms and autonomous improvements are needed, similar to that modeled by the Intent-Based Networking paradigm [157].

Efficient monitoring plays a vital role in NGNI. *In-band network telemetry* technology emerged upon the developments of SDN and programmable data plane technologies; it relies on collecting hop-by-hop network status information through packets to achieve end-to-end visualization of network services [158,159]. While having advantages such as flexible programming or real-time and path-level network status perception, in-band network telemetry has the challenges of adding overhead to packets [160]; these schemes need to be improved in terms of resource efficiency and accuracy. Additionally, in executing network telemetry operations, the collaboration among the different types of entities with diverse resource constraints in the network must be coordinated [161].

Dynamic adaptation to changing conditions is achieved via network orchestration [51]. For instance, when a sudden spike in demand occurs, computing resources can be dynamically allocated and caching strategies adjusted accordingly. 3C orchestration can leverage a hierarchical design where local optimizations serve system-wide performance goals. Alternatively, the pooling of distributed computing resources can be realized in a peer-to peer fashion where end-user devices form virtual fogs capable of providing their 3C resources to a 5G ecosystem [162]. Moreover, enabling the orchestration of 3C resources across the continuum is further challenging when considering various types of devices (e.g., MEC, in-network computing, and caching nodes) and their constraints in an NGNI environment. Therefore, proper orchestration of heterogeneous 3C resources is critical for scalability as computing and caching capabilities can be distributed across the network.

Ensuring low latency data access is a highly complex problem when the geographical distribution of application traffic demands and the network topology are considered [163]. In addition to dynamically *acquiring* the information of available resources, proactive solutions using learning algorithms could also be employed, to *predict* available resources and/or service requests [164] in a dynamic topology [165]. The learning algorithms could work on individual nodes and optimize for that node, or could work in a distributed fashion to optimize for the network [166,167].

Services that utilize 3C in the continuum need to be elastic; i.e., can adapt to changing demands or resources. Network autonomy, or automation, plays a crucial role in providing flexibility, resilience, cost-efficiency, scalability (across diverse deployments and services), and enhanced security [168]. Network softwarization technologies such as SDN and INC have facilitated a paradigm shift toward *programmable and automated network management*. To fully support the high level of automation required to enable next-generation network use cases such as ITS or Industrial IoT, *autonomous network management* that takes into account the interdependence between computing, caching, communication, and control is crucial, as well as challenging [78,169,170]. *Network tomography*, an emerging monitoring approach, estimates network performance based on external measurements [171]. Network tomography, Knowledge-defined Networking (KDN) [172], and Intent-Based Networking (IBN) [173] are among the requirements and challenges that are needed to realize *autonomous, "self-driving" networks* that can self-adapt to dynamic network changes with minimal to no human intervention [42]. Proper mechanisms leveraging these mechanisms are needed, in order to provide closed-loop management for stability and scalability in the face of changing resource and request dynamics.

6.1.2. Effective experimentation

The integration of computing, caching, and communication infrastructures in the cloud-to-edge continuum involves complex research problems such as resource allocation, load prediction, and function placement. Artificial intelligence has been widely used to address these problems. However, these algorithms can be effective only if they are trained and evaluated in real environments and on large, real data sets. Conversely, these AI-based algorithms, when evaluated against inefficient data, can generate non-optimal decisions or predictions. Therefore, it is very important that real-world data is collected efficiently from integrated 3C infrastructures to achieve effective results [174].

On the other hand, algorithms that address decision-making problems such as reinforcement learning and deep reinforcement learning can also leverage simulators that are assumed to reflect the dynamics and complexity of a real environment. However, an integrated 3C environment can be very complex and very difficult to model. Several events and details are assumed by these simulators, which results in inefficient experimentation of the algorithms, leading to low performance when used in real environments [126].

These challenges can be solved by efficiently collecting real-world data in a cloud-to-edge environment. The collected data can therefore be used to train and test machine learning algorithms and also be fed into simulators to simulate a near real-world environment. However, it is important for the research community to design and implement efficient experimental test labs that can be used to experiment and evaluate the performance of the proposed algorithms in a real-world environment.

6.1.3. Federation

Caching and computing resources will be distributed across the cloud-to-edge continuum, with their inter-connections supporting different data transfer rates. Cooperation among these entities is fundamental in terms of providing the best set of resources to requesting applications, as well as utilizing the available resources most efficiently.

The idea of a federation of computing resources is not new. Cloud-based Mobile Augmentation (CMA) for different user requirements and diverse constraints has been investigated in [175]. While CMA approaches provide augmentation of processing resources on various cloud components, the federation of computing in the cloud-to-edge continuum is a superset of this problem.

A recently proposed paradigm for efficient cloud resource utilization is *sky computing* [176]. The partitioned cloud computing resources and the competition between their owners in the cloud market has led users away from the vision of utility computing. Sky computing has been proposed in order to transform cloud computing into a commodity-like service. In this paradigm, cloud service providers establish reciprocal peering between them in order to offer users vendor-agnostic access to cloud services. Achieving this vision necessitates a compatibility layer that hides the differences in implementation between clouds, an intercloud layer to automatically find the best price/performance for various services, and reciprocal peering to facilitate fast and cost-free data movements. Addressing these challenges is crucial for ensuring a seamless and efficient utilization of diverse cloud resources.

6.1.4. Economic aspects

As the owner of a device with 3C capabilities in the cloud-to-edge continuum may be a single person, an institution, or a company providing other 3C services, the task of including them in a conventional centralized pricing system is not trivial. Considering the geographical dispersion of the nodes, their mobility, and the potential flexibility in the QoS demands, a *dynamic* financial model is more suitable for NGNI. Note that in a simple dynamic pricing model, a computational node, which has processed an offloaded task, bills a computing device usually based on the processing time duration multiplied by a fixed cost per time processing unit. Depending on how far from the edge this device is positioned, it may request an additional incentive that is proportionate to the reduction in latency provided by the shorter network propagation times.

Additionally, nonmonetary incentives should also be considered for providing integrated 3C services in the NGNI as devices with 3C capabilities may belong to different and possibly competing entities. Different dynamic incentive mechanisms, such as those based on deep learning and game theory, can be formulated and applied to incentivize computing nodes to collaborate on the computation of tasks.

6.1.5. Privacy and security

The integration of 3C in the cloud-to-edge continuum will create a dynamic environment that involves the interaction of multiple agents, whether they are service providers, infrastructure providers, or knowledge providers. Such a complex environment may be faced with privacy and security issues, since data and functions may exist in different administrative domains under the supervision of different entities. As an example, the deployment of NGNI in the healthcare sector poses challenges in the areas of security and privacy. The sensitive nature of patients' health information requires the implementation of reliable and secure computing, caching, and communication services. Additionally, trust and security mechanisms are necessary to ensure the authenticity of collected health information and to prevent malicious attacks.

On the one hand, the cloud-to-edge continuum presents a practical solution that can achieve near-ideal protection of privacy. Because there are no protocol boundaries between devices when they are all part of the same broadcast domain, imposing security policies is difficult. By breaking up a single large broadcast domain into multiple smaller broadcast domains, a boundary between devices is created. On the other hand, the integration of computing and caching functions would require more consideration of security aspects. Distributed service placement across the network creates vulnerability to various attacks. The interaction between the different functions must be designed securely to prevent any violation of the computing and caching services.

6.1.6. Regulatory issues

Along with the privacy considerations, the distributed computation of private personal information within the continuum also raises important legal considerations. The Internet is a globally distributed network comprising many voluntarily interconnected autonomous networks. It operates without a central governing body with each constituent network setting and enforcing its own policies. The central pillar in distributed governance of the Internet is *net neutrality*, which states that all data traffic on a network should be treated indiscriminately, and Internet service providers would be restricted from blocking, slowing down, or speeding up the delivery of online content at their discretion. Today, some countries have regulations about net neutrality for networking, but not about cloud computing, which has been the key to enable the development of a plethora of cloud applications, but has disintermediated the telecommunications operators from the value chain.

It must be noted that the regulations will affect the rules on which computing tasks should be transferred to computing devices within the network and which should be retained. This can be handled by tools such as the complex event processing (CEP) tool proposed in [78], by transforming the application-specific function into simple match-action processes for capturing and analyzing data streams to figure out what went wrong (i.e., complex events). Another example is a system created in [177] that pools mid-path storage and connects resources to transport data across the end-to-end path. Along the way from the original server to the client, data can transfer from one hop to the next with this new system. It contains a set of mechanisms that ensures: (*i*) zero packet loss in intermediate network routers, (*ii*) network stability, and (*iii*) higher link utilization.

6.1.7. Safety and trust

A 3C node must ensure that service requests are generated from genuine end-user devices. In fact, a two-way trust relationship is desirable to develop a trusted interaction between 3C nodes and end-user devices. An opinion-based model [178] could be helpful in choosing a 3C service. However, reliability will become an essential factor to be considered. Service-Level Agreement (SLA) between a cloud service and end-users is limited if the service is processed on the edge. A professional and licensed third-party should monitor SLA verification for end-users and small organizations that lack technical capability. Although a service provider offers attributes to measure the trust in the service, at the same time, the verification and monitoring of these attributes are not yet studied in detail. Providing Blockchain-based trust systems is a potential key enabler for 3C service distribution [179].

6.2. Future research directions

Below we elaborate on selected directions for future research, that drive the developments of 3C integration in the continuum.

6.2.1. Integrated 4C: Adding the control dimension to 3C

4C is the extended version of 3C with the addition of the control. The control, referring to a collection of functions that operate at the management plane, manages the computing, caching, and communication functions and controls their functioning to ensure their efficient execution and operation. Examples of these control functions could be lifecycle management, monitoring, orchestration, and service chaining functions. Furthermore, these control functions could be deployed as microservices and placed optimally within the infrastructure alongside the 3C functions. For instance, a lifecycle management function can be deployed close to a compute function to optimize the control and management of that function on the run, eliminating the delay induced when having a remote management entity.

Although some research efforts have started the investigation on 4C [45,51], it is far from being concluded. Besides, there is a need for more studies on intelligent solutions to optimally orchestrate and integrate 4C functionalities across the cloud-to-edge continuum. Indeed, some applications and use cases are expected to be facilitated by the convergence of computing, caching, communication, control, and sensing in 6G networks [29]. Thus, the collaboration and tradeoffs of all 4C dimensions should be further investigated in order to fully leverage 4C integration in next-generation networks.

6.2.2. Computation-caching oriented context-aware communications

Current and future applications such as federated learning and extended reality have unique requirements that contribute in shifting the focus of network infrastructure researchers from communication-only features to multiple services and functions. Authors of [29] advocate that, starting from 6G, network infrastructures must deliver multiple services (i.e., communication, computing, control, and others). That means there must be a tight integration and management of these multiple services while meeting the target performance requirements (e.g., computing latency, stability) for each service. This shift from communication-focused networking has also been supported by the authors of [180], where they proposed a Computation Oriented Communication (COC) service type. The proposed service can provide an acceptable rate-latency-reliability balance on computation tasks, in order to achieve a certain computational accuracy. The same authors also proposed a Contextually Agile eMBB (enhanced Mobile BroadBand) Communication (CAeC) service type, which can adapt the provision of eMBB services based on the network context (i.e., link congestion, network topology, or even surrounding physical location and mobility).

That said, we believe the requirements of NGNI will be even more stringent. By activating INC and the tight integration of 3C/4C, it would be possible for NGNI to offer 3C/4C services throughout the whole cloud-to-edge continuum. The provision of these services must be managed according to the availability of 3C resources, but also according to other aspects related to the network and physical context (i.e., link congestion, network topology, mobility). Given that NGNI will be composed of many distributed and heterogeneous devices belonging to different stakeholders, the latter would also need to cooperate with each other to realize an optimal context-aware 4C service provision. These factors suggest that realizing a computation-caching oriented context-aware communication that offers 4C services throughout the whole continuum is still an open challenge.

6.2.3. Beyond Shannon communications

In their seminal work, Shannon and Weaver [181] categorized communication into three levels based on the purpose of communication; *Level A: Technical*, *Level B: Semantic*, and *Level C: Effectiveness*. At the time, Shannon focused on the technical problem of communicating symbols accurately, and so did the researchers for the following seven decades. Recently, with the increasing differentiation of applications envisioned for 6G, and the soaring need for high bandwidth and low latency, there is an emerging interest in the other two types of communications. Semantic communications focus on transferring only the bits that are significant to convey the meaning. With effective communication, the focus is on delivering only the bits that are significant to ensure the receiver takes the action intended by the sender. With both approaches, the core concern is context-specific rather than the error-free delivery of all information symbols, and it is possible to reduce the number of bits transferred by eliminating the bits that are redundant for the application.

3C continuum establishes an opportunity for semantic and goal-oriented communications in the NGNI. With such communications, the number of bits transferred is reduced, as AI-based methods utilize computational and caching resources available at different segments of the network. One approach for reducing the transferred bits is via *semantic encoding*, i.e., choosing the symbol that maximizes the possibility of receiver's accurate interpretation of the message [182]. Another approach is to determine the next symbol for the sender, based on an estimate of the receiver's understanding of the message. This would

be possible via constructing a knowledge base [183] or a knowledge graph [184] that is shared between the communicating entities. Such a system typically consists of a computational ontology (providing symbolic representations), facts, rules, and constraints [184], as well as a reasoning engine built on the rules and constraints associated with the given application domain [183].

The knowledge, which is available at the receiver to interpret the received symbols, is a dynamic notion. In [185], an AI model is used to represent this relationship, and it is trained via coordinated collaboration among edge servers that perform encoding and decoding based on the commonly shared knowledge base. In such architecture, different edge servers can establish and maintain shared AI models without exposing any of their local data uploaded by the users.

The cognition of the amount and type of knowledge available to the intended recipient, i.e., the interlocutor, is a determining factor in the message symbols that will be transmitted. In [186], contextual reasoning is infused into semantic-native communication, such that each speaker agent runs internal simulations by locally and iteratively reasoning about the communication context of the target receiver. Measuring the communication reliability by the receiver's success in correctly recognizing and interpreting the intended message, the expected semantic representation bit length is derived, which quantifies the extracted effective semantics. As such, post-Shannon communications necessitate the effective coordination of computational and storage, to alleviate the pressure on the communication.

7. Conclusion

In this paper, we have shed light on the main distributed computing landscape in the cloud-to-edge continuum. In particular, we have laid out a comprehensive state-of-the-art literature review on Computing, Caching, and Communication (3C) integration in the cloud-to-edge continuum, highlighting its role in shaping Next-Generation Network Infrastructures (NGNI) and identifying several research challenges that shall be addressed. We began our discussion by analyzing motivations for enabling 3C integration in the cloud-to-edge continuum to fulfill the stringent requirements of emerging applications, which are illustrated by several use cases. Then, we examined the synergy between 3C technologies to enable 3C integration and promote the design of next-generation network architectures. Next, we provided a detailed review of recent works addressing 3C integration and resource allocation across the continuum, highlighting AI/ML and collaborative-based studies. Finally, we concluded with several opportunities and future research directions to achieve the stringent requirements of futuristic applications and successful deployment of NGNI.

CRediT authorship contribution statement

Adyson Maia: Writing – original draft, Writing – review & editing. **Akram Boutouchent:** Writing – original draft, Writing – review & editing. **Youcef Kardjadja:** Writing – original draft, Writing – review & editing. **Manel Gherari:** Writing – original draft, Writing – review & editing. **Ece Gelal Soyak:** Writing – original draft, Writing – review & editing. **Muhammad Saqib:** Writing – original draft. **Kacem Boussekar:** Writing – original draft. **Idil Cilbir:** Writing – original draft. **Sama Habibi:** Writing – original draft. **Soukaina Ouledsidi Ali:** Writing – original draft. **Wessam Ajib:** Conceptualization, Supervision, Writing – original draft, Writing – review & editing. **Halima Elbiaze:** Conceptualization, Supervision, Writing – original draft, Writing – review & editing. **Ozgur Erçetin:** Conceptualization, Supervision, Writing – original draft, Writing – review & editing. **Yacine Ghamri-Doudane:** Conceptualization, Project administration, Supervision, Writing – original draft, Writing – review & editing. **Roch Glitho:** Conceptualization, Supervision, Writing – review & editing.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

No data was used for the research described in the article.

Acknowledgment

This work was partially supported by ANR in France [grant number ANR-19-CHR3-0004], TÜBİTAK in Turkey [grant number 119E353], and FRQNT in Québec – Canada [grant number 287319] through the CHIST-ERA programme (CHIST-ERA-18-SDCDN-005). It was also partially funded by the Région Nouvelle-Aquitaine in France [grant number AAPR2022-2022-17522710] through the GOVERNED-B5G project.

References

[1] Y. Wang, Z. Su, N. Zhang, R. Xing, D. Liu, T.H. Luan, X. Shen, A survey on metaverse: Fundamentals, security, and privacy, IEEE Commun. Surv. Tutor. 25 (1) (2023) 319–352, http://dx.doi.org/10.1109/COMST.2022.3202047.

[2] A. Clemm, M.T. Vega, H.K. Ravuri, T. Wauters, F.D. Turck, Toward truly immersive holographic-type communication: Challenges and solutions, IEEE Commun. Mag. 58 (1) (2020) 93–99, http://dx.doi.org/10.1109/MCOM.001.1900272.

[3] R. Hussain, S. Zeadally, Autonomous cars: Research results, issues, and future challenges, IEEE Commun. Surv. Tutor. 21 (2) (2019) 1275–1313, http://dx.doi.org/10.1109/COMST.2018.2869360.

[4] F. Bonomi, R. Milito, J. Zhu, S. Addepalli, Fog computing and its role in the internet of things, in: Proceedings of the First Edition of the MCC Workshop on Mobile Cloud Computing, MCC '12, Association for Computing Machinery, New York, NY, USA, 2012, pp. 13–16, http://dx.doi.org/10.1145/2342509.2342513.

[5] Y.C. Hu, M. Patel, D. Sabella, N. Sprecher, V. Young, Mobile edge computing a key technology towards 5G, Tech. rep., ETSI, Sophia Antipolis, France, p. 16.

[6] I. Kunze, K. Wehrle, D. Trossen, M.-J. Montpetit, X. de Foy, D. Griffin, M. Rio, Use cases for in-network computing, Internet-draft, Internet Engineering Task Force, 2024, URL https://datatracker.ietf.org/doc/draft-irtf-coinrg-use-cases/05/, work in progress.

[7] R. Boutaba, M.A. Salahuddin, N. Limam, S. Ayoubi, N. Shahriar, F. Estrada-Solano, O.M. Caicedo, A comprehensive survey on machine learning for networking: evolution, applications and research opportunities, J. Int. Serv. Appl. 9 (1) (2018) 16, http://dx.doi.org/10.1186/s13174-018-0087-2.

[8] F. Firouzi, B. Farahani, A. Marinšek, The convergence and interplay of edge, fog, and cloud in the AI-driven Internet of Things (IoT), Inf. Syst. (2021) 101840, http://dx.doi.org/10.1016/j.is.2021.101840.

[9] M. Bendechache, S. Svorobej, P. Takako Endo, T. Lynn, Simulating resource management across the cloud-to-thing continuum: A survey and future directions, Future Internet 12 (6) (2020) http://dx.doi.org/10.3390/fi12060095.

[10] S. Kianpisheh, T. Taleb, A survey on in-network computing: Programmable data plane and technology specific applications, IEEE Commun. Surv. Tutor. (2022) 1, http://dx.doi.org/10.1109/COMST.2022.3213237.

[11] G.A.S. Cassel, V.F. Rodrigues, R. da Rosa Righi, M.R. Bez, A.C. Nepomuceno, C. André da Costa, Serverless computing for internet of things: A systematic literature review, Future Gener. Comput. Syst. 128 (2022) 299–316, http://dx.doi.org/10.1016/j.future.2021.10.020.

[12] X. Kong, Y. Wu, H. Wang, F. Xia, Edge computing for internet of everything: A survey, IEEE Internet Things J. (2022) 1, http://dx.doi.org/10.1109/JIOT.2022.3200431.

[13] C. Wang, Y. He, F.R. Yu, Q. Chen, L. Tang, Integration of networking, caching, and computing in wireless systems: A survey, some research issues, and challenges, IEEE Commun. Surv. Tutor. 20 (1) (2018) 7–38.

[14] J. Schleier-Smith, V. Sreekanti, A. Khandelwal, J. Carreira, N.J. Yadwadkar, R.A. Popa, J.E. Gonzalez, I. Stoica, D.A. Patterson, What serverless computing is and should become: The next phase of cloud computing, Commun. ACM 64 (5) (2021) 76–84, http://dx.doi.org/10.1145/3406011.

[15] H. Liu, Z. Chen, L. Qian, The three primary colors of mobile systems, IEEE Commun. Mag. 54 (9) (2016) 15–21, http://dx.doi.org/10.1109/MCOM.2016.7565182.

[16] W. Zhuang, Q. Ye, F. Lyu, N. Cheng, J. Ren, SDN/NFV-Empowered future IoV with enhanced communication, computing, and caching, Proc. IEEE 108 (2) (2020) 274–291, http://dx.doi.org/10.1109/JPROC.2019.2951169.

[17] F. Yu, T. Huang, Y. Liu, Integrated Networking, Caching, and Computing, CRC Press, 2018.

[18] S. Wang, X. Zhang, Y. Zhang, L. Wang, J. Yang, W. Wang, A survey on mobile edge networks: Convergence of computing, caching and communications, IEEE Access 5 (2017) 6757–6779, http://dx.doi.org/10.1109/ACCESS.2017.2685434.

[19] M. Mehrabi, D. You, V. Latzko, H. Salah, M. Reisslein, F.H.P. Fitzek, Device-enhanced MEC: Multi-access edge computing (MEC) aided by end device computation and caching: A survey, IEEE Access 7 (2019) 166079–166108, http://dx.doi.org/10.1109/ACCESS.2019.2953172.

[20] P. Mach, Z. Becvar, Mobile edge computing: A survey on architecture and computation offloading, IEEE Commun. Surv. Tutor. 19 (3) (2017) 1628–1656, http://dx.doi.org/10.1109/COMST.2017.2682318.

[21] Q. Luo, S. Hu, C. Li, G. Li, W. Shi, Resource scheduling in edge computing: A survey, IEEE Commun. Surv. Tutor. 23 (4) (2021) 2131–2165, http://dx.doi.org/10.1109/COMST.2021.3106401.

[22] X. Wang, Y. Han, V.C.M. Leung, D. Niyato, X. Yan, X. Chen, Convergence of edge computing and deep learning: A comprehensive survey, IEEE Commun. Surv. Tutor. 22 (2) (2020) 869–904, http://dx.doi.org/10.1109/COMST.2020.2970550.

[23] L. Bittencourt, R. Immich, R. Sakellariou, N. Fonseca, E. Madeira, M. Curado, L. Villas, L. DaSilva, C. Lee, O. Rana, The internet of things, fog and cloud continuum: Integration and challenges, Int. Things 3 (2018) 134–155.

[24] J. Ren, D. Zhang, S. He, Y. Zhang, T. Li, A survey on end-edge-cloud orchestrated network computing paradigms: Transparent computing, mobile edge computing, fog computing, and cloudlet, ACM Comput. Surv. 52 (6) (2019) http://dx.doi.org/10.1145/3362031.

[25] J. Santos, T. Wauters, B. Volckaert, F. De Turck, Towards low-latency service delivery in a continuum of virtual resources: State-of-the-art and research directions, IEEE Commun. Surv. Tutor. 23 (4) (2021) 2557–2589, http://dx.doi.org/10.1109/COMST.2021.3095358.

[26] S. Duan, D. Wang, J. Ren, F. Lyu, Y. Zhang, H. Wu, X. Shen, Distributed artificial intelligence empowered by end-edge-cloud computing: A survey, IEEE Commun. Surv. Tutor. 25 (1) (2023) 591–624, http://dx.doi.org/10.1109/COMST.2022.3218527.

[27] N.C. Luong, D.T. Hoang, S. Gong, D. Niyato, P. Wang, Y.-C. Liang, D.I. Kim, Applications of deep reinforcement learning in communications and networking: A survey, IEEE Commun. Surv. Tutor. 21 (4) (2019) 3133–3174, http://dx.doi.org/10.1109/COMST.2019.2916583.

[28] Y. Zhao, W. Wang, Y. Li, C. Colman Meixner, M. Tornatore, J. Zhang, Edge computing and networking: A survey on infrastructures and applications, IEEE Access 7 (2019) 101213–101230, http://dx.doi.org/10.1109/ACCESS.2019.2927538.

[29] W. Saad, M. Bennis, M. Chen, A vision of 6G wireless systems: Applications, trends, technologies, and open research problems, IEEE Netw. 34 (3) (2020) 134–142, http://dx.doi.org/10.1109/MNET.001.1900287.

[30] L. Bariah, L. Mohjazi, S. Muhaidat, P.C. Sofotasios, G.K. Kurt, H. Yanikomeroglu, O.A. Dobre, A prospective look: Key enabling technologies, applications and open research topics in 6G networks, IEEE Access 8 (2020) 174792–174820, http://dx.doi.org/10.1109/ACCESS.2020.3019590.

[31] E.V. Tonkikh, K.D. Burobina, A.A. Shurakhov, Possible applications of sixth generation communication networks, in: 2020 Systems of Signals Generating and Processing in the Field of on Board Communications, 2020, pp. 1–6, http://dx.doi.org/10.1109/IEEECONF48371.2020.9078581.

[32] Z.A. Hmitti, H.B. Ammar, E.G. Soyak, Y. Kardjadja, S. Malektaji, S.O. Ali, M. Rayani, M. Saqib, S. Taghizadeh, W. Ajib, H. Elbiaze, O. Ercetin, Y. Ghamri-Doudane, R. Glitho, SCORING: Towards Smart Collaborative cOmputing, caching and netwoRking paradIgm for Next Generation communication infrastructures, in: 2022 International Conference on Computer Communications and Networks (ICCCN), 2022, pp. 1–10, http://dx.doi.org/10.1109/ICCCN54977.2022.9868940.

[33] J. Crowcroft, P. Eardley, D. Kutscher, E.M. Schooler, Compute-first networking (dagstuhl seminar 21243), Dagstuhl Reports 11 (5) (2021) 54–75.

[34] F. Guo, F.R. Yu, H. Zhang, X. Li, H. Ji, V.C.M. Leung, Enabling massive IoT toward 6G: A comprehensive survey, IEEE Internet Things J. 8 (15) (2021) 11891–11915, http://dx.doi.org/10.1109/JIOT.2021.3063686.

[35] D. Mourtzis, V. Zogopoulos, E. Vlachou, Augmented reality application to support remote maintenance as a service in the robotics industry, Proc. Cirp 63 (2017) 46–51.

[36] N. Promwongsa, A. Ebrahimzadeh, D. Naboulsi, S. Kianpisheh, F. Belqasmi, R. Glitho, N. Crespi, O. Alfandi, A comprehensive survey of the tactile internet: State-of-the-art and research directions, IEEE Commun. Surv. Tutor. 23 (1) (2020) 472–523.

[37] M.S. Elbamby, C. Perfecto, M. Bennis, K. Doppler, Toward low-latency and ultra-reliable virtual reality, IEEE Netw. 32 (2) (2018) 78–84, http://dx.doi.org/10.1109/MNET.2018.1700268.

[38] X. Xu, Y. Pan, P.P.M.Y. Lwin, X. Liang, 3D holographic display and its data transmission requirement, in: 2011 International Conference on Information Photonics and Optical Communications, IEEE, 2011, pp. 1–4.

[39] A. Sapio, M. Canini, C.-Y. Ho, J. Nelson, P. Kalnis, C. Kim, A. Krishnamurthy, M. Moshref, D.R.K. Ports, P. Richtárik, Scaling distributed machine learning with in-network aggregation, 2020, arXiv:1903.06701.

[40] T.Q. Dinh, D.N. Nguyen, D.T. Hoang, P.T. Vu, E. Dutkiewicz, In-network computation for large-scale federated learning over wireless edge networks, 2021, arXiv:2109.10903.

[41] M. Liu, L. Luo, J. Nelson, L. Ceze, A. Krishnamurthy, K. Atreya, Incbricks: Toward in-network computation with an in-network cache, in: Proceedings of the Twenty-Second International Conference on Architectural Support for Programming Languages and Operating Systems, 2017, pp. 795–809.

[42] M.C. Luizelli, R. Canofre, A.F. Lorenzon, F.D. Rossi, W. Cordeiro, O.M. Caicedo, In-network neural networks: Challenges and opportunities for innovation, IEEE Netw. 35 (6) (2021) 68–74.

[43] R.A. Dziyauddin, D. Niyato, N.C. Luong, A.A.A. Mohd Atan, M.A. Mohd Izhar, M.H. Azmi, S. Mohd Daud, Computation offloading and content caching and delivery in vehicular edge network: A survey, Comput. Netw. 197 (2021) 108228, http://dx.doi.org/10.1016/j.comnet.2021.108228.

[44] L.N. Huynh, E.-N. Huh, Envisioning edge computing in future 6G wireless networks, in: 2021 Fifth World Conference on Smart Trends in Systems Security and Sustainability (WorldS4), IEEE, 2021, pp. 307–311.

[45] A. Ndikumana, N.H. Tran, T.M. Ho, Z. Han, W. Saad, D. Niyato, C.S. Hong, Joint communication, computation, caching, and control in big data multi-access edge computing, IEEE Trans. Mob. Comput. 19 (6) (2020) 1359–1374, http://dx.doi.org/10.1109/TMC.2019.2908403.

[46] M. Chen, Y. Hao, L. Hu, M.S. Hossain, A. Ghoneim, Edge-cocaco: Toward joint optimization of computation, caching, and communication on edge cloud, IEEE Wirel. Commun. 25 (3) (2018) 21–27, http://dx.doi.org/10.1109/MWC.2018.1700308.

[47] Y. Hao, Y. Miao, L. Hu, M.S. Hossain, G. Muhammad, S.U. Amin, Smart-edge-cocaco: AI-enabled smart edge with joint computation, caching, and communication in heterogeneous IoT, IEEE Netw. 33 (2) (2019) 58–64, http://dx.doi.org/10.1109/MNET.2019.1800235.

[48] Y. Wei, F.R. Yu, M. Song, Z. Han, Joint optimization of caching, computing, and radio resources for fog-enabled IoT using natural actor–critic deep reinforcement learning, IEEE Internet Things J. 6 (2) (2019) 2061–2073, http://dx.doi.org/10.1109/JIOT.2018.2878435.

[49] T. Dang, M. Peng, Joint radio communication, caching, and computing design for mobile virtual reality delivery in fog radio access networks, IEEE J. Sel. Areas Commun. 37 (7) (2019) 1594–1607, http://dx.doi.org/10.1109/JSAC.2019.2916486.

[50] Y. Zhou, F.R. Yu, J. Chen, Y. Kuo, Resource allocation for information-centric virtualized heterogeneous networks with in-network caching and mobile edge computing, IEEE Trans. Veh. Technol. 66 (12) (2017) 11339–11351, http://dx.doi.org/10.1109/TVT.2017.2737028.

[51] Y. Cai, J. Llorca, A.M. Tulino, A.F. Molisch, Joint compute-caching-communication control for online data-intensive service delivery, IEEE Trans. Mob. Comput. (2023).

[52] J. Chakareski, VR/AR immersive communication: Caching, edge computing, and transmission trade-offs, in: Proceedings of the Workshop on Virtual Reality and Augmented Reality Network '17, Association for Computing Machinery, New York, NY, USA, 2017, pp. 36–41, http://dx.doi.org/10.1145/3097895.3097902.

[53] C. Wang, C. Liang, F.R. Yu, Q. Chen, L. Tang, Computation offloading and resource allocation in wireless cellular networks with mobile edge computing, IEEE Trans. Wireless Commun. 16 (8) (2017) 4924–4938, http://dx.doi.org/10.1109/TWC.2017.2703901.

[54] I. Alam, K. Sharif, F. Li, Z. Latif, M.M. Karim, S. Biswas, B. Nour, Y. Wang, A survey of network virtualization techniques for internet of things using SDN and NFV, ACM Comput. Surv. 53 (2) (2020) http://dx.doi.org/10.1145/3379444.

[55] Q. Chen, F.R. Yu, T. Huang, R. Xie, J. Liu, Y. Liu, Joint resource allocation for software-defined networking, caching, and computing, IEEE/ACM Trans. Netw. 26 (1) (2018) 274–287.

[56] J. Okwuibe, J. Haavisto, I. Kovacevic, E. Harjula, I. Ahmad, J. Islam, M. Ylianttila, SDN-enabled resource orchestration for industrial IoT in collaborative edge-cloud networks, IEEE Access 9 (2021) 115839–115854.

[57] H. Li, L. Wang, Z. Zhu, Y. Chen, Z. Lu, X. Wen, Multicast service function chain orchestration in SDN/NFV-Enabled networks: Embedding, readjustment, and expanding, IEEE Trans. Netw. Serv. Manag. (2023).

[58] Q. Wang, Y. Lu, E. Xu, J. Li, Y. Chen, J. Shu, Concordia: Distributed shared memory with {in-network} cache coherence, in: 19th USENIX Conference on File and Storage Technologies (FAST 21), 2021, pp. 277–292.

[59] B. Ahlgren, C. Dannewitz, C. Imbrenda, D. Kutscher, B. Ohlman, A survey of information-centric networking, IEEE Commun. Mag. 50 (7) (2012) 26–36.

[60] A. Afanasyev, J. Burke, T. Refaei, L. Wang, B. Zhang, L. Zhang, A brief introduction to named data networking, in: MILCOM 2018-2018 IEEE Military Communications Conference, MILCOM, IEEE, 2018, pp. 1–6.

[61] G. Gür, A. Kalla, C. de Alwis, Q.-V. Pham, K.-H. Ngo, M. Liyanage, P. Porambage, Integration of ICN and MEC in 5G and beyond networks: Mutual benefits, use cases, challenges, standardization, and future research, IEEE Open J. Commun. Soc. 3 (2022) 1382–1412.

[62] Q. Chen, F.R. Yu, T. Huang, R. Xie, J. Liu, Y. Liu, An integrated framework for software defined networking, caching, and computing, IEEE Netw. 31 (3) (2017) 46–55, http://dx.doi.org/10.1109/MNET.2017.1600083NM.

[63] Q. Chen, H. Wang, N. Liu, Integrating networking, storage, and computing for resilient battlefield networks, IEEE Commun. Mag. 57 (8) (2019) 56–63, http://dx.doi.org/10.1109/MCOM.2019.1900186.

[64] M. Amadeo, C. Campolo, G. Ruggeri, A. Molinaro, A. Iera, SDN-managed provisioning of named computing services in edge infrastructures, IEEE Trans. Netw. Serv. Manag. 16 (4) (2019) 1464–1478, http://dx.doi.org/10.1109/TNSM.2019.2945497.

[65] R. Huo, F.R. Yu, T. Huang, R. Xie, J. Liu, V.C. Leung, Y. Liu, Software defined networking, caching, and computing for green wireless networks, IEEE Commun. Mag. 54 (11) (2016) 185–193, http://dx.doi.org/10.1109/MCOM.2016.1600485CM.

[66] B. Li, Y. Wang, R. Wang, C. Tai, R. Iyer, Z. Zhou, A. Herdrich, T. Zhang, A. Haj-Ali, I. Stoica, et al., RLDRM: Closed loop dynamic cache allocation with deep reinforcement learning for network function virtualization, in: 2020 6th IEEE Conference on Network Softwarization (NetSoft), IEEE, 2020, pp. 335–343.

[67] J. Chen, J. Chen, R. Hu, H. Zhang, QMORA: A Q-learning based multi-objective resource allocation scheme for NFV orchestration, in: 2020 IEEE 91st Vehicular Technology Conference (VTC2020-Spring), IEEE, 2020, pp. 1–6.

[68] A. El Amine, O. Brun, A game-theoretic algorithm for the joint routing and VNF placement problem, in: NOMS 2022-2022 IEEE/IFIP Network Operations and Management Symposium, IEEE, 2022, pp. 1–9.

[69] J. Tao, Z. Lu, Y. Chen, J. Wu, P. Yu, C. Lei, Adaptive VNF scaling approach with proactive traffic prediction in NFV-enabled clouds, in: Proceedings of the ACM Turing Award Celebration Conference-China, 2021, pp. 166–172.

[70] Y. Yu, X. Bu, K. Yang, H.K. Nguyen, Z. Han, Network function virtualization resource allocation based on joint benders decomposition and ADMM, IEEE Trans. Veh. Technol. 69 (2) (2019) 1706–1718.

[71] Y.C. Hu, M. Patel, D. Sabella, N. Sprecher, V. Young, Mobile edge computing–a key technology towards 5G, ETSI white paper, 11, (11) 2015, pp. 1–16.

[72] J.B. Moreira, H. Mamede, V. Pereira, B. Sousa, Next generation of microservices for the 5G service-based architecture, Int. J. Netw. Manag. 30 (6) (2020) e2132, http://dx.doi.org/10.1002/nem.2132, arXiv:https://onlinelibrary.wiley.com/doi/pdf/10.1002/nem.2132, URL https://onlinelibrary.wiley.com/doi/abs/10.1002/nem.2132, e2132 nem.2132.

[73] S.D.A. Shah, M.A. Gregory, S. Li, Cloud-native network slicing using software defined networking based multi-access edge computing: A survey, IEEE Access 9 (2021) 10903–10924, http://dx.doi.org/10.1109/ACCESS.2021.3050155.

[74] Z. Wang, Y. Wei, F.R. Yu, Z. Han, Utility optimization for resource allocation in multi-access edge network slicing: A twin-actor deep deterministic policy gradient approach, IEEE Trans. Wireless Commun. 21 (8) (2022) 5842–5856, http://dx.doi.org/10.1109/TWC.2022.3143949.

[75] M. Rayani, A. Ebrahimzadeh, R.H. Glitho, H. Elbiaze, Ensuring profit and QoS when dynamically embedding delay-constrained ICN and IP slices for content delivery, IEEE Trans. Netw. Sci. Eng. 9 (2) (2021) 769–782.

[76] I. Benkacem, M. Bagaa, T. Taleb, Q. Nguyen, T. Toshitaka, T. Sato, Integrated ICN and CDN slice as a service, in: 2018 IEEE Global Communications Conference, GLOBECOM, IEEE, 2018, pp. 1–7.

[77] N.H. Chu, D.T. Hoang, D.N. Nguyen, K.T. Phan, E. Dutkiewicz, D. Niyato, T. Shu, Metaslicing: A novel resource allocation framework for metaverse, IEEE Trans. Mob. Comput. (2023).

[78] T. Mai, H. Yao, S. Guo, Y. Liu, In-network computing powered mobile edge: Toward high performance industrial IoT, IEEE Netw. 35 (1) (2021) 289–295, http://dx.doi.org/10.1109/MNET.021.2000318.

[79] Y. Tokusashi, H.T. Dang, F. Pedone, R. Soulé, N. Zilberman, The case for in-network computing on demand, in: Proceedings of the Fourteenth EuroSys Conference 2019, 2019, pp. 1–16.

[80] A.A. Albalawi, A. Chakraborti, C. Westphal, D. Kutscher, J. He, Q. Hoole, INCA: An architecture for in-network computing, in: Proceedings of the 1st ACM CoNEXT Workshop on Emerging in-Network Computing Paradigms, ENCP '19, Association for Computing Machinery, New York, NY, USA, 2019, pp. 56–62.

[81] A. Sapio, I. Abdelaziz, A. Aldilaijan, M. Canini, P. Kalnis, In-network computation is a dumb idea whose time has come, in: Proceedings of the 16th ACM Workshop on Hot Topics in Networks, in: HotNets-XVI, Association for Computing Machinery, New York, NY, USA, 2017, pp. 150–156.

[82] F. Yang, Z. Wang, X. Ma, G. Yuan, X. An, SwitchAgg: A further step towards in-network computation, 2019, arXiv preprint arXiv:1904.04024.

[83] X. Jin, X. Li, H. Zhang, R. Soulé, J. Lee, N. Foster, C. Kim, I. Stoica, Netcache: Balancing key-value stores with fast in-network caching, in: Proceedings of the 26th Symposium on Operating Systems Principles, 2017, pp. 121–136.

[84] C.-L. Chen, C.G. Brinton, V. Aggarwal, Latency minimization for mobile edge computing networks, IEEE Trans. Mob. Comput. (2021) 1, http://dx.doi.org/10.1109/TMC.2021.3117511.

[85] K. Wang, W. Chen, J. Li, Y. Yang, L. Hanzo, Joint task offloading and caching for massive MIMO-aided multi-tier computing networks, IEEE Trans. Commun. 70 (3) (2022) 1820–1833, http://dx.doi.org/10.1109/TCOMM.2022.3142162.

[86] G. Zhang, S. Zhang, W. Zhang, Z. Shen, L. Wang, Joint service caching, computation offloading and resource allocation in mobile edge computing systems, IEEE Trans. Wireless Commun. 20 (8) (2021) 5288–5300, http://dx.doi.org/10.1109/TWC.2021.3066650.

[87] Z. Chen, Z. Zhou, C. Chen, Code caching-assisted computation offloading and resource allocation for multi-user mobile edge computing, IEEE Trans. Netw. Serv. Manag. 18 (4) (2021) 4517–4530, http://dx.doi.org/10.1109/TNSM.2021.3103533.

[88] S. Bi, L. Huang, Y.-J.A. Zhang, Joint optimization of service caching placement and computation offloading in mobile edge computing systems, IEEE Trans. Wireless Commun. 19 (7) (2020) 4947–4963, http://dx.doi.org/10.1109/TWC.2020.2988386.

[89] P. Liu, G. Xu, K. Yang, K. Wang, X. Meng, Jointly optimized energy-minimal resource allocation in cache-enhanced mobile edge computing systems, IEEE Access 7 (2019) 3336–3347, http://dx.doi.org/10.1109/ACCESS.2018.2889815.

[90] W. Wen, Y. Cui, T.Q.S. Quek, F.-C. Zheng, S. Jin, Joint optimal software caching, computation offloading and communications resource allocation for mobile edge computing, IEEE Trans. Veh. Technol. 69 (7) (2020) 7879–7894, http://dx.doi.org/10.1109/TVT.2020.2993359.

[91] W. Fan, J. Han, Y. Su, X. Liu, F. Wu, B. Tang, Y. Liu, Joint task offloading and service caching for multi-access edge computing in WiFi-cellular heterogeneous networks, IEEE Trans. Wireless Commun. 21 (11) (2022) 9653–9667, http://dx.doi.org/10.1109/TWC.2022.3178541.

[92] S.-W. Ko, S.J. Kim, H. Jung, S.W. Choi, Computation offloading and service caching for mobile edge computing under personalized service preference, IEEE Trans. Wireless Commun. 21 (8) (2022) 6568–6583, http://dx.doi.org/10.1109/TWC.2022.3151131.

[93] L. Zhang, Y. Sun, Z. Chen, S. Roy, Communications-caching-computing resource allocation for bidirectional data computation in mobile edge networks, IEEE Trans. Commun. 69 (3) (2021) 1496–1509, http://dx.doi.org/10.1109/TCOMM.2020.3041343.

[94] L.N.T. Huynh, Q.-V. Pham, T.D.T. Nguyen, M.D. Hossain, Y.-R. Shin, E.-N. Huh, Joint computational offloading and data-content caching in NOMA-MEC networks, IEEE Access 9 (2021) 12943–12954, http://dx.doi.org/10.1109/ACCESS.2021.3051278.

[95] R. Basir, S.B. Qaisar, M. Ali, M. Naeem, K.C. Joshi, J. Rodriguez, Latency-aware resource allocation in green fog networks for industrial IoT applications, in: 2020 IEEE International Conference on Communications Workshops (ICC Workshops), 2020, pp. 1–6, http://dx.doi.org/10.1109/ICCWorkshops49005.2020.9145351.

[96] X. Wang, W. Cheng, C. Ren, Multi-objective joint optimization of communication-computation-caching resources in mobile edge computing, in: 2021 IEEE/CIC International Conference on Communications in China (ICCC Workshops), 2021, pp. 94–99, http://dx.doi.org/10.1109/ICCCWorkshops52231.2021.9538887.

[97] L. Dong, H. Yao, C. Fang, T. Dong, S. Xu, Y. Liu, Edge cache-aided computation offloading for mobile cloud computing, in: 2021 IEEE/CIC International Conference on Communications in China, ICCC, 2021, pp. 1048–1053, http://dx.doi.org/10.1109/ICCC52777.2021.9580264.

[98] K. Kamran, E. Yeh, Q. Ma, DECO: Joint computation scheduling, caching, and communication in data-intensive computing networks, IEEE/ACM Trans. Netw. (2021) 1–15, http://dx.doi.org/10.1109/TNET.2021.3136157.

[99] C. Liang, Y. He, F.R. Yu, N. Zhao, Energy-efficient resource allocation in software-defined mobile networks with mobile edge computing and caching, in: 2017 IEEE Conference on Computer Communications Workshops (INFOCOM WKSHPS), 2017, pp. 121–126, http://dx.doi.org/10.1109/INFCOMW.2017.8116363.

[100] C. Liu, H. Zhang, H. Ji, X. Li, MEC-assisted flexible transcoding strategy for adaptive bitrate video streaming in small cell networks, China Commun. 18 (2) (2021) 200–214, http://dx.doi.org/10.23919/JCC.2021.02.013.

[101] C. Wang, D. Feng, S. Zhang, Q. Chen, Video caching and transcoding in wireless cellular networks with mobile edge computing: A robust approach, IEEE Trans. Veh. Technol. 69 (8) (2020) 9234–9238, http://dx.doi.org/10.1109/TVT.2020.2997344.

[102] Y.F. Yeznabad, M. Helfert, G.-M. Muntean, Backhaul traffic and QoE joint optimization approach for adaptive video streaming in MEC-enabled wireless networks, in: 2022 IEEE International Symposium on Broadband Multimedia Systems and Broadcasting, BMSB, 2022, pp. 1–6, http://dx.doi.org/10.1109/BMSB55706.2022.9828728.

[103] W. Liu, H. Zhang, H. Ding, D. Yuan, Delay and energy minimization for adaptive video streaming: A joint edge caching, computing and power allocation approach, IEEE Trans. Veh. Technol. 71 (9) (2022) 9602–9612, http://dx.doi.org/10.1109/TVT.2022.3179696.

[104] G. Zhou, L. Zhao, Y. Wang, G. Zheng, L. Hanzo, Energy efficiency and delay optimization for edge caching aided video streaming, IEEE Trans. Veh. Technol. 69 (11) (2020) 14116–14121, http://dx.doi.org/10.1109/TVT.2020.3029742.

[105] C. Liang, Y. He, F.R. Yu, N. Zhao, Enhancing video rate adaptation with mobile edge computing and caching in software-defined mobile networks, IEEE Trans. Wireless Commun. 17 (10) (2018) 7013–7026, http://dx.doi.org/10.1109/TWC.2018.2865354.

[106] X. Huang, L. He, L. Wang, F. Li, Towards 5G: Joint optimization of video segment caching, transcoding and resource allocation for adaptive video streaming in a multi-access edge computing network, IEEE Trans. Veh. Technol. 70 (10) (2021) 10909–10924, http://dx.doi.org/10.1109/TVT.2021.3108152.

[107] Y. Jin, Y. Wen, C. Westphal, Towards joint resource allocation and routing to optimize video distribution over future internet, in: 2015 IFIP Networking Conference (IFIP Networking), 2015, pp. 1–9, http://dx.doi.org/10.1109/IFIPNetworking.2015.7145311.

[108] H. Hu, Y. Jin, Y. Wen, C. Westphal, Orchestrating caching, transcoding and request routing for adaptive video streaming over ICN, ACM Trans. Multimedia Comput. Commun. Appl. 15 (1) (2019) http://dx.doi.org/10.1145/3289184.

[109] T. Dang, M. Peng, Y. Liu, C. Liu, Joint bandwidth, caching, and computing resource allocation for mobile VR delivery in F-RANs, in: 2019 IEEE Global Communications Conference, GLOBECOM, 2019, pp. 1–6, http://dx.doi.org/10.1109/GLOBECOM38437.2019.9013251.

[110] Z. Gu, H. Lu, C. Zou, Horizontal and vertical collaboration for VR delivery in MEC-enabled small-cell networks, IEEE Commun. Lett. 26 (3) (2022) 627–631, http://dx.doi.org/10.1109/LCOMM.2021.3140072.

[111] S. Li, P. Lin, J. Song, Q. Song, Computing-assisted task offloading and resource allocation for wireless VR systems, in: 2020 IEEE 6th International Conference on Computer and Communications, ICCC, 2020, pp. 368–372, http://dx.doi.org/10.1109/ICCC51575.2020.9345178.

[112] Z. Tan, F.R. Yu, X. Li, H. Ji, V.C.M. Leung, Virtual resource allocation for heterogeneous services in full duplex-enabled SCNs with mobile edge computing and caching, IEEE Trans. Veh. Technol. 67 (2) (2018) 1794–1808, http://dx.doi.org/10.1109/TVT.2017.2764002.

[113] Z. Tan, F.R. Yu, X. Li, H. Ji, V.C.M. Leung, Virtual resource allocation for heterogeneous services in full duplex-enabled small cell networks with cache and MEC, in: 2017 IEEE Conference on Computer Communications Workshops (INFOCOM WKSHPS), 2017, pp. 163–168, http://dx.doi.org/10.1109/INFCOMW.2017.8116370.

[114] C. Wang, C. Liang, F.R. Yu, Q. Chen, L. Tang, Joint computation offloading, resource allocation and content caching in cellular networks with mobile edge computing, in: 2017 IEEE International Conference on Communications, ICC, 2017, pp. 1–6, http://dx.doi.org/10.1109/ICC.2017.7996857.

[115] S. Kim, 5G network communication, caching, and computing algorithms based on the two-tier game model, ETRI J. 40 (1) (2018) 61–71, http://dx.doi.org/10.4218/etrij.2017-0023.

[116] S. Luo, X. Chen, Z. Zhou, S. Yu, Fog-enabled joint computation, communication and caching resource sharing for energy-efficient IoT data stream processing, IEEE Trans. Veh. Technol. 70 (4) (2021) 3715–3730, http://dx.doi.org/10.1109/TVT.2021.3062664.

[117] R.A. Cooke, S.A. Fahmy, A model for distributed in-network and near-edge computing with heterogeneous hardware, Future Gener. Comput. Syst. 105 (2020) 395–409.

[118] N. Hu, Z. Tian, X. Du, M. Guizani, An energy-efficient in-network computing paradigm for 6G, IEEE Trans. Green Commun. Netw. (2021) 1, http://dx.doi.org/10.1109/TGCN.2021.3099804.

[119] G. Wang, L. Wang, J. Chuan, W. Xie, H. Zhang, A. Fei, LRA-3C: Learning based resource allocation for communication-computing-caching systems, in: 2019 International Conference on Internet of Things (IThings) and IEEE Green Computing and Communications (GreenCom) and IEEE Cyber, Physical and Social Computing (CPSCom) and IEEE Smart Data (SmartData), 2019, pp. 828–833, http://dx.doi.org/10.1109/iThings/GreenCom/CPSCom/SmartData.2019.00150.

[120] Z. Ning, K. Zhang, X. Wang, L. Guo, X. Hu, J. Huang, B. Hu, R.Y.K. Kwok, Intelligent edge computing in internet of vehicles: A joint computation offloading and caching solution, IEEE Trans. Intell. Transp. Syst. 22 (4) (2021) 2212–2225, http://dx.doi.org/10.1109/TITS.2020.2997832.

[121] Z. Yang, Y. Liu, Y. Chen, N. Al-Dhahir, Cache-aided NOMA mobile edge computing: A reinforcement learning approach, IEEE Trans. Wireless Commun. 19 (10) (2020) 6899–6915, http://dx.doi.org/10.1109/TWC.2020.3006922.

[122] Z. Yang, Y. Fu, Y. Liu, Y. Chen, J. Zhang, A new look at AI-driven NOMA-F-RANs: Features extraction, cooperative caching, and cache-aided computing, IEEE Wirel. Commun. 29 (3) (2022) 123–130, http://dx.doi.org/10.1109/MWC.112.2100264.

[123] R. Wang, M. Li, L. Peng, Y. Hu, M.M. Hassan, A. Alelaiwi, Cognitive multi-agent empowering mobile edge computing for resource caching and collaboration, Future Gener. Comput. Syst. 102 (2020) 66–74, http://dx.doi.org/10.1016/j.future.2019.08.001.

[124] A. Ndikumana, N.H. Tran, D.H. Kim, K.T. Kim, C.S. Hong, Deep learning based caching for self-driving cars in multi-access edge computing, IEEE Trans. Intell. Transp. Syst. 22 (5) (2021) 2862–2877, http://dx.doi.org/10.1109/TITS.2020.2976572.

[125] J. Chen, H. Xing, X. Lin, A. Nallanathan, S. Bi, Joint resource allocation and cache placement for location-aware multi-user mobile-edge computing, IEEE Internet Things J. 9 (24) (2022) 25698–25714, http://dx.doi.org/10.1109/JIOT.2022.3196908.

[126] W. Chen, X. Qiu, T. Cai, H.-N. Dai, Z. Zheng, Y. Zhang, Deep reinforcement learning for internet of things: A comprehensive survey, IEEE Commun. Surv. Tutor. 23 (3) (2021) 1659–1692, http://dx.doi.org/10.1109/COMST.2021.3073036.

[127] Y. He, N. Zhao, H. Yin, Integrated networking, caching, and computing for connected vehicles: A deep reinforcement learning approach, IEEE Trans. Veh. Technol. 67 (1) (2018) 44–55, http://dx.doi.org/10.1109/TVT.2017.2760281.

[128] Y. He, F.R. Yu, N. Zhao, V.C.M. Leung, H. Yin, Software-defined networks with mobile edge computing and caching for smart cities: A big data deep reinforcement learning approach, IEEE Commun. Mag. 55 (12) (2017) 31–37, http://dx.doi.org/10.1109/MCOM.2017.1700246.

[129] T. Liu, L. Tang, W. Wang, X. He, Q. Chen, X. Zeng, H. Jiang, Resource allocation in DT-assisted internet of vehicles via edge intelligent cooperation, IEEE Internet Things J. 9 (18) (2022) 17608–17626, http://dx.doi.org/10.1109/JIOT.2022.3156100.

[130] L.U. Khan, Z. Han, W. Saad, E. Hossain, M. Guizani, C.S. Hong, Digital twin of wireless systems: Overview, taxonomy, challenges, and opportunities, IEEE Commun. Surv. Tutor. 24 (4) (2022) 2230–2254, http://dx.doi.org/10.1109/COMST.2022.3198273.

[131] Z. Qin, S. Leng, J. Zhou, S. Mao, Collaborative edge computing and caching in vehicular networks, in: 2020 IEEE Wireless Communications and Networking Conference, WCNC, 2020, pp. 1–6, http://dx.doi.org/10.1109/WCNC45663.2020.9120600.

[132] K. Zhang, J. Cao, H. Liu, S. Maharjan, Y. Zhang, Deep reinforcement learning for social-aware edge computing and caching in urban informatics, IEEE Trans. Ind. Inform. 16 (8) (2020) 5467–5477, http://dx.doi.org/10.1109/TII.2019.2953189.

[133] L.T. Tan, R.Q. Hu, Mobility-aware edge caching and computing in vehicle networks: A deep reinforcement learning, IEEE Trans. Veh. Technol. 67 (11) (2018) 10190–10203, http://dx.doi.org/10.1109/TVT.2018.2867191.

[134] H. Peng, X.S. Shen, DDPG-based resource management for MEC/UAV-Assisted vehicular networks, in: 2020 IEEE 92nd Vehicular Technology Conference (VTC2020-Fall), 2020, pp. 1–6, http://dx.doi.org/10.1109/VTC2020-Fall49728.2020.9348633.

[135] D. Lan, A. Taherkordi, F. Eliassen, L. Liu, Deep reinforcement learning for computation offloading and caching in fog-based vehicular networks, in: 2020 IEEE 17th International Conference on Mobile Ad Hoc and Sensor Systems, MASS, 2020, pp. 622–630, http://dx.doi.org/10.1109/MASS50613.2020.00081.

[136] F. Xu, F. Yang, S. Bao, C. Zhao, DQN inspired joint computing and caching resource allocation approach for software defined information-centric internet of things network, IEEE Access 7 (2019) 61987–61996, http://dx.doi.org/10.1109/ACCESS.2019.2916178.

[137] J. Ren, T. Hou, H. Wang, H. Tian, H. Wei, H. Zheng, X. Zhang, Collaborative task offloading and resource scheduling framework for heterogeneous edge computing, Wirel. Netw. (2021) http://dx.doi.org/10.1007/s11276-021-02768-y.

[138] Y. He, C. Liang, F.R. Yu, V.C. Leung, Integrated computing, caching, and communication for trust-based social networks: A big data DRL approach, in: 2018 IEEE Global Communications Conference, GLOBECOM, 2018, pp. 1–6, http://dx.doi.org/10.1109/GLOCOM.2018.8647548.

[139] Y. Sun, M. Peng, S. Mao, Deep reinforcement learning-based mode selection and resource management for green fog radio access networks, IEEE Internet Things J. 6 (2) (2019) 1960–1971, http://dx.doi.org/10.1109/JIOT.2018.2871020.

[140] C. Fang, T. Zhang, J. Huang, H. Xu, Z. Hu, Y. Yang, Z. Wang, Z. Zhou, X. Luo, A DRL-driven intelligent optimization strategy for resource allocation in cloud-edge-end cooperation environments, Symmetry 14 (10) (2022) http://dx.doi.org/10.3390/sym14102120.

[141] W. Chen, Q. Song, P. Lin, L. Guo, A. Jamalipour, Proactive 3C resource allocation for wireless virtual reality using deep reinforcement learning, in: 2021 IEEE Global Communications Conference, GLOBECOM, 2021, pp. 1–6, http://dx.doi.org/10.1109/GLOBECOM46510.2021.9685438.

[142] Q. Chen, Z. Kuang, L. Zhao, Multiuser computation offloading and resource allocation for cloud–edge heterogeneous network, IEEE Internet Things J. 9 (5) (2022) 3799–3811, http://dx.doi.org/10.1109/JIOT.2021.3100117.

[143] A. Ndikumana, S. Ullah, T. LeAnh, N.H. Tran, C.S. Hong, Collaborative cache allocation and computation offloading in mobile edge computing, in: 2017 19th Asia-Pacific Network Operations and Management Symposium, APNOMS, 2017, pp. 366–369, http://dx.doi.org/10.1109/APNOMS.2017.8094149.

[144] D. Ren, X. Gui, K. Zhang, Adaptive request scheduling and service caching for MEC-assisted IoT networks: An online learning approach, IEEE Internet Things J. 9 (18) (2022) 17372–17386, http://dx.doi.org/10.1109/JIOT.2022.3157677.

[145] S. Chen, L. Rui, Z. Gao, W. Li, X. Qiu, Cache-assisted collaborative task offloading and resource allocation strategy: A metareinforcement learning approach, IEEE Internet Things J. 9 (20) (2022) 19823–19842, http://dx.doi.org/10.1109/JIOT.2022.3168885.

[146] S. Yu, R. Langar, X. Fu, L. Wang, Z. Han, Computation offloading with data caching enhancement for mobile edge computing, IEEE Trans. Veh. Technol. 67 (11) (2018) 11098–11112, http://dx.doi.org/10.1109/TVT.2018.2869144.

[147] A.H. Khan, N. Ul Hassan, C. Yuen, J. Zhao, D. Niyato, Y. Zhang, H.V. Poor, Blockchain and 6G: The future of secure and ubiquitous communication, IEEE Wirel. Commun. 29 (1) (2022) 194–201, http://dx.doi.org/10.1109/MWC.001.2100255.

[148] X. Ye, M. Li, P. Si, R. Yang, Z. Wang, Y. Zhang, Collaborative and intelligent resource optimization for computing and caching in IoV with blockchain and MEC using A3C approach, IEEE Trans. Veh. Technol. 72 (2) (2023) 1449–1463, http://dx.doi.org/10.1109/TVT.2022.3210570.

[149] Y. Zhou, X. Li, H. Ji, H. Zhang, Blockchain-based trustworthy service caching and task offloading for intelligent edge computing, in: 2021 IEEE Global Communications Conference, GLOBECOM, 2021, pp. 1–6, http://dx.doi.org/10.1109/GLOBECOM46510.2021.9685168.

[150] Y. Sun, S. Chen, Z. Wang, S. Mao, A joint learning and game-theoretic approach to multi-dimensional resource management in fog radio access networks, IEEE Trans. Veh. Technol. 72 (2) (2023) 2550–2563, http://dx.doi.org/10.1109/TVT.2022.3214075.

[151] Q. Li, D. Wang, H. Lu, A cooperative caching and computing-offloading method for 3C trade-off in VR video services, IEEE Access 9 (2021) 124010–124022, http://dx.doi.org/10.1109/ACCESS.2021.3110741.

[152] T. Yang, Z. Tan, Y. Xu, S. Cai, Collaborative edge caching and transcoding for 360° video streaming based on deep reinforcement learning, IEEE Internet Things J. 9 (24) (2022) 25551–25564, http://dx.doi.org/10.1109/JIOT.2022.3197798.

[153] Z. Liu, N. Garg, T. Ratnarajah, Multi-agent federated reinforcement learning strategy for mobile virtual reality delivery networks, IEEE Trans. Netw. Sci. Eng. 11 (1) (2024) 100–114, http://dx.doi.org/10.1109/TNSE.2023.3292570.

[154] T.X. Tran, P. Pandey, A. Hajisami, D. Pompili, Collaborative multi-bitrate video caching and processing in mobile-edge computing networks, in: 2017 13th Annual Conference on Wireless on-demand Network Systems and Services (WONS), 2017, pp. 165–172.

[155] Y. Zhang, B. Feng, W. Quan, A. Tian, K. Sood, Y. Lin, H. Zhang, Cooperative edge caching: A multi-agent deep learning based approach, IEEE Access 8 (2020) 133212–133224, http://dx.doi.org/10.1109/ACCESS.2020.3010329.

[156] T. Zhang, Z. Wang, Y. Liu, W. Xu, A. Nallanathan, Joint resource, deployment, and caching optimization for ar applications in dynamic UAV NOMA networks, IEEE Trans. Wireless Commun. 21 (5) (2022) 3409–3422, http://dx.doi.org/10.1109/TWC.2021.3121584.

[157] L. Velasco, M. Signorelli, O.G. De Dios, C. Papagianni, R. Bifulco, J.J.V. Olmos, S. Pryor, G. Carrozzo, J. Schulz-Zander, M. Bennis, R. Martinez, F. Cugini, C. Salvadori, V. Lefebvre, L. Valcarenghi, M. Ruiz, End-to-end intent-based networking, IEEE Commun. Mag. 59 (10) (2021) 106–112, http://dx.doi.org/10.1109/MCOM.101.2100141.

[158] L. Tan, W. Su, W. Zhang, J. Lv, Z. Zhang, J. Miao, X. Liu, N. Li, In-band network telemetry: A survey, Comput. Netw. 186 (2021) 107763.

[159] P. Manzanares-Lopez, J.P. Muñoz-Gea, J. Malgosa-Sanahuja, Passive in-band network telemetry systems: The potential of programmable data plane on network-wide telemetry, IEEE Access 9 (2021) 20391–20409.

[160] R. Ben Basat, S. Ramanathan, Y. Li, G. Antichi, M. Yu, M. Mitzenmacher, PINT: Probabilistic in-band network telemetry, in: Proceedings of the Annual Conference of the ACM Special Interest Group on Data Communication on the Applications, Technologies, Architectures, and Protocols for Computer Communication, 2020, pp. 662–680.

[161] Q. Huang, H. Sun, P.P. Lee, W. Bai, F. Zhu, Y. Bao, Omnimon: Re-architecting network telemetry with resource efficiency and full accuracy, in: Proceedings of the Annual Conference of the ACM Special Interest Group on Data Communication on the Applications, Technologies, Architectures, and Protocols for Computer Communication, 2020, pp. 404–421.

[162] E.K. Markakis, K. Karras, A. Sideris, G. Alexiou, E. Pallis, Computing, caching, and communication at the edge: The cornerstone for building a versatile 5G ecosystem, IEEE Commun. Mag. 55 (11) (2017) 152–157.

[163] T. Taleb, A. Boudi, L. Rosa, L. Cordeiro, T. Theodoropoulos, K. Tserpes, P. Dazzi, A. Protopsaltis, R. Li, Towards supporting XR services: Architecture and enablers, IEEE Internet Things J. (2022) 1, http://dx.doi.org/10.1109/JIOT.2022.3222103.

[164] T.-V. Nguyen, N.-N. Dao, V. Dat Tuong, W. Noh, S. Cho, User-aware and flexible proactive caching using LSTM and ensemble learning in IoT-MEC networks, IEEE Internet Things J. 9 (5) (2022) 3251–3269, http://dx.doi.org/10.1109/JIOT.2021.3097768.

[165] M. Manalastas, H. Farooq, S.M. Asad Zaidi, A. Imran, Where to go next?: A realistic evaluation of AI-assisted mobility predictors for HetNets, in: 2020 IEEE 17th Annual Consumer Communications & Networking Conference, CCNC, 2020, pp. 1–6, http://dx.doi.org/10.1109/CCNC46108.2020.9045127.

[166] O. Gupta, R. Raskar, Distributed learning of deep neural network over multiple agents, 2018, http://dx.doi.org/10.48550/ARXIV.1810.06060.

[167] T. Li, A.K. Sahu, A. Talwalkar, V. Smith, Federated learning: Challenges, methods, and future directions, IEEE Signal Process. Mag. 37 (3) (2020) 50–60.

[168] Network automation, White Paper, 2019, URL http://bit.ly/2RGBDbc.

[169] I. Kunze, R. Glebke, J. Scheiper, M. Bodenbenner, R.H. Schmitt, K. Wehrle, Investigating the applicability of in-network computing to industrial scenarios, in: 2021 4th IEEE International Conference on Industrial Cyber-Physical Systems, ICPS, 2021, pp. 334–340, http://dx.doi.org/10.1109/ICPS49255.2021.9468247.

[170] Z. Ning, K. Zhang, X. Wang, M.S. Obaidat, L. Guo, X. Hu, B. Hu, Y. Guo, B. Sadoun, R.Y. Kwok, Joint computing and caching in 5G-envisioned internet of vehicles: A deep reinforcement learning-based traffic control system, IEEE Trans. Intell. Transp. Syst. 22 (8) (2020) 5201–5212.

[171] G. Kakkavas, A. Stamou, V. Karyotis, S. Papavassiliou, Network tomography for efficient monitoring in SDN-enabled 5G networks and beyond: Challenges and opportunities, IEEE Commun. Mag. 59 (3) (2021) 70–76.

[172] J. Hyun, J.W.-K. Hong, Knowledge-defined networking using in-band network telemetry, in: 2017 19th Asia-Pacific Network Operations and Management Symposium, APNOMS, IEEE, 2017, pp. 54–57.

[173] A. Kretsis, P. Kokkinos, P. Soumplis, J.J.V. Olmos, M. Fehér, M. Sipos, D.E. Lucani, D. Khabi, D. Masouros, K. Siozios, P. Bourgos, S. Tsekeridou, F. Zyulkyarov, E. Karanastasis, E. Chondrogiannis, V. Andronikou, A.F. Gomez, S. Panica, G. Iuhasz, A. Nanos, C. Chalios, M. Varvarigos, SERRANO: Transparent application deployment in a secure, accelerated and cognitive cloud continuum, in: 2021 IEEE International Mediterranean Conference on Communications and Networking (MeditCom), 2021, pp. 55–60, http://dx.doi.org/10.1109/MeditCom49071.2021.9647689.

[174] N. Kato, B. Mao, F. Tang, Y. Kawamoto, J. Liu, Ten challenges in advancing machine learning technologies toward 6G, IEEE Wirel. Commun. 27 (3) (2020) 96–103, http://dx.doi.org/10.1109/MWC.001.1900476.

[175] S. Abolfazli, Z. Sanaei, E. Ahmed, A. Gani, R. Buyya, Cloud-based augmentation for mobile devices: Motivation, taxonomies, and open challenges, IEEE Commun. Surv. Tutor. 16 (1) (2014) 337–368, http://dx.doi.org/10.1109/SURV.2013.070813.00285.

[176] I. Stoica, S. Shenker, From cloud computing to sky computing, in: Proceedings of the Workshop on Hot Topics in Operating Systems, HotOS '21, Association for Computing Machinery, 2021, pp. 26–32.

[177] S. Rene, O. Ascigil, I. Psaras, G. Pavlou, A congestion control framework based on in-network resource pooling, IEEE/ACM Trans. Netw. (2021) 1–15, http://dx.doi.org/10.1109/TNET.2021.3128384.

[178] R. Ureña, G. Kou, Y. Dong, F. Chiclana, E. Herrera-Viedma, A review on trust propagation and opinion dynamics in social networks and group decision making frameworks, Inform. Sci. 478 (2019) 461–475.

[179] W. Zheng, Z. Zheng, X. Chen, K. Dai, P. Li, R. Chen, NutBaaS: A blockchain-as-a-service platform, IEEE Access 7 (2019) 134422–134433.

[180] K.B. Letaief, W. Chen, Y. Shi, J. Zhang, Y.-J.A. Zhang, The roadmap to 6G: AI empowered wireless networks, IEEE Commun. Mag. 57 (8) (2019) 84–90, http://dx.doi.org/10.1109/MCOM.2019.1900271.

[181] C.E. Shannon, W. Weaver, The Mathematical Theory of Communication, University of Illinois Press, 1949.

[182] C. Chaccour, W. Saad, M. Debbah, Z. Han, H.V. Poor, Less data, more knowledge: Building next generation semantic communication networks, 2022, http://dx.doi.org/10.48550/ARXIV.2211.14343.

[183] E. Calvanese Strinati, S. Barbarossa, 6G networks: Beyond Shannon towards semantic and goal-oriented communications, Comput. Netw. 190 (2021) 107930, http://dx.doi.org/10.1016/j.comnet.2021.107930.

[184] M. Chein, M.-L. Mugnier, Graph-Based Knowledge Representation: Computational Foundations of Conceptual Graphs, in: Advanced Information and Knowledge Processing, Springer, 2008, p. 427, http://dx.doi.org/10.1007/978-1-84800-286-9.

[185] G. Shi, Y. Xiao, Y. Li, X. Xie, From semantic communication to semantic-aware networking: Model, architecture, and open problems, IEEE Commun. Mag. 59 (8) (2021) 44–50, http://dx.doi.org/10.1109/MCOM.001.2001239.

[186] H. Seo, J. Park, M. Bennis, M. Debbah, Semantics-native communication with contextual reasoning, 2021, arXiv:2108.05681.