Unidirectional Communication Model for Hyper-Low Latency in 6G Networks

Haydar Çukurtepe
Dept. of Computer and Information Sciences
Valparaiso University
Valparaiso, IN USA
0000-0002-4670-4877

***Abstract*—Next-generation networks are expected to connect an excessive number of devices and provide ubiquitous intelligence to user equipment. To achieve this goal, 6G networks aim numerous capabilities: 3D global coverage, ultra-dense connectivity, high-precision positioning, ultra-reliable and secure connectivity, high energy efficiency, hyper-low latency, and more. These key capabilities require specialized communication models to meet the expected network standards. In this study, a unidirectional communication model is proposed for hyper-low latency to provide ubiquitous intelligence to the user equipment. The proposed communication model contributes to various key capabilities of 6G networks, such as hyper-low latency, ubiquitous intelligence, ultra-dense connectivity, high mobility, sustainability, and reliability.**

***Index Terms*—6G Networks, Hyper-Low Latency, Edge Computing, Ubiquitous Intelligence, and Joint Sensing.**

I. INTRODUCTION

Wireless communication reached its peak with significant improvements brought by 5G. The world is more interconnected ever before, but 5G is still far from meeting the low latency requirements of ubiquitous intelligence. Key capabilities defined by ITU-R in [1] still apply for the next-generation networks, but 6G will both extend the existing ones and aim more capabilities with higher performance standards such as: 3D global coverage, sensing-related capabilities, AI-related capabilities, sustainability, interoperability, high-precision positioning, ultra-dense connectivity, higher spectrum efficiency, ultra-reliable and secure connectivity, high energy efficiency, hyper-low latency and ubiquitous intelligence [2].

Ultra-dense connectivity refers to the massive number of devices being connected, including various types of devices, such as: smart meters, smart devices, vehicles, home appliances, industrial equipment, wearables, sensors, readers, actuators, and etc. Connectivity also refers to necessary software and communication models that enable connected-devices exchanging data and providing joint sensing and intelligence over the network. Together with the extensive computing power and advanced artificial intelligence/machine learning (AI/ML) models, 6G networks will not only provide communication services but also other services to combine communications, computing, control, sensing, localization, storage and ubiquitous wireless intelligence.

Running large-scale AI/ML models on a single small device is difficult due to the restricted computing and storage capabilities of smart devices. However, thanks to advanced AI/ML models, small devices can still contribute to collectively produced intelligence. The collective capability of intensive computation and distributed storage relies heavily on underlying networks. Current networking models are not designed to support hyper-low latency requirements for pervasive intelligence. 6G networks require specialized design and communication models to achieve the expected key performance capabilities. Thus, a convenient network architecture and communication model that enables small devices to contribute to AI/ML models to meet future expectations is the key to 6G networks. This study focuses on the networking requirements of ubiquitous intelligence of 6G networks. A generic unidirectional communication model is proposed, allowing various 6G network elements to collaborate in order to transmit intelligence to user equipment (UE). The proposed communication model is applicable and it contributes to various key capabilities of 6G networks, such as hyper-low latency, ubiquitous intelligence, ultra-dense connectivity, high mobility, reliability, and sustainability.

This study is organized as follows: The global vision for 6G network architecture is explained in Section II together with expected features and capabilities. Expected ubiquitous intelligence and enabling features of 6G is discussed further in Section III. The proposed communication model, data transfer models, and detailed tasks of different network components are given at Section IV. The expected improvements and contributions of proposed model are discussed in Section V. Section VI concludes the paper.

II. 6G NETWORK ARCHITECTURE GLOBAL VISION

The first vision document about the 6G was published in 2019 [3]. At this early stage of 6G network development, organizations, institutions, and researchers articulate their visions [4]–[12]. The articulated visions has broad similarity about ”ubiquitous wireless intelligence” being key to enabling enabler of 6G network networks, and ”ultra-dense connectivity with hyper-low latency” being the key driver.

The authors propose an infrastructure-free service architecture that extends the 5G service-based architecture (SBA) in [13]. The proposed architecture aims to realize full software-based service architecture. In [14], the authors propose an architecture composed of 3 layers and 4 planes. The layers are

979-8-3503-3094-6/24/$31.00

the resource layer, network function layer, service and application layer serving at data collection, security, AI, and cooperation planes . Huawei foresees 6G intelligence-of-everything where everything will be sensed, connected, and intelligent [12]. In [15], hierarchical and distributed intelligence, in-depth convergence of connection and intelligence, and collaborative exposure of connection and intelligence capabilities in 6G are reviewed. The IMT-2030 promotion group expressed their 6G vision in [10] as an intelligent connection of everything and digital twins, where the digital world reflects the physical world in real-time. Joint communications, sensing, and computation enabled 6G intelligent networks concepts were proposed in [16]. Connected AI was proposed in [17]. In [18], a joint cloud edge computing model was proposed to improve the computing performance. Sideloading and mesh networking protocols are adapted to create an automated device-to-device network in [19] that would work as an auxiliary. A comprehensive survey of 6G studies and foreseen architectures is presented in [20].

ITU-R published a perspective framework in July 2023 [2], which was a draft guideline for all initiatives for what to expect, and what to develop. This framework document incorporates 6G visions for realistic goals. ITU-R framework document defines 6 usage scenarios: Immersive communication, massive communication, Hyper Reliable and Low-Latency Communication, Ubiquitous Connectivity, Integrated AI and Communication, and Integrated Sensing and Communication. It also defined 4 overarching aspects that will be commonly applicable to all usage scenarios. Overarching aspects are Sustainability, Connecting the unconnected, Ubiquitous Intelligence, Security/Privacy/Resilience. The standards of key capabilities defined by ITU-R [2] are shown in Figure-1 and further information about the defined capabilities are given below:

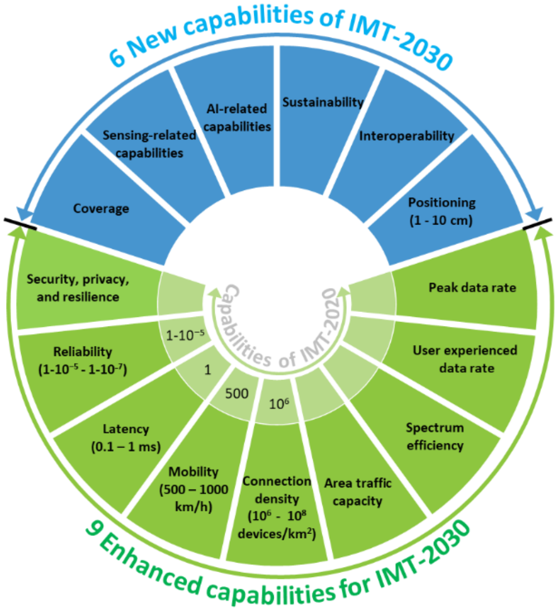

Fig. 1. 6G Capabilities [2].

- Ubiquitous coverage in 3D: Terrestrial mobile communications will be expanded to cover all the users (e.g., including space, air, ocean, deserts, even underground). The total coverage percentage of 6G network is expected to increase from 70% to 99% [21].
- High data rates (e.g. Tbps level peak data rates, and 10-100Gbps user experience rates using orbital angular momentum, super-large scale antenna, visible light, new coding and modulation, AI [4], and millimeter wave (mmW) spectrum [3], [5].)
- Hyper-low latency (e.g. Near real time latency, 0.1 ms or even zero delay [3])
- Ultra-High Reliability (e.g. the services and the network will always be reachable with 99.999999% [8]).
- Multi-domain collaboration, mobility in high speed (e.g. the network will support user equipment in motion up to 1000 km/h [4], [21]).
- High precision positioning (e.g., 1m outdoor, 1cm [8] indoor accuracy).
- Massive connection density (e.g. 10s of millions of UE connections per km^2 will be connected to the network [4], [8]).
- Energy and Spectrum Efficiency (e.g. IoT devices would require extremely low energy to keep alive for longer lifetimes. 6G is expected to provide 10–100 times energy efficiency and 5–10 times spectrum efficiency compared to 5G [5]).
- Higher network security (e.g. Invulnerable, resilience, data governance, and secure circulation considerations).
- AI-Related capabilities (e.g. AI applications will dominate the network traffic [5]).
- Interoperability (e.g. plug and play service based architecture [13] of 5G can be tailored and continue to be used in control plane. User plane requires more generic approach that can achieve hyper-low latency.)

These capabilities cannot be achieved simply by adapting technological advances to the existing 5G network architecture. A network architecture and a compatible communication model are the key to meet future expectations.

This study focuses on the practical application of intelligence flow to meet high standards of 6G. The proposed communication model is based on multi-layer computation and bringing intelligence before it is requested. In the proposed model, cloud computing uses high-capacity resources in the cloud with appropriate application and service provision. Edge computing, on the other hand, uses resources close to the UE for time-critical applications to adjust intelligence just before it is served. Although this general structure has been considered in different studies [4], [5], [9]–[12] for 6G networks, none of them proposed a practical implementation of to meet the emerging standards.

III. ENABLERS AND CHALLENGES OF HYPER-LOW LATENCY

A. Ubiquitous Intelligence

Artificial intelligence/machine learning (AI/ML) algorithms are one of the main drivers and enablers of next-generation networks. AI/ML applications continue to improve at an unprecedented pace. AI/ML models grow to include more data for better precision. Large-scale AI/ML models present new opportunities that consider more parameters and using massive data, thus, achieve superior performance or handle more complex tasks. For example, Switch Transformer NLP model reaches 1.6 trillion parameters, where GPT-3 has ≈ 175 billion parameters trained on nearly 45 TB of text data [22].

Recent advances in AI/ML models promise not only the strengthening of production and reasoning capabilities, but also the ubiquity of intelligence across the network (e.g. zero- and few-step prediction success) [23], [24].

Large-scale AI/ML implementations are extremely compute intensive. The process includes proper model selection, training the model with proper data, evaluate the process, and update when necessary. 3GPP defines the AI/ML functional framework for the NR air interface in [25]. The life-cycle management of the AI/ML is characterized by following aspects: data collection, model training, identification, transfer, inference, selection, monitoring, and update. Figure-2 shows edge nodes' contribution in various aspects. The framework lays the foundation for future air-interface use cases leveraging AI/ML techniques, where the final step of the life-cycle is to provide UEs with AI-related capability. The massive data used in AI/ML models will not only be the stored data but it will also be the collected data through millions of connected devices. 6G networks will be senses for large-scale AI/ML models.

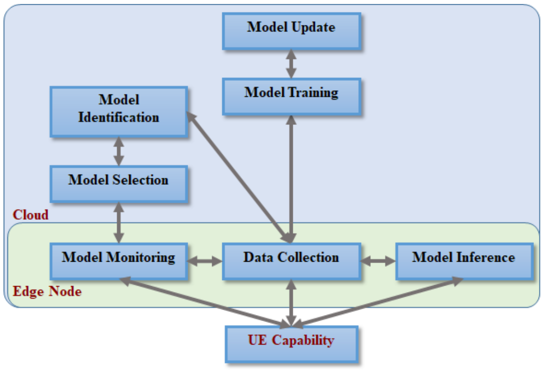

Fig. 2. AI/ML Life-cycle Framework

B. Network Challenges

The latency requirements of 6G radio are very strict, i.e. less than 0.1 ms. The latency-sensitive applications (e.g. instant security, tactile messaging, and etc.) do not tolerate delays on the network. The large-scale AI/ML models running on the cloud, joint network sensing, multi-dimensions of the problems, and the demand for a higher amount of data makes the latency requirements more challenging.

Request/response based service model suffers from delays such as carrying message to the proper server on the cloud, processing of request, running the AI/ML model on demand, and carrying the reply messages every time. If the network resources are not allocated from source to destination then the latency will also be affected from traffic load at the time. Thus, widespread deployment of edge nodes cannot address the emerging requirements alone.

The strict latency requirement do not tolerate collisions on accessing the network as well. Medium access control, which is a difficult problem in wired networks, is even more challenging in wireless networks. Collision detection is difficult in wireless networks. Signals attenuate more in wireless transmission. Two nodes having packets to send to the third node within their range, may not be aware of each other (Hidden node problem). Another problem with signal detection is that most radio devices are half duplex. They can only send or receive at a time. Millions of devices deployed in km^2 area equipped with wireless communication capability require more careful and detailed scheduling. Network access should be carefully planned and managed by a coordinator i.e. edge node.

IV. UNIDIRECTIONAL COMMUNICATION MODEL FOR HYPER-LOW LATENCY

The goal of proposed model is to transmit the intelligence to the user equipment with hyper-low latency. And, the intelligence is created by a massive number of deployed devices, and a necessary number of edge nodes. The envisioned communication between devices, edge node and the UEs is in the form of continuous service providing and consuming the services, instead of a request/response-based service model. The continuously offered services transmit the intelligence at the scheduled time across the network, which is carried to the last mile of the network before it is requested, updated by with the latest sensed data.

The components of underlying network services is shown in Figure-3. This is an updated version of current SBA [26]. Access Mobility Management Functionality (AMF) will also track the location information of UEs for registered services. AMF will coordinate AI/ML logic with Application Function(AF). In this communication model, the intelligence is either at the UEs or at the edge nodes, which are powered by a continuous energy sources. Responsibilities of these devices are explained in the following sections. Data transfer models marked as numbers in Figure-3 are explained in the following section.

A. Data Transfer Model

This study focuses on transfer of ubiquitous intelligence to UEs. Except for the edge node-cloud server communication, all the data transfer is shown for radio network, but also applies for wired connections. The radio network data flow will be

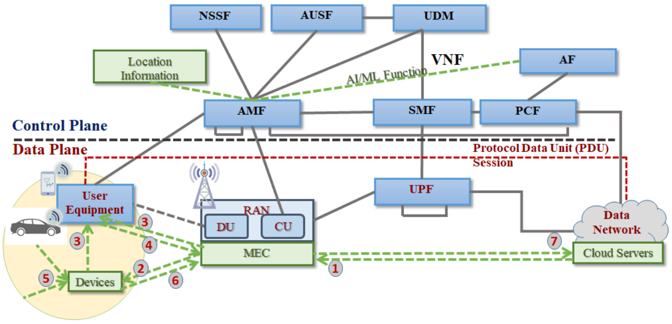

Fig. 3. Reference Network Architecture

unidirectional and will not require acknowledgment, except for the network join requests during the contention access period.

The data transfer models are as follows:

- *1.Data transfer from cloud servers*: Data prepared using large scale AI/ML algorithms are sent to the edge nodes, and the UEs through core network. Data coming from servers is formatted data for registered and unregistered services. The data includes: timestamp, service type, provider information, and etc.
- *2. Data transfer to device*: Edge node shares the data with the devices to be updated and transmitted. Different types of services require associate data and devices. The content of the data changes according to the services.
- *3. Data transfer to UE*: The devices are the leaves of this structure to transmit the data to UEs. Edge node also transmits certain types of data. Data for unregistered services is plain and accessible by any UE. The registered services' data on the other hand is encrypted and only the registered UEs are able to read it. Homomorphic encryption schemes allow small devices to contribute to data without decrypting it [27]. Registered services' data are also served at the location of registered UEs. Again, the content of the data changes according to the services.
- *4. Data transfer from UE*: UEs introduce themselves to edge node to receive personalized service.
- *5. Sensing*: Devices continuously collect information from their environment.
- *6. Data transfer to Edge*: Edge nodes are aware of the devices, their capabilities, UEs in their proximity, and their registered services.
- *7. Data transfer to cloud*: Edge node sends the collected data to the cloud servers through core network.

Detailed scheduling steps of data transfer models and responsibilities of radio network components are given in the following section.

B. Edge Nodes

Edge node is the key element of 6G network architecture with numerous tasks. It can be implemented either as a software-only entity that runs on top of a virtualised infrastructure [28], or as a hardware component at the edge of the network. The edge node communicates with each network entity to achieve the following tasks, as shown in Figure-4.

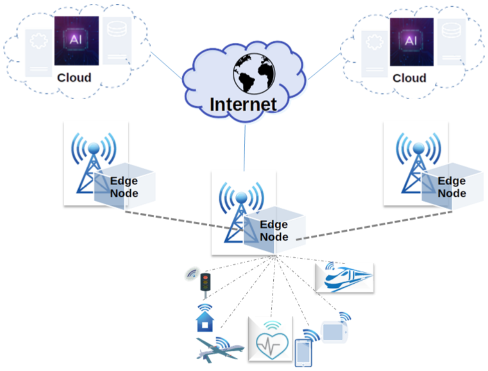

Fig. 4. Edge Node Connections.

1) Communication with Cloud Servers: Edge nodes get the output of the AI/ML algorithms from cloud servers. Data reaches multiple edge node in the area.

2) Running AI/ML Algorithms: Thanks to modern AI/ML algorithms and frameworks, the data prepared at cloud servers can be updated at the edge of the network. Edge nodes are not as capable as cloud servers but they can run the large-scale AI models partially, or re-run the algorithms with limited data entry.

3) Keep Registry of Devices and Services in the Proximity: Nodes maintain a dynamic registry to keep information about the small devices and the UEs in their vicinity.

4) Collect data: Edge nodes collect data from small devices and UEs. Collected data at the last mile of the network is sent to cloud servers. It includes the location of the UEs for required registered services.

5) Data Transfer to Devices: Edge nodes send data to the devices to be transmitted further, with an update if possible.

6) Scheduling and access management: Edge nodes are coordinators of the last mile of the network. They are responsible for scheduling and access management. This task is the key of the proposed communication model, where edge node dynamically manages devices' access to network. Scheduling is done using cells which are elemental units of network resources. The cells in this model are similar to standardized medium access protocol for Industrial IoT [29]. Time is discrete and divided into timeslots. A cell is defined by the spectrum and the timeslot offset pair. Each cell is large enough for a device to transmit data or receive data from edge node. ACK is not necessary, because the data flow is continuous and unidirectional. Small devices will be assigned a timeslot and a frequency by the edge nodes. All components are synchronized by the edge node on a slotframe. Cells are allocated in a schedule. Edge nodes run dynamic scheduling algorithms to decide when and how many cells are allocated for each device in the proximity. Unidirectional intelligence flow is orchestrated through dynamically maintained schedules. Small device know if they have the right to transmit data or not in a specific cell, and they operate in the allocated cell. Depending on the capabilities of connected small devices and

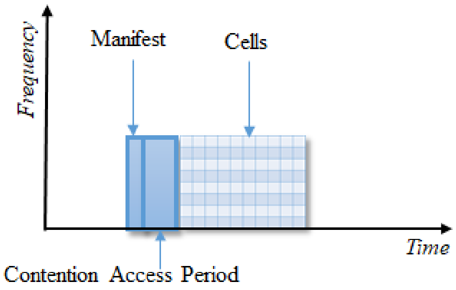

Fig. 5. Superframe structure.

the implementation of applications, connected devices may also receive data from their neighbors, besides sensing.

Different types of devices at the edge will operate for different purposes. They will need to be assigned a different number of cells for transmission. According to the format of the data being transferred the cell size can be adjusted by the coordinator edge node.

The edge node will prepare a manifest file based on: the number of connected devices, the capabilities of nodes, the services offered to the connected user equipment in the proximity, and link statistics. The manifest file will be updated at the beginning of each slotframe. Each device and UE in the proximity will receive this manifest, adjust their clock cycle with slotframe, learn about the contention access period (CAP), and learn about the details of cell allocation. CAP is the period where the UEs and devices will reach the edge node to register themselves. Association and neighbor discovery are other challenging topics in wireless communication. Thanks to the manifest file, devices will learn about their neighbors and offered services.

C. User Equipment (UE)

User equipment are capable devices (e.g. smartphones, smart cars, fast track trains, drones, etc.) with high computing power, necessary storage, and continuous energy supply.

UE will have more than one interface to communicate with the network. One interface will be for regular interaction over the Internet (e.g. live video streaming, and etc.) and the other interface is to receive intelligence for a specific service through a radio network.

There will be 2 types of services: registered and unregistered services. The unregistered services (e.g. measuring tire condition for public safety in coordination with weather services) will be provided to any UE in the proximity. On the other hand, registered services will provide intelligence locally only to the registered UEs. The intelligence will be prepared in a chain starting from cloud servers, edge nodes, and small devices. Registered services will only be available when the UE joins the network. When a UE joins the network its location information will be sent to the service provider through the edge node. Registered service intelligence will be prepared by the cloud server and delivered to the edge nodes that are close to UE or will be close soon enough. The data will be adjusted by the edge nodes and then shared with small devices and UE. The small devices will either transmit the data without a change or update it first using the sensed and calculated data.

D. Devices

A device is a small hardware (i.e. sensors, actuators, cameras, vehicles, and etc.) that is connected to the network, and has a radio interface to receive and transmit data. The devices are equipped with sensors or readers to sense their environment, and limited capability to update the data.

When they are deployed to the network, they select an edge node as coordinator and introduce themselves (e.g. reading capabilities, message format and length, spectrum usage, lifetime, etc.). They synchronize their clock with the coordinator, receive the latest data for the service they are deployed for, start sensing and reading data, and make themselves ready to transmit data at the assigned cell.

Devices have two types of input data: received from edge nodes and collected by themselves. They collect data continuously or at discrete times according to their capability and internal design. Devices are able to repeat the data received from edge nodes or they are also able to update the received data using sensors and calculations before transmitting. Since the data flow is unidirectional, devices do not require to get ACK for transmissions. They keep the recent data for the next cycle. In the next cycle, they will either use the current data, or the data received from the edge node, or updated data with the collected data. The devices then put themselves into sleep mode to save energy.

V. EXPECTED OUTCOMES

UEs will interact with their environment by consuming the data provided by the devices around them. Edge devices will provide intelligence to their proximity, and the UEs will consume the provided service.

The proposed communication model contributes directly or indirectly to various 6G capabilities defined in [2] as given in Table-I.

A request/response model will suffer from multiple delays on the network. Thus, hyper-low latency requirements of 6G can only be achieved with a communication model that brings the data before it is requested.

Ubiquitous intelligence can be achieved with collective work of centralized high computing power and final touches of distributed sensing. The proposed communication model lays the foundation for enabling ubiquitous intelligence.

6G will facilitate multi-service (i.e. location and registered services) collaboration to support mobility at high speed. The proposed communication model contributes to increased mobility by bringing data to the proximity before it is needed.

6G Networks are expected to host a massive number of devices per unit area. Massive communication in a limited area cannot tolerate collisions for network access. The proposed model provides collision-free radio communication via detailed access scheduling. In accordance with the schedule,

Capability of 6G	**How does the proposed model contribute?**	**Direct/Indirect Contribution**
Hyper-Low Latency	The communication model aims to provide the intelligence to the proximity before it is requested.	**D**
Ubiquitous Intelligence	Generated intelligence is transmitted to the UEs using a unidirectional data flow model.	**D**
Massive connectivity	Collision-free wireless communication enables massive number of devices connected.	**D**
Mobility	Generated ubiquitous data will be delivered to multiple edge nodes in the area to facilitate mobility.	**D**
Reliability	More than one edge node in the proximity will be involved in providing service.	**I**
Sustainability	Small devices are mostly battery-powered. They will be able to put themselves into sleep in idle times.	**I**

TABLE I
CONTRIBUTION TO 6G CAPABILITIES.

the devices will be able to put themselves into sleep to save energy for better sustainability.

The proposed model contributes indirectly to the reliability of the network providing the ubiquitous intelligence to the proximity. Thus, instantaneous network issues on the core network will not be noticed by the user.

The proposed model contributes most of the capabilities defined for 6G networks. It relies heavily on edge nodes to collect data from devices, schedule network access, and provide intelligence to their proximity.

VI. CONCLUSION

This study proposes a generic unidirectional communication model that allows network components to cooperate for hyper-low latency intelligence delivery to UE. The proposed model essentially considers the intelligence flow as a unidirectional data flow. The services provided to registered users are conveyed to their location, updated accordingly, and then delivered ubiquitously. Intelligence that is valuable at a particular location is served as a public service to every UE in that area.

The proposed communication model is applicable and contributes to various key capabilities of 6G networks, such as hyper-low latency, ubiquitous intelligence, ultra-dense connectivity, high mobility, reliability, and sustainability.

REFERENCES

[1] ITU-R, “IMT vision–framework and overall objectives of the future development of imt for 2020 and beyond,” *Recommendation ITU*, vol. 2083, no. 0, 2015.

[2] W. D. M. T. ITU-R, “The ITU-R framework for IMT-2030,” 2023.

[3] M. Latva-aho and K. Leppänen, “Key drivers and research challenges for 6G ubiquitous wireless intelligence,” Sep 2019. [Online]. Available: http://urn.fi/urn:isbn:9789526223544

[4] G. Liu, Y. Huang, N. Li, J. Dong, J. Jin, Q. Wang, and N. Li, “Vision, requirements and network architecture of 6G mobile network beyond 2030,” *China Communications*, vol. 17, no. 9, pp. 92–104, 2020.

[5] X. You, C.-X. Wang, J. Huang, X. Gao, Z. Zhang, M. Wang, Y. Huang, C. Zhang, Y. Jiang, J. Wang, *et al.*, “Towards 6G wireless communication networks: Vision, enabling technologies, and new paradigm shifts,” *Science China Information Sciences*, vol. 64, pp. 1–74, 2021.

[6] D. L. Quan Zhao, Narothum Saxena, “6G drivers and vision,” Apr 2021. [Online]. Available: https://www.ngmn.org/wp-content/uploads/NGMN-6G-Drivers-and-Vision-V1.0-final-New.pdf

[7] “From cloud AI to network AI, a view from 6GANA,” may 2021. [Online]. Available: https://www.6g-ana.com/upload/file/20210619/6375969458505193666851527.pdf

[8] C. J. Bernardos and M. A. Uusitalo, “European vision for the 6G network ecosystem,” jun 2021. [Online]. Available: https://doi.org/10.5281/zenodo.5007671

[9] “Artificial intelligence in next-generation connected systems,” White Paper, 2021. [Online]. Available: https://www.ericsson.com/en/reports-and-papers/white- papers/artificial-intelligence-in-next-generation-connected-systems

[10] “White paper on 6G vision and candidate technologies,” White Paper, Jun 2021. [Online]. Available: http://www.caict.ac.cn/kxyj/qwfb/ztbg/202106/P020210604552573543918.pdf

[11] E.-K. Hong, I. Lee, B. Shim, Y.-C. Ko, S.-H. Kim, S. Pack, K. Lee, S. Kim, J.-H. Kim, Y. Shin, Y. Kim, and H. Jung, “6G R&D vision: Requirements and candidate technologies,” *Journal of Communications and Networks*, vol. 24, no. 2, pp. 232–245, 2022.

[12] W. Tong and P. Zhu, “6G: The next horizon,” *TITLES*, p. 54, 2022.

[13] M. Corici, E. Troudt, and T. Magedanz, “An organic 6G core network architecture,” in *2022 25th Conference on Innovation in Clouds, Internet and Networks (ICIN)*, 2022, pp. 1–7.

[14] G. Liu, N. Li, J. Deng, Y. Wang, J. Sun, and Y. Huang, “The SOLIDS 6G mobile network architecture: driving forces, features, and functional topology,” *Engineering*, vol. 8, pp. 42–59, 2022.

[15] Y. Liu, Y. He, Y. Lin, and L. Tang, “Toward native artificial intelligence in 6G,” in *2022 IEEE International Symposium on Broadband Multimedia Systems and Broadcasting (BMSB)*. IEEE, 2022, pp. 1–6.

[16] Z. Feng, Z. Wei, X. Chen, H. Yang, Q. Zhang, and P. Zhang, “Joint communication, sensing, and computation enabled 6G intelligent machine system,” *IEEE Network*, vol. 35, no. 6, pp. 34–42, 2021.

[17] X. Shen, J. Gao, W. Wu, M. Li, C. Zhou, and W. Zhuang, “Holistic network virtualization and pervasive network intelligence for 6G,” *IEEE Communications Surveys & Tutorials*, vol. 24, no. 1, pp. 1–30, 2021.

[18] A. Alnoman and A. Anpalagan, “Computing-aware base station sleeping mechanism in h-cran-cloud-edge networks,” *IEEE Transactions on Cloud Computing*, vol. 9, no. 3, pp. 958–967, 2019.

[19] Q. Bi, “P-RAN: A distributed solution for cellular systems in high frequency bands,” *IEEE Network*, vol. 36, no. 4, pp. 86–91, 2022.

[20] C.-X. Wang, X. You, X. Gao, X. Zhu, Z. Li, C. Zhang, H. Wang, Y. Huang, Y. Chen, H. Haas, *et al.*, “On the road to 6G: Visions, requirements, key technologies and testbeds,” *IEEE Communications Surveys & Tutorials*, 2023.

[21] S. Chen, Y.-C. Liang, S. Sun, S. Kang, W. Cheng, and M. Peng, “Vision, requirements, and technology trend of 6G: How to tackle the challenges of system coverage, capacity, user data-rate and movement speed,” *IEEE Wireless Communications*, vol. 27, no. 2, pp. 218–228, 2020.

[22] H.-Y. Lin, “Large-scale artificial intelligence models,” *Computer*, vol. 55, no. 05, pp. 76–80, 2022.

[23] M. Xu, C. Song, S. Ilager, S. S. Gill, J. Zhao, K. Ye, and C. Xu, “Coscal: Multifaceted scaling of microservices with reinforcement learning,” *IEEE Transactions on Network and Service Management*, vol. 19, no. 4, pp. 3995–4009, 2022.

[24] R. Singh and S. S. Gill, “Edge ai: a survey,” *Internet of Things and Cyber-Physical Systems*, 2023.

[25] 3GPP, “Technical specification group radio access network;study on artificial intelligence (AI)/machine learning (ML) for NR air interface,” *Technical Report*, 2023. [Online]. Available: https://www.3gpp.org/ftp/Specs/archive/38_series/38.843/

[26] ——, “5G; 5G system; technical realization of service based architecture; stage 3,” *Technical Report*, 2020.

[27] C. Gentry, “Computing arbitrary functions of encrypted data,” *Communications of the ACM*, vol. 53, no. 3, pp. 97–105, 2010.

[28] M. ETSI, “Multi-access edge computing (MEC) framework and reference architecture,” *ETSI GS MEC*, vol. 3, p. V3, 2022-03.

[29] IEEE, “IEEE standard for low-rate wireless networks,” *IEEE Std 802.15.4-2020 (Revision of IEEE Std 802.15.4-2015)*, pp. 1–800, 2020.