

Review

Computational Architectures for 6G Networks: Integrating Distributed Computing and Edge Artificial Intelligence

Evelio Astaiza Hoyos¹, Héctor Fabio Bermúdez-Orozco^{1,*}  and Nasly Cristina Rodríguez-Idrobo² ¹ Electronic Engineering Program, Faculty of Engineering, University of Quindío, Armenia 630004, Quindío, Colombia; estaiza@uniquindio.edu.co² Occupational Health and Safety Program, Faculty of Health Sciences, University of Quindío, Armenia 630004, Quindío, Colombia; ncrodriguezi@uniquindio.edu.co

* Correspondence: hfbermudez@uniquindio.edu.co; Tel.: +57-3206671107

Abstract

This paper investigates the integration of distributed computing and edge Artificial Intelligence (edge AI) as foundational enablers of sixth-generation (6G) mobile networks. Through a systematic review following the Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) guidelines, encompassing over 200 peer-reviewed papers, architectural proposals, and standardization documents retrieved from IEEE Xplore, Scopus, Web of Science, MDPI, arXiv, ITU-R, 3GPP, and ETSI, this study provides a structured computational analysis of architectural approaches that integrate distributed computing paradigms and edge AI as core enablers of 6G. The analysis examines the evolution from cloud-centric to edge-centric computing, key edge AI techniques—including Federated Learning (FL), Split Learning (SL), and edge-adapted Large AI Models (LAMs)—and their role in enabling intelligent orchestration, resource optimization, and context-aware services. The comparative analysis demonstrates that edge computing architectures reduce end-to-end latency by 85–95% relative to cloud-centric deployments (under conditions of MEC servers within 1 km and 5G NR fronthaul), while federated learning with gradient compression achieves communication overhead reductions of up to 99% under IID data distributions and stable channel conditions. The results indicate that the tight integration of distributed computing and edge AI enhances network responsiveness, scalability, and adaptability, while also revealing persistent challenges related to orchestration complexity, resource constraints, security, and interoperability. The study concludes that holistic computational architectures and AI-native design principles are essential for the effective realization of 6G networks and for guiding future research and standardization efforts.



Academic Editors: Stefano Scanzio,
Pedro Santos, Ramez Daoud and
Jetmir Haxhibeqiri

Received: 21 March 2026

Revised: 31 May 2026

Accepted: 3 June 2026

Published: 5 June 2026

Copyright: © 2026 by the authors.
Licensee MDPI, Basel, Switzerland.
This article is an open access article
distributed under the terms and
conditions of the [Creative Commons
Attribution \(CC BY\)](https://creativecommons.org/licenses/by/4.0/) license.

Keywords: distributed computing; edge artificial intelligence; 6G network architectures; computational frameworks; AI-native networks; intelligent orchestration

1. Introduction

The evolution of mobile communication systems is increasingly characterized by the convergence of networking, distributed computing and artificial intelligence, transforming communication infrastructures into large-scale computational systems. As network complexity, data volume and service heterogeneity continue to grow, the efficient distribution and orchestration of computational and intelligent functions across the network have become central design challenges. In this context, future mobile networks are no longer conceived merely as platforms for data transmission, but as computation-driven and

intelligence-native infrastructures in which communication, computing and data processing are tightly integrated. This computational perspective provides the necessary foundation for understanding the architectural transformations required to support the next generation of mobile systems and frames the discussion of sixth-generation (6G) networks presented in the following sections.

1.1. The 6G Vision: Beyond Connectivity

Mobile networks have evolved dramatically from the first generation (1G), which introduced mobile telephony, to the fifth generation (5G), which began incorporating intelligence into mobile communications [1]. The sixth generation (6G), expected to reach commercial deployment around 2030, promises an even deeper transformation, marking the transition from “mobile intelligence” to an “Ubiquitous Intelligent Mobile Society” [2], opening new horizons for the integration of comfort, security, and intelligence into everyday life [3]. The 6G vision transcends incremental improvements in connectivity and aspires instead to achieve a seamless fusion of the physical, digital and human worlds [4], where artificial intelligence (AI) is not an overlaying application but an intrinsic capability of the network itself [5].

The International Telecommunication Union (ITU), through its Radiocommunication Sector (ITU-R), has established the foundational framework for 6G under the designation “IMT-2030” [6]. This framework defines several evolved use scenarios compared with 5G, such as immersive communication (enabling interactive experiences like XR and holographic telepresence), massive communication (supporting the Internet of Everything, IoE), and high-reliability, low-latency communication (HRLLC) for mission-critical applications [6]. It also introduces new scenarios enabled by emerging capabilities, including integrated sensing and communication (ISAC), native AI integration and ubiquitous connectivity spanning terrestrial, aerial, space and underwater environments [7].

To support these scenarios, 6G sets key performance indicators (KPIs) that are significantly more ambitious than those of 5G. These include peak data rates on the order of 1 Tbps [8]; user-experienced data rates of up to 10 Gbps [2]; end-to-end (E2E) latency below 1 ms, and even in the range of 0.1 ms to 100 μ s for specific cases [7]; connection densities potentially reaching 10^8 devices per km^2 [8]; extremely high reliability (success probability of 0.99999 to 0.9999999) [9]; mobility support for speeds up to 1000 km/h [2]; and substantial improvements in energy and spectral efficiency [7]. New capabilities, such as centimetre-level or even sub-centimetre-level positioning accuracy [7] and integrated sensing [1], also form an integral part of the vision.

Realizing this vision and meeting such demanding KPIs critically depend on the development and integration of radically new enabling technologies and network architectures [6,10–12]. Technologies such as terahertz (THz) and visible-light communication (VLC), reconfigurable intelligent surfaces (RIS), ultra-massive MIMO (UM-MIMO), non-terrestrial networks (NTN) and AI are considered fundamental pillars [1,12].

1.2. The Critical Role of Distributed Computing and Edge AI

Among the most crucial enabling technologies for 6G are distributed computing, particularly in the form of edge computing and multi-access edge computing (MEC), and edge AI [6]. Edge computing consists of bringing computational and storage capabilities closer to end users or data sources by situating them at the edge of the access network [4]. This contrasts with the traditional centralized cloud model and is essential for mitigating latency and reducing the load on transport networks [13].

Edge AI, in turn, refers to the execution of AI algorithms, both for training and inference, directly on edge nodes or even on end devices [14]. The convergence of dis-

tributed edge computing with AI is essential for realizing the 6G vision of “connected intelligence” [15]. This synergy not only enables intelligent and autonomous optimization of the 6G network itself but also supports a new generation of services requiring ultra-low latency, extensive local data processing and customized, context-aware AI capabilities. The deep integration of AI into the network architecture, often termed “AI-native” [5], fundamentally depends on the distributed computing infrastructure at the edge.

This study investigates the integration of distributed computing and edge AI for enabling 6G networks, with a focus on reducing latency, optimizing resources, and providing context-aware services. Specifically, it aims to provide an expert and comprehensive analysis of existing proposals for integrating distributed computing and edge AI as key elements for implementing 6G networks. It seeks to identify the synergies between these two technological domains, examine the inherent challenges of their joint deployment within the 6G context, and explore future perspectives and directions for research and standardization.

Figure 1 provides a conceptual taxonomy of the key enabling technologies and distributed computing paradigms discussed throughout this work.

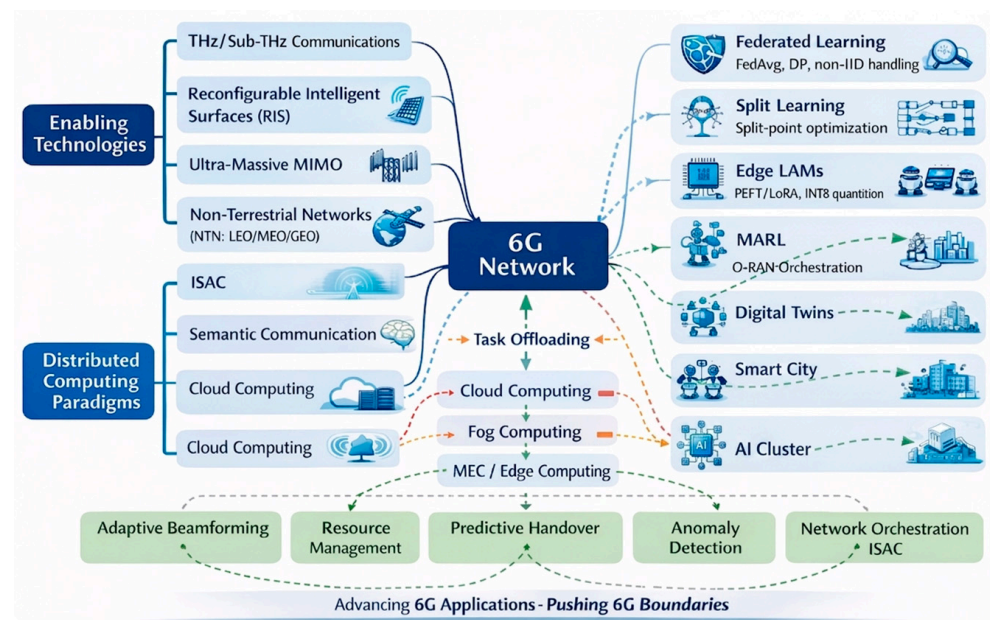


Figure 1. Taxonomy of 6G Enabling Technologies and Distributed Computing Architectures.

Section 2 examines the distributed computing paradigms relevant to 6G, tracing their evolution from cloud to edge and highlighting the benefits and challenges of edge computing. Section 3 delves into the concept of edge AI, describing key techniques (such as FL, SL and edge LAMs), their applications for optimizing the 6G network and enabling intelligent services, and the associated challenges. Section 4 analyses various architectural and orchestration proposals that aim to effectively integrate distributed computing and AI into 6G networks, including interactions with technologies such as Digital Twins and ISAC. Section 5 reviews the current state of standardization within major organizations (ITU-R, 3GPP, ETSI) and discusses open challenges and future research directions. Section 6 presents the main conclusions of the analysis, synthesizing the key findings and reinforcing the transformative role of distributed computing and edge AI for the future of mobile communications.

1.3. Contributions of This Work

This review article makes the following key contributions to the understanding and development of 6G networks:

1. **Comprehensive Synthesis:** It provides a structured integration of distributed computing paradigms (cloud, Fog, edge, MEC) and edge AI techniques (Federated Learning, Split Learning, edge LAMs) within the specific context of 6G requirements and use cases.
2. **Architectural Analysis:** The work systematically examines multiple architectural proposals, including evolutionary, revolutionary (TONA), satellite–terrestrial (Space–Air–Ground Integrated Network (SAGIN)), and O-RAN-based approaches, identifying their design principles, strengths, and applicability to different 6G scenarios.
3. **Critical Evaluation Framework:** Unlike purely descriptive surveys, this study evaluates the practical implications of integrating distributed intelligence, discussing trade-offs in terms of latency reduction, bandwidth optimization, privacy enhancement, and the challenges of orchestration complexity, resource constraints, and interoperability.
4. **Standardization Landscape:** The article maps the current state of 6G standardization efforts (ITU-R IMT-2030, 3GPP Releases 19–21, ETSI MEC), providing researchers and industry practitioners with a clear reference point for ongoing and future activities.
5. **Research Roadmap:** Based on the analysis of existing proposals and persistent challenges, the work identifies critical open research directions, including AI-native architecture design, trustworthy distributed AI, holistic security, and sustainability, that will guide the next phase of 6G development.

To better position the contribution of this survey within the existing literature, Table 1 presents a comparative analysis of representative review articles related to distributed computing, Edge AI, and 6G network architectures. The comparison demonstrates that, unlike previous surveys, the present work simultaneously addresses distributed computing paradigms, federated and split learning, edge Large AI Models, O-RAN, SAGIN/NTN integration, standardisation activities, and wireless channel considerations.

Table 1. Comparison of this survey with closely related reviews (2022–2025).

Survey	Year	Venue	Distrib. Computing	Edge AI (FL/SL)	Edge LAMs	O-RAN	SAGIN/NTN	Standardization	Wireless Channel Analysis
Wang et al. [16]	2020	IEEE CST	✓	Partial	✗	✗	✗	Partial	✗
Letaief et al. [17]	2022	IEEE JSAC	Partial	✓	✗	Partial	✗	✗	✓
Abreha et al. [18]	2022	Sensors	✗	✓ (FL)	✗	✗	✗	✗	✗
Al-Ansi et al. [4]	2021	Future Internet	✓	Partial	✗	✗	✗	✗	✗
Chen et al. [19]	2023	arXiv	Partial	✓	✓	Partial	✗	Partial	✗
Kairouz et al. [20]	2021	Found. ML	✗	✓ (FL)	✗	✗	✗	✗	✗
This work	2026	JSAN	✓	✓	✓	✓	✓	✓	✓

Legend: ✓ = comprehensive coverage; Partial = limited coverage; ✗ = not covered.

These contributions collectively position distributed intelligence as a foundational design principle for 6G networks and provide actionable insights for both academic research and industrial implementation, as further elaborated in the Section 6.

- Novelty and Differentiation from Conventional Approaches

This work distinguishes itself from the conventional 5G and early 6G literature in several critical dimensions:

Shift from Connectivity-Centric to Intelligence-Centric Architecture: Traditional mobile network reviews focus on radio access technologies (spectrum, MIMO, beamforming) and connectivity improvements (throughput, coverage). This work repositions the architectural discourse around distributed intelligence and edge computing as primary design drivers, reflecting 6G's fundamental departure from previous generations. While 5G introduced edge computing as an optional enhancement, 6G mandates it as a core architectural principle to meet sub-millisecond latency and support AI-native services.

Holistic Integration versus Component-Level Analysis: Unlike studies that examine AI, edge computing, or network architecture in isolation, this review provides an integrative framework showing how distributed computing paradigms (MEC, Fog, edge) and edge AI techniques (FL, SL, edge LAMs) must be co-designed with network architecture, orchestration mechanisms, and enabling technologies (Digital Twins, ISAC, Blockchain). This systems-level perspective is essential for understanding 6G's complexity but is often absent in the component-focused literature.

Critical Evaluation versus Descriptive Cataloguing: Conventional surveys often catalogue technologies and proposals without rigorous comparative analysis. This work explicitly evaluates trade-offs (complexity vs. performance, scalability limits, deployment feasibility, economic costs), identifies gaps between vision and implementation, and grounds the discussion in quantitative performance benchmarks and real-world testbed results. This critical stance provides actionable guidance for researchers and practitioners rather than merely documenting the state-of-the-art.

Standardization-Aware Perspective: By tracking ongoing efforts across ITU-R (IMT-2030), 3GPP (Releases 19–21), and ETSI MEC, this review bridges academic research and industry standardization, highlighting where consensus exists and where open questions remain. This dual focus is particularly valuable as 6G transitions from vision to concrete specification.

Forward-Looking Research Roadmap: Rather than concluding with generic recommendations, this work identifies specific, actionable research challenges—trustworthy distributed AI, AI-native architecture design, holistic security, and sustainability—that are both technically precise and aligned with industry priorities. These directions emerge organically from the critical analysis rather than being imposed as afterthoughts.

In summary, the novelty lies not in proposing a new architecture or algorithm, but in providing a comprehensive, critical, and actionable synthesis that positions distributed intelligence as the foundational principle for 6G, backed by performance analysis, real-world validation, and clear identification of research frontiers.

1.4. Methodology and Article Selection Criteria

This systematic review follows the Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) guidelines [PRISMA 2020] to ensure a rigorous and transparent methodology for identifying, selecting, and analyzing the relevant literature on distributed computing and edge AI for 6G networks.

1.4.1. Article Selection Criteria:

The literature corpus was constructed through a comprehensive search of major scientific databases, including IEEE Xplore, Scopus, Web of Science, MDPI, and arXiv, as well as standardization documents from ITU-R, 3GPP, and ETSI. The search strategy employed the following inclusion criteria:

1. Temporal Scope: Publications from 2020 to 2025, covering the period of active 6G research and standardization.
2. Technical Relevance: Articles addressing distributed computing paradigms (cloud, Fog, edge, MEC), edge AI techniques (Federated Learning, Split Learning, edge Large AI Models), 6G architectural proposals, or related enabling technologies (Digital Twins, ISAC, Blockchain, Multi-Agent Reinforcement Learning).
3. Scientific Merit: Peer-reviewed journal articles, conference papers from recognized venues (e.g., IEEE, ACM), white papers from standardization bodies, and high-quality preprints from established repositories (arXiv).
4. Language: Publications in English.

Exclusion criteria included: purely theoretical AI work without network context, studies focused exclusively on previous generations (4G, 5G) without forward-looking 6G relevance, and non-technical opinion pieces lacking substantive architectural or algorithmic content.

1.4.2. Search Strategy

Keywords and Boolean operators were used to identify relevant works: ("6G" OR "sixth generation" OR "IMT-2030") AND ("Edge AI" OR "Federated Learning" OR "Split Learning" OR "Edge Computing" OR "MEC" OR "distributed intelligence" OR "AI-native") AND ("architecture" OR "framework" OR "orchestration" OR "optimization").

1.4.3. Screening Process

Initial screening based on titles and abstracts yielded approximately 227 candidate papers. Full-text review and quality assessment reduced this to the final corpus of cited works (approximately 83 primary references), distributed as follows: approximately 25 papers on MEC and edge computing architectures, 20 on Federated Learning and Split Learning for 6G, 15 on O-RAN and orchestration frameworks, 12 on edge LAMs and large-scale AI at the edge, and 11 on standardization and enabling technologies (ISAC, Digital Twins, Blockchain). This selection ensures diversity across architectural proposals, AI techniques, use cases, and standardization activities. Data synthesis followed a qualitative thematic analysis approach, grouping findings by research question and technology domain. A formal meta-analysis was not performed due to the significant heterogeneity of study designs, performance metrics, and experimental conditions across the reviewed works. Priority was given to papers providing concrete architectural frameworks, experimental validation, performance analysis, or authoritative standardization guidance.

1.4.4. Research Questions

This systematic review is structured around five core research questions (RQs) that guide the literature selection and analysis:

- RQ1: What are the dominant distributed computing paradigms (cloud, Fog, MEC/edge) proposed for 6G networks, and how do their latency and bandwidth characteristics align with 6G performance requirements?
- RQ2: Which edge AI techniques (Federated Learning, Split Learning, edge LAMs) have been proposed for resource-constrained 6G edge environments, and what are their comparative advantages and limitations in terms of communication overhead, privacy, and convergence under wireless channel impairments?
- RQ3: What architectural frameworks (TONA, SAGIN, O-RAN-based, Intent-Based Networking (IBN)) have been proposed to integrate distributed computing and Edge AI into 6G networks, and what are the key design trade-offs?

- RQ4: What is the current state of standardization (ITU-R, 3GPP, ETSI, O-RAN Alliance) for AI-native and distributed computing capabilities in 6G, and what gaps remain between research proposals and adopted standards?
- RQ5: What are the principal open research challenges and future directions for realizing AI-native 6G networks with distributed intelligence?

Figure 2 presents the PRISMA flow diagram summarizing the systematic literature selection process.

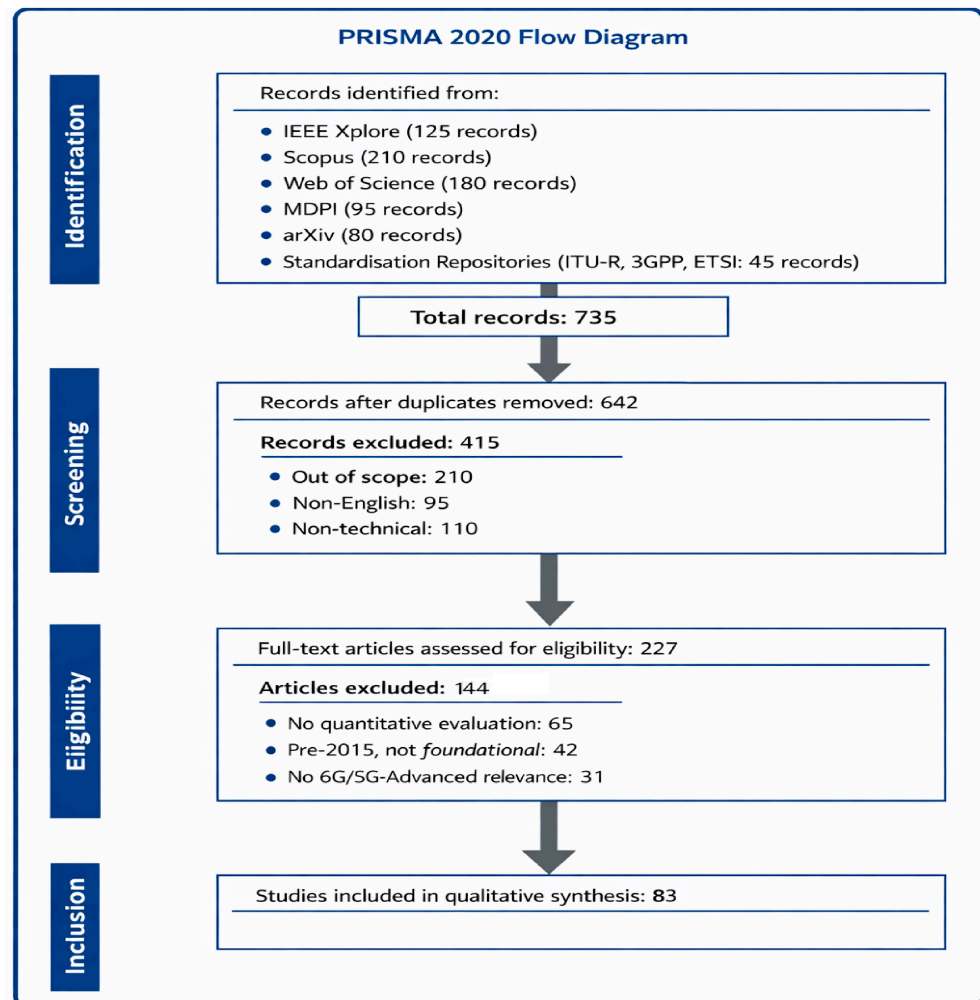


Figure 2. PRISMA Flow Diagram.

This methodological rigour ensures that the review comprehensively covers the state-of-the-art while maintaining focus on the integration of distributed computing and edge AI as foundational enablers of 6G networks.

2. Distributed Computing Paradigms in the 6G Context

The evolution towards sixth-generation (6G) mobile networks involves not only advances in radio technologies, but also a fundamental transformation in how computation, data processing and intelligence are distributed across the network. From a computational perspective, 6G can be understood as a large-scale, heterogeneous distributed computing system in which communication, computing and storage resources must be jointly orchestrated to meet extreme performance requirements. In this context, distributed computing, particularly edge-oriented paradigms, emerges as a core architectural shift required to support ultra-low latency, massive scalability and intelligent network operation.

These paradigms redefine where computational tasks are executed, how resources are allocated and how system-level efficiency, responsiveness and reliability are achieved in next-generation mobile networks.

2.1. Historical Evolution: From Centralized Cloud to the Distributed Edge

For the past few decades, the dominant paradigm has been cloud computing. According to the NIST definition, this is a model that enables ubiquitous, convenient and on-demand access to a shared pool of configurable computing resources (networks, servers, storage, applications and services) that can be rapidly provisioned and released with minimal management effort or service provider interaction [15]. This centralized model has been fundamental for large-scale data processing and the deployment of massive applications [21].

However, the growing proliferation of connected devices (Internet of Things, IoT) and the emergence of applications with strict real-time requirements, such as augmented/virtual reality (AR/VR), autonomous driving and industrial automation, have highlighted the inherent limitations of the centralized cloud model [4]. The physical distance between end devices and cloud data centres introduces significant latency, which is incompatible with the millisecond- or even microsecond-level requirements of many 6G applications [21]. Furthermore, transmitting the massive volumes of data generated at the edge, estimated to represent 75% of enterprise data by 2025 [22], to the cloud consumes substantial bandwidth and raises serious concerns regarding the privacy and security of sensitive information [4].

Fog computing emerged in response to these limitations. This paradigm introduces an intermediate layer of distributed computing infrastructure positioned between end devices and the centralized cloud [13]. Fog nodes, although more resource-constrained than the cloud [23], are geographically closer to users, enabling the local processing of latency-sensitive tasks or smaller data volumes, thereby reducing dependence on the cloud [24]. The typical architecture of Fog computing is often described as comprising three layers: the terminal/IoT layer, the Fog layer and the cloud layer [24,25].

Edge computing represents the logical extension of this decentralization trend, bringing computational and storage capabilities even closer to the end user, directly to the edge of the access network [4,13]. Within this broad concept, MEC is a key initiative, particularly relevant for mobile networks.

This progression from centralized cloud to distributed edge is not merely another technological option but a necessary and inevitable response to the fundamental requirements of the 6G vision. Applications that will define 6G, such as immersive XR experiences [6], cooperative autonomous driving [26] and the Tactile Internet [1], impose ultra-low latency requirements (<1 ms) [8] and generate massive locally produced data streams [27]. The physics of signal propagation and the bandwidth limitations of transport networks make it infeasible to satisfy these requirements using distant centralized cloud infrastructures [21]. The need to process critical information in real time and to preserve the privacy of sensitive data (such as biometric data in XR or vehicular data) [21] further reinforces the imperative of bringing computation closer to the source. Therefore, the adoption of paradigms such as Fog computing and, especially, edge computing becomes a *sine qua non* condition for enabling many of the most transformative 6G use cases.

2.2. Multi-Access Edge Computing (MEC) as a Key Enabler in 6G

The European Telecommunications Standards Institute (ETSI) defines MEC as a technology that provides application developers and content providers with cloud-computing capabilities and an IT service environment at the edge of the multi-access network (includ-

ing mobile networks such as the RAN, as well as Wi-Fi and others) [28]. This environment is characterized by ultra-low latency, high bandwidth and real-time access to radio network information, which can be exploited by applications [28].

Current Deployment Status of MEC: Commercial MEC deployments are already active in LTE and 5G networks worldwide. ETSI MEC Phase 3 standards (ISG MEC 003, 010, 016) provide the normative framework for MEC platform APIs and application lifecycle management. Major operators, including Deutsche Telekom, Verizon, and NTT DOCOMO, have deployed MEC servers at cellular base station sites, primarily supporting low-latency video processing, V2X, and industrial automation use cases [29–31]. The O-RAN Alliance's MEC integration specifications are enabling vendor-neutral MEC deployments across disaggregated RAN environments, providing a practical foundation for the more advanced, AI-native MEC capabilities envisioned for 6G.

MEC, already a significant component in the evolution of 5G [1], is considered even more critical for 6G [8]. The MEC architecture is expected to undergo significant transformation in 6G, becoming more highly distributed, extending further towards the “Far Edge” (computing nodes located closer to the user, potentially within radio units or cellular sites) [32], and adopting principles of openness and disaggregation [33].

The deep integration of MEC into the 6G architecture is essential for supporting the most demanding use cases [34]. Examples are presented as follows.

2.2.1. Connected and Autonomous Vehicles (V2X)

MEC provides the low latency required for safety-critical V2V and V2I communications (collision avoidance, coordinated driving) and enables the local processing of sensor data for environmental perception [4].

2.2.2. Extended Reality (XR) and the Metaverse

Immersive experiences require intensive graphical rendering and extremely low-latency responses to user actions. MEC enables the offloading of part of this computation to the edge, improving Quality of Experience (QoE) and reducing the processing burden on end devices [13].

2.2.3. Industrial Automation and Robotics

Real-time control of robots and cyber-physical systems (CPS) in Industry 4.0 demands microsecond-level latency and high reliability, both of which MEC can support [8].

2.2.4. Drones and UAVs

Fleet management, autonomous navigation and onboard sensor-data processing can be enhanced through computational support provided by terrestrial or aerial MEC nodes [1,35,36].

2.2.5. Telemedicine

Applications such as remote surgery or AI-assisted diagnosis require low latency and secure processing of sensitive medical data, making MEC an ideal enabler [8].

Moreover, the evolution towards Open-Source MEC (OS-MEC) is actively being explored [33]. Traditional MEC architectures may be rigid, as they depend on specialized hardware integrated with proprietary software. OS-MEC proposes decoupling MEC functions (software) from underlying resources (hardware) using technologies such as Network Function Virtualization (NFV) and Software-Defined Networking (SDN). This would allow for the dynamic reconfiguration of functions and resources to create customized MEC services tailored to specific 6G scenarios, promoting innovation and flexibility [33].

The performance of MEC in 6G is fundamentally conditioned by the wireless radio channel connecting user equipment to the MEC-hosting base station or distributed unit. Unlike fixed backhaul, the radio access link is subject to multipath fading (Rayleigh/Rician models), shadowing, Doppler shifts at high mobility (up to 1000 km/h for 6G [37–39], and inter-cell interference. These channel impairments directly affect the transmission latency L_i^{trans} in the E2E latency model: for a time-varying channel $h_{ij}(t)$ with instantaneous SNR $\gamma_{ij}(t) = p_{ij}|h_{ij}(t)|^2 / (N_0 b_{ij})$, the expected transmission latency becomes $E[L_i^{\text{trans}}] = D_i / E[b_{ij} \log_2(1 + \gamma_{ij}(t))]$, which is strictly greater than the latency computed using the average SNR due to Jensen's inequality. Furthermore, handover events in high-mobility scenarios introduce additional latency spikes, with interruption times reported in the literature ranging from approximately 10 to 50 ms in LTE-Advanced networks, whereas optimized 5G NR deployments can achieve significantly lower interruption times. The corresponding mobility management procedures are specified in 3GPP TS 36.331 for LTE-A and in 3GPP TS 38.331, Section 5.3.5, for NR [40]. These mobility-induced latency variations must therefore be considered in MEC task-offloading schedulers. Channel-aware offloading algorithms that jointly optimize task placement and radio resource allocation under fading conditions are therefore essential for meeting 6G URLLC targets [38,41].

2.3. Fundamental Benefits of Edge Computing for 6G

The widespread adoption of edge computing in 6G networks offers several essential advantages.

2.3.1. Drastic Reduction in Latency

By processing data and executing applications closer to the user or the data source, the round-trip time to a centralized cloud is eliminated. This is absolutely critical for meeting the ultra-low latency requirements (<1 ms, and even 0.1 ms) of 6G scenarios such as HURLLC, the Tactile Internet, immersive XR, cloud gaming, remote surgery and autonomous driving [4].

Mathematical Latency Model:

The following E2E latency decomposition is a standard formulation widely used in the edge and mobile cloud computing literature [4,38,42]; the specific numerical examples and adaptation for 6G use cases are original contributions of this work.

The end-to-end latency for task execution can be decomposed as follows:

$$L_{E2E} = L_{\text{trans}} + L_{\text{prop}} + L_{\text{queue}} + L_{\text{comp}} + L_{\text{backhaul}} \quad (1)$$

where the following hold:

- $L_{\text{trans}} = \frac{D}{R_{\text{data}}}$: transmission time for data size D at rate R_{data} .
- $L_{\text{prop}} = \frac{d}{c}$: propagation delay over distance d (speed of light $c \approx 3 \times 10^8$ m/s).
- L_{queue} : queuing delay at intermediate nodes.
- $L_{\text{comp}} = \frac{C}{f}$: computation time for C operations at frequency f .
- L_{backhaul} : backhaul network delay.

Cloud vs. Edge Latency Comparison

For cloud computing (data centre at distance $d_{\text{cloud}} \approx 1000$ km):

$$L_{\text{cloud}} = \frac{D}{R} + \frac{2 \times d_{\text{cloud}}}{c} + L_{\text{queue}}^{\text{cloud}} + \frac{C}{f_{\text{cloud}}} + L_{\text{backhaul}} \quad (2)$$

with the following typical values:

- Propagation: $\frac{2 \times 1000 \times 10^3}{3 \times 10^8} \approx 6.7$ ms (round-trip).

- Backhaul: $L_{\text{backhaul}} \approx 10\text{--}50$ ms [4,38,43,44].
- Queuing: $L_{\text{queue}}^{\text{cloud}} \approx 5\text{--}20$ ms [4,38] (M/M/1 queuing model estimates).
- Total: $L_{\text{cloud}} \approx 20\text{--}80$ ms. (Note: The total cloud latency of 20–80 ms is computed as the sum of transmission (1.7–10 ms for $D = 1\text{--}10$ MB over $R = 100$ Mbps–1 Gbps), backhaul (10–50 ms), and queuing (5–20 ms) delays, assuming typical parameter ranges. The lower bound 16.7 ms is rounded to 20 ms to account for additional processing overhead.) (Note: The edge latency of 0.5–5 ms is computed with co-located MEC (backhaul ≈ 0), edge queuing (0.1–0.5 ms), and reduced transmission delay due to proximity.)
- For edge computing (MEC server at distance $d_{\text{edge}} \approx 1\text{--}5$ km):

$$L_{\text{edge}} = \frac{D}{R} + \frac{2 \times d_{\text{edge}}}{c} + L_{\text{queue}}^{\text{edge}} + \frac{C}{f_{\text{edge}}} \quad (3)$$

with the following typical values:

- Propagation: $\frac{2 \times 5 \times 10^3}{3 \times 10^8} \approx 0.033$ ms (negligible).
- No backhaul delay.
- Queuing: $L_{\text{queue}}^{\text{edge}} \approx 0.1\text{--}2$ ms (lower load).
- Total: $L_{\text{edge}} \approx 0.5\text{--}5$ ms.

Latency Reduction Factor:

$$\text{Reduction} = \frac{L_{\text{cloud}} - L_{\text{edge}}}{L_{\text{cloud}}} \times 100\% \approx 85\text{--}95\% \quad (4)$$

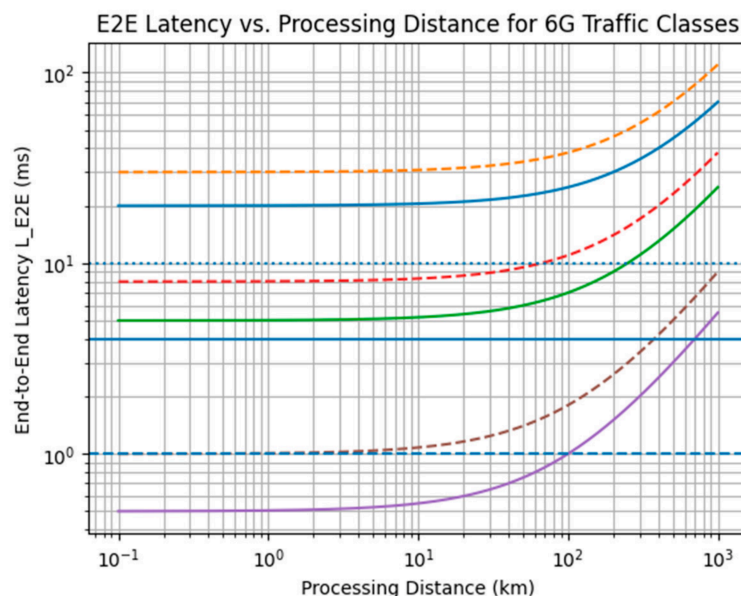
6G Ultra-Low Latency Achievement:

- For HRLLC requiring $L_{\text{E2E}} < 1$ ms, the following hold.
- Wireless transmission (5G NR, mini-slot): $L_{\text{trans}} \approx 0.125$ ms (5G NR mini-slot scheduling: one OFDM mini-slot with 2 symbols at 30 kHz SCS [1]).
- Propagation (<1 km): $L_{\text{prop}} \approx 0.003$ ms (from $L_{\text{prop}} = d/c$ with $d < 1$ km).
- Edge computation (GPU/NPU): $L_{\text{comp}} \approx 0.2\text{--}0.5$ ms [38,41,45,46] (GPU inference benchmarks on edge hardware).
- Processing delay: $L_{\text{proc}} \approx 0.1$ ms [38,45,46] (GPU inference benchmarks on edge hardware).
- Total: $L_{\text{edge}}^{\text{6G}} \approx 0.4\text{--}0.8$ ms ✓ (meets requirement).
- Cloud deployment would require $L_{\text{cloud}} \geq 20$ ms, failing the <1 ms requirement by $20\times$.

Assumptions and Sensitivity Analysis for the Latency Model: The E2E latency decomposition above relies on several simplifying assumptions. First, queuing delay is modelled under an M/M/1 assumption (Poisson arrivals, exponential service times), yielding typical values of 1–3 ms for lightly loaded edge nodes; under an M/D/1 model (deterministic service times), queuing delay is halved for the same server utilization. Second, backhaul latency assumes typical fibre-optic transport (1–5 ms for metropolitan distances). Third, the 85–95% latency reduction is specifically achieved when: (i) the MEC server is co-located within 1 km of the user equipment, (ii) a 5G NR fronthaul link is used, (iii) task data size is 1–10 MB, and (iv) edge server utilization is below 70%. Results may vary plus or minus 20–30% under stochastic traffic conditions (bursty Poisson arrivals) and wireless channel impairments such as Rayleigh fading, Doppler spreading at high mobility, inter-cell interference, and handover latency spikes of 5–50 ms. These handover-induced delays are modelled as additive stochastic terms in the E2E latency (Equation (2)), affecting the transmission delay component t_{tx} during handover events. Suburban or rural scenarios with longer MEC server distances and higher server load may yield lower latency improvements

in the range of 60–80%, highlighting the importance of careful deployment planning for 6G edge infrastructure.

Figure 3 illustrates E2E latency as a function of processing distance for the three computing paradigms under different 6G traffic classes. Note: The parameters $B = 100$ MHz and $SNR = 20$ dB shown in Figure 3 correspond to the Shannon capacity bound used in Equation (1) ($C = B \cdot \log_2(1 + SNR)$) to compute the achievable data rate R_{data} for the transmission delay component of the latency model.



Values derived from the analytical model in Eq. (1) with: $B = 100$ MHz, $SNR = 20$ dB, server utilisation $\rho = 0.7$.

Figure 3. E2E Latency versus Processing Distance for Cloud, Fog, and MEC under 6G Traffic Classes.

2.3.2. Optimized Bandwidth Usage

Local processing significantly reduces the amount of data that must be transmitted across backhaul networks and the network core [4]. This is vital in 6G, where an explosion in data volume is expected from billions of IoT devices, sensors and multimedia applications [27]. Alleviating backhaul congestion improves overall network performance.

Mathematical Bandwidth Optimization Model:

The following bandwidth reduction framework is adapted from standard edge offloading models reported in the literature [4,38]. Specifically, the formulation builds upon the edge-computing architectural concepts discussed in [4] (Section 4) and the communication-efficient resource management framework presented in [38] (Section III), while the 6G-oriented parametrization, analytical assumptions, and illustrative smart-city case study are original contributions of this work.

Consider a 6G network with N IoT devices generating raw data at rate r_i (bps per device):

Cloud-Only Scenario:

All raw data transmitted to cloud:

$$B_{\text{backhaul}}^{\text{cloud}} = \sum_{i=1}^N r_i \quad (5)$$

Edge Computing Scenario:

Data processed at edge, only aggregated results/alerts transmitted:

$$B_{\text{backhaul}}^{\text{edge}} = \sum_{i=1}^N c_i \cdot r_i \quad (6)$$

where $c_i \in [0, 1]$ is the compression/aggregation ratio for device i .

Bandwidth Reduction:

$$\Delta B = \frac{B_{\text{backhaul}}^{\text{cloud}} - B_{\text{backhaul}}^{\text{edge}}}{B_{\text{backhaul}}^{\text{cloud}}} = 1 - \frac{\sum_i c_i r_i}{\sum_i r_i} \quad (7)$$

Concrete Example—Smart City Video Surveillance:

- $N = 1000$ cameras.
- Raw video rate: $r_i = 25$ Mbps (4K video at 30 fps).
- Total raw data: $B_{\text{cloud}} = 1000 \times 25 = 25,000$ Mbps = 25 Gbps.

with edge AI processing (object detection, anomaly detection):

- Transmitted data: metadata + alerts only.
- Compression ratio: $c_i \approx 0.001$ (only events transmitted).
- Aggregated data: $B_{\text{edge}} = 1000 \times 25 \times 0.001 = 25$ Mbps.

Bandwidth Reduction:

$$\Delta B = \frac{25,000 - 25}{25,000} = 0.999 = 99.9\% \quad (8)$$

This saves 24.975 Gbps of backhaul bandwidth, reducing infrastructure costs and enabling scalability.

General Edge Bandwidth Model:

For heterogeneous traffic types (video, sensor data, control signals):

$$B_{\text{edge}} = \sum_{k \in \text{types}} N_k \cdot r_k \cdot c_k \cdot \alpha_k \quad (9)$$

where the following hold:

- N_k : number of devices of type k .
- r_k : raw data rate per device.
- c_k : compression ratio after edge processing.
- α_k : activity factor (fraction of time transmitting).

Typical compression ratios:

- Video (with edge AI): $c_k \approx 0.001$ – 0.01 (99–99.9% reduction) [4,16,47,48].
- Sensor data (aggregated): $c_k \approx 0.05$ – 0.1 (90–95% reduction) [4,41].
- Control signals: $c_k \approx 0.1$ – 0.5 (50–90% reduction) [4,41].

Enhanced Privacy and Security

Keeping sensitive data (personal, medical, industrial, vehicular) at the edge, without the need to transmit it to a potentially less trustworthy or more exposed central cloud, intrinsically enhances privacy [5]. Local processing reduces the attack surface associated with long-distance data transmission.

Mathematical Privacy Model:

The following information-theoretic privacy framework is based on the mutual-information formulation introduced by Letaief et al. [17] and the privacy-leakage analysis based on mutual-information bounds presented by Rassouli and Gündüz [49]. These con-

cepts are combined with standard differential privacy theory. The healthcare-data example and its 6G edge-AI parametrization are original contributions of this work.

Define privacy leakage as the mutual information between raw data X and transmitted information Y :

$$\text{PrivacyLeakage} = I(X; Y) = H(X) - H(X|Y) \quad (10)$$

where $H(\cdot)$ is entropy and $H(X|Y)$ is conditional entropy.

Cloud Scenario: Raw data transmitted $Y = X$:

$$I(X; X) = H(X) - H(X|X) = H(X) - 0 = H(X) \quad (11)$$

(maximum leakage).

Edge Scenario: Only processed features $Y = f(X)$ transmitted:

$$I(X; f(X)) = H(X) - H(X|f(X)) < H(X) \quad (12)$$

The reduction in privacy leakage is:

$$\Delta_{\text{privacy}} = H(X) - I(X; f(X)) = H(X|f(X)) \quad (13)$$

For effective edge processing, $H(X|f(X))$ should be large (high uncertainty about X given $f(X)$).

Example—Healthcare Data:

- Raw medical images: $H(X) \approx 10^6$ bits (high entropy) (illustrative example: a 1 Mpixel greyscale image at 1 bit/pixel entropy yields $\sim 10^6$ bits [17]).
- Edge-processed diagnosis: $H(f(X)) \approx 10$ bits (binary or categorical) ($\log_2(C)$ bits for C disease classes [17]).
- Privacy improvement: $\Delta_{\text{privacy}} \approx 10^6 - 10 \approx 10^6$ bits (derived from $H(X) - H(f(X))$ as defined above).

2.3.3. Enablement of Context Awareness and Localization

Physical proximity allows edge applications and services to access and react to local information in real time, such as radio channel conditions, precise device location or events in the physical environment [4]. This is crucial for personalized and adaptive services.

Mathematical Context-Aware Service Model:

The context-aware decision function below is an original formulation introduced in this work to model the edge computing advantage in providing real-time, location-aware, and context-sensitive service optimization in 6G environments.

Context-aware decision function:

$$a^* = \underset{a \in \mathcal{A}}{\operatorname{argmax}} U(a | c_{\text{local}}, c_{\text{global}}) \quad (14)$$

where the following hold:

- a : action/service configuration.
- c_{local} : local context (location, channel, nearby devices).
- c_{global} : global context (network state).
- $U(\cdot)$: utility function.

Edge advantage: Access to c_{local} with latency $L_{\text{local}} < 1$ ms vs. cloud requiring $L_{\text{cloud}} > 20$ ms.

For time-critical contexts (e.g., vehicle collision avoidance), only edge can react within safety margins.

2.3.4. Greater Scalability and Reliability

Distributed edge computing architectures can be inherently more scalable, allowing capacity to be added incrementally where needed. Moreover, they avoid the single points of failure typical of centralized systems, thereby improving overall system resilience [13].

Mathematical Reliability Model:

The following reliability and scalability model is an original formulation introduced in this work to quantify the resilience advantage of distributed edge over centralized cloud deployments for 6G use cases.

System reliability for M independent edge nodes each with reliability p :

$$R_{edge} = 1 - (1 - p)^M \quad (15)$$

For $M = 10$ nodes with $p = 0.99$:

$$R_{edge} = 1 - (1 - 0.99)^{10} = 1 - 10^{-20} \approx 1 \cdot (\text{extremely high})$$

Centralized cloud with single data centre reliability $p_{cloud} = 0.9999$:

$$R_{cloud} = 0.9999$$

Even with high individual reliability, distributed edge achieves better overall system reliability through redundancy.

Scalability—Linear vs. Centralized:

- Edge: Adding capacity scales linearly: $C_{total} = M \cdot C_{node}$.
- Cloud: Centralized bottleneck limits scalability beyond infrastructure capacity.

2.4. Inherent Challenges of Distributed Edge Computing

Despite its advantages, the implementation of distributed edge computing for 6G presents significant challenges that must be addressed.

2.4.1. Resource Constraints

Edge devices and nodes (ranging from sensors and smartphones to MEC servers) possess considerably lower computational (CPU, GPU, NPU), storage and energy capabilities than centralized cloud data centres [4]. These limitations hinder the execution of computationally intensive tasks, particularly complex AI algorithms such as large language models (LLMs) [13].

2.4.2. Complex Management and Orchestration

Managing an ecosystem of heterogeneous, geographically distributed and dynamic computing resources is inherently complex [6]. It requires sophisticated solutions for resource discovery, task allocation (offloading), load balancing, application lifecycle management and Quality of Service (QoS) assurance [4].

2.4.3. Security and Privacy

Although the edge enhances privacy by localizing data, the distributed and often open nature of edge infrastructures increases the attack surface [50]. Challenges include the physical security of edge nodes, the heterogeneity of operating systems and protocols, limited user interfaces on IoT devices, weak computational capabilities at peripheral devices for robust defences, container security in MEC, vulnerabilities in SDN (centralized controller, switch saturation), complex access management in mobile and multi-domain environments, the risks associated with open interfaces (such as in O-RAN) and multi-

vendor interoperability issues [50]. Robust mechanisms for authentication, authorization, encryption, data protection, intrusion detection and trust management are essential [13].

New Attack Surfaces in Distributed Edge AI: Beyond traditional network security challenges, AI deployment at distributed edge nodes introduces novel attack vectors. (1) **Model Poisoning Attacks:** In federated settings, compromised clients can inject malicious gradient updates designed to degrade global model performance or introduce backdoors (causing misclassification of specific inputs). Byzantine-robust aggregation algorithms—including Krum, coordinate-wise median, and FLTrust—detect and exclude outlier updates, reducing poisoning success rates to less than 5% even with 30% malicious clients [51,52]. (2) **Adversarial Inference Attacks at Edge Nodes:** Edge servers performing inference are vulnerable to adversarial examples—crafted inputs that cause misclassification. In 6G network management (e.g., AI-based intrusion detection), adversarial inputs could mask malicious traffic. Adversarial training and certified robustness techniques provide defences but increase inference latency by 10–30% [53]. (3) **Compromised Edge Infrastructure:** Physical or software compromise of edge nodes (via supply-chain attacks or insider threats) can expose model parameters, user data, and network telemetry. Mitigation strategies include: Trusted Execution Environments (TEE, e.g., Intel SGX, ARM TrustZone) that isolate AI model execution in hardware-protected enclaves; differential privacy (DP), which provides formal bounds on information leakage from model outputs or gradient updates; and Byzantine-robust aggregation combined with DP and TEE for defence-in-depth in distributed AI systems. These attack surfaces underscore the need for a holistic security architecture co-designed with the distributed AI framework from the outset.

2.4.4. Mobility and Intermittent Connectivity

Ensuring service continuity and maintaining the application state for mobile users and devices as they move across different access points or edge nodes is a major challenge [4]. This is particularly critical in V2X and UAV scenarios, where network topology changes rapidly [26]. Intermittent connectivity can also disrupt distributed tasks such as AI model training [54].

2.4.5. Interoperability and Standardization

The diversity of hardware, software and protocols in the edge ecosystem can lead to interoperability issues unless clear and open standards are established [50]. The lack of standardization hinders the creation of a unified market and limits application portability [55].

The transition towards more open and disaggregated architectures, such as Open-Source MEC (OS-MEC) [33] or Open RAN (O-RAN)-based frameworks [50], while promising in terms of flexibility, innovation and reduced vendor lock-in, intensifies several of these challenges. Open interfaces, by definition, increase the surface exposed to potential attacks [50]. Multi-vendor component management introduces additional complexity in ensuring security and interoperability [50]. Functional decoupling requires more sophisticated orchestration mechanisms to compose and manage end-to-end services [33]. This dynamic highlights a fundamental tension: the pursuit of openness and flexibility at the 6G edge creates an even more urgent need for advanced and automated solutions for managing security, trust and orchestration, possibly through AI-driven approaches or distributed ledger technologies such as blockchain [13]. Table 2 summarizes the key characteristics of the computing paradigms discussed.

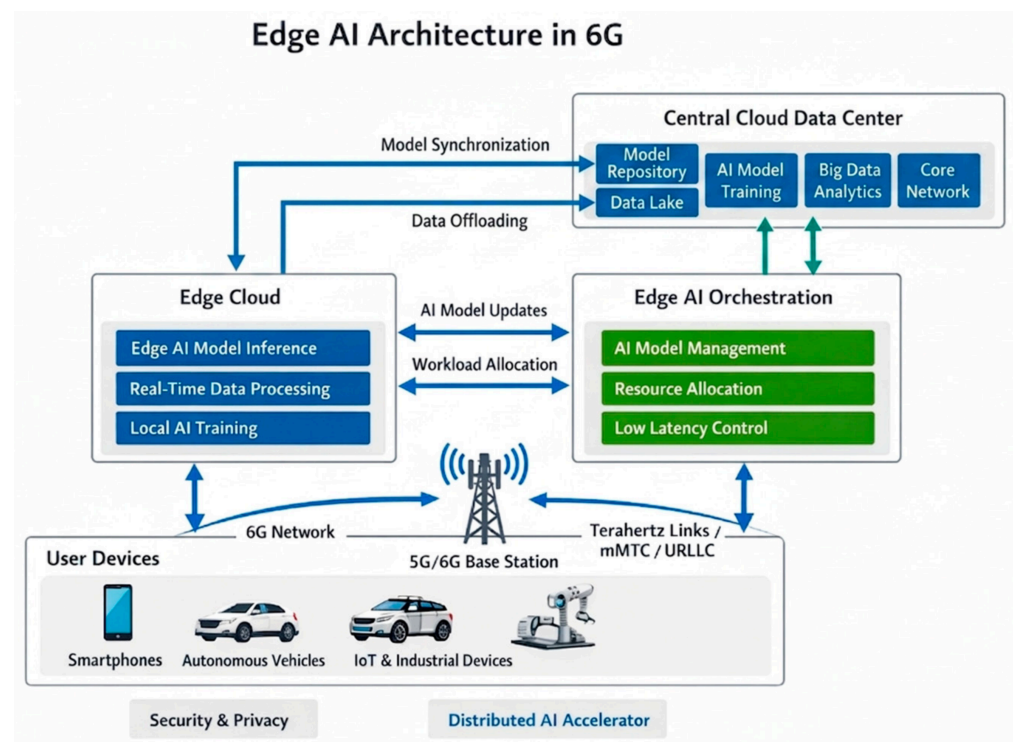
Table 2. Comparison of Computing Paradigms: Cloud, Fog and Edge.

Characteristic	Cloud Computing	Fog Computing	Edge Computing (MEC)
Location of Compute/Storage	Centralized, remote data centres	Distributed nodes between cloud and edge	At/near the network edge (e.g., RAN, cellular sites)
Typical Latency	High (tens to hundreds of ms) [38]	Medium/low (ms to tens of ms) [56]	Ultra-low (sub-ms to a few ms) [5]
Bandwidth (Backhaul)	High consumption [57]	Reduced consumption compared with cloud) [56]	Minimal/optimized consumption [58]
Management/Orchestration	Centralized, mature	Distributed, more complex	Highly distributed, complex, requires automation [59]
Scalability	Very high (virtually unlimited resources)	Moderate	High (distributed), but limited by local resources
Main Limitations	Latency, bandwidth, privacy [57]	Per-node resource constraints, complexity [56]	Very limited per-node resources, mobility, security [59]
Typical 6G Use Cases	Large-scale offline processing, storage	Industrial IoT, local video analytics [60]	XR, V2X, robotics, gaming, real-time AI [5]

Note: Evaluation scores (1–10 scale) reflect suitability for typical 6G use cases requiring ultra-low latency (<10 ms), high privacy, and context awareness. Cloud scores lower despite computational advantages due to latency constraints. Edge scores highest despite deployment complexity due to alignment with 6G requirements.

3. Edge AI: Artificial Intelligence Integration at the 6G Network Edge

The convergence of artificial intelligence with edge computing infrastructure is one of the most significant conceptual and technological pillars of 6G. Edge AI is not merely an application running on the edge network; it is an intrinsic capability that redefines both network operation and the services it can deliver, as is illustrated in Figure 4.

**Figure 4.** Possible Edge AI Architecture in 6G.

3.1. Concept and Relevance of Edge AI in 6G

Edge AI refers to the deployment and execution of artificial intelligence algorithms, including both training and inference phases, directly on edge devices (such as smartphones, sensors or vehicles) or on infrastructure nodes located close to the data source (such as MEC servers) [17]. This approach represents a paradigm shift compared with traditional cloud-centred AI.

Its relevance to 6G is fundamental. The 6G vision of “connected intelligence” [15] or “AI-native networks” [5] implies that AI must permeate all layers and domains of the network [19]. Edge AI is the disruptive technology that enables this deep integration, fostering a synergy among communication, computing, sensing and intelligence [6]. Pervasive AI is considered a key candidate for managing the complexity of 6G networks, enabling dynamic resource allocation, adaptive traffic flow management and advanced signal processing in interference-rich environments [61].

The key advantages of edge AI over cloud-centered AI are particularly significant in the 6G context [5]:

- Lower Latency: Local execution eliminates communication delays with the cloud, which is crucial for real-time 6G applications.
- Reduced Bandwidth Consumption: Large volumes of raw data do not need to be transmitted to the cloud, as they are processed locally.
- Enhanced Privacy and Security: Sensitive data remain on the local device or edge node, reducing exposure and attack surface.
- Context Awareness: AI models can exploit real-time, locally available information to make more adaptive and context-specific decisions.
- Offline Operation: Intelligent services can function even without a continuous connection to the cloud.

3.2. Key Models and Techniques for Edge AI in 6G

Given the resource constraints at the edge, not all AI models and techniques are directly applicable. Specific approaches have therefore been developed and adapted for edge AI, with the most relevant for 6G being the following:

3.2.1. Federated Learning (FL)

The following mathematical framework is derived from and extends the canonical FL formulation of McMahan et al. [62], with adaptations for the wireless 6G channel environment.

Federated Learning is a distributed learning framework in which multiple clients (edge devices) collaboratively train a global model coordinated by a central server, without sharing their private local data [63]. Only model updates (e.g., gradients or parameters) are exchanged [18].

Mathematical Framework: The FL optimization problem can be formally stated as minimizing the global objective function:

$$\min_{w \in \mathbb{R}^d} C(w) = \sum_{i=1}^N \frac{n_i}{n} L_i(w) \quad (16)$$

where w represents the global model parameters, N is the number of participating clients, n_i is the number of data samples at client i , $n = \sum_{i=1}^N n_i$ is the total number of

samples, and $L_i(w)$ is the local loss function at client i computed over its private dataset $\mathcal{D}_i$. Each local loss is typically defined as:

$$L_i(w) = \frac{1}{n_i} \sum_{(x,y) \in \mathcal{D}_i} \ell(f(x;w), y) \quad (17)$$

where $\ell(\cdot, \cdot)$ is a sample-wise loss function (e.g., cross-entropy for classification) and $f(x;w)$ is the model's prediction.

θ will be used for model parameters in the FedAvg algorithm to avoid notational conflict with beamforming vectors f_k , introduced later.

Federated Averaging (FedAvg) Algorithm: The canonical FL algorithm, FedAvg, operates in iterative communication rounds $t = 1, 2, \dots, T$. At each round: The FL optimization framework presented in Equations (20)–(23) follows the FedAvg formulation of McMahan et al. [62] with convergence bounds from Kairouz et al. [20], extended with communication efficiency terms from [18] and large model adaptations from [63].

- Server broadcasts the global model $\theta^{(t)}$ to a subset $\mathcal{S}^{(t)} \subseteq \{1, \dots, N\}$ of clients.
- Local training: Each selected client $i \in \mathcal{S}^{(t)}$ performs E local epochs of SGD on its local data:

$$\theta_i^{(t+1)} = \theta_i^{(t)} - \eta \nabla \mathcal{L}_i(\theta_i^{(t)}) \quad (18)$$

where η is the local learning rate.

- Aggregation: The server aggregates the updated local models:

$$\theta^{(t+1)} = \sum_{i \in \mathcal{S}^{(t)}} \frac{n_i}{\sum_{j \in \mathcal{S}^{(t)}} n_j} \theta_i^{(t+1)} \quad (19)$$

where θ denotes model parameters (to avoid conflict with beamforming vectors).

Convergence Analysis: Under convexity assumptions, FedAvg achieves a convergence rate of $\mathcal{O}(1/\sqrt{T})$ for convex objectives and $\mathcal{O}(T^{-1/3})$ for non-convex objectives, where T is the number of communication rounds. The convergence bound incorporates client drift due to heterogeneous data:

$$\mathbb{E}[C(w^{(T)})] - C(w^*) \leq \mathcal{O}\left(\frac{1}{\sqrt{T}} + \frac{E\eta^2\mathbb{V}}{T}\right) \quad (20)$$

where:

- $\mathcal{L}$: global loss function.
- T : total communication rounds.
- E : local epochs per round.
- η : learning rate.
- $\sigma^2 = \frac{1}{N} \sum_{i=1}^N \|\nabla \mathcal{L}_i(\theta^*) - \nabla \mathcal{L}(\theta^*)\|^2$: gradient variance capturing data heterogeneity.

Under IID data distribution ($\sigma^2 = 0$), convergence reduces to $\mathcal{O}(1/\sqrt{T})$, matching standard SGD rates [20].

Differential Privacy Integration: To provide formal privacy guarantees, FL can be enhanced with differential privacy mechanisms. A standard two-step approach first clips and then perturbs local gradients before aggregation:

$$\tilde{g}_i = \text{clip}(g_i, S) + \mathcal{N}(0, \sigma_{DP}^2 S^2 I) \quad (21)$$

$$\theta^{(t+1)} = \theta^{(t)} - \eta \tilde{g}_i \quad (22)$$

where $g_i = \nabla \mathcal{L}_i(\theta^{(t)})$ is the local gradient, $\text{clip}(g_i, S) = g_i \cdot \min\left(1, \frac{S}{\|g_i\|_2}\right)$ is gradient clipping with sensitivity bound $S > 0$, and $\sigma_{\text{DP}} > 0$ is the noise multiplier controlling the privacy–utility trade-off. This provides (ϵ, δ) -differential privacy, where the privacy budget ϵ grows with composition across rounds according to:

$$\epsilon_{\text{total}} \leq \sqrt{2T \ln\left(\frac{1}{\delta}\right)} \cdot \epsilon + T\epsilon \cdot (e^\epsilon - 1) \quad (23)$$

under basic composition, or tighter bounds under advanced composition theorems. This quantifies the fundamental trade-off between privacy (lower ϵ) and model accuracy.

Impact of Wireless Channel on FL Convergence

A critical dimension often overlooked in theoretical FL analyses is the effect of the wireless channel on gradient aggregation [64,65]. In practical 6G deployments, gradient updates are transmitted over time-varying fading channels, introducing three main impairments: (i) gradient errors due to channel noise that corrupt the aggregated model $\theta^{(t+1)}$; (ii) stale gradients caused by variable transmission delays under fading; and (iii) partial participation due to deep fades preventing some clients from completing a round, exacerbating data heterogeneity effects.

Over-the-Air Computation (AirComp) exploits the superposition property of the wireless multiple-access channel: all clients transmit simultaneously, and the aggregation server receives the naturally summed signal [66,67]:

$$\hat{\theta}^{(t+1)} = \frac{1}{N} \sum_{i=1}^N \theta_i^{(t+1)} + n_{\text{agg}} \quad (24)$$

where n_{agg} represents aggregation noise (thermal noise plus channel estimation error). This reduces spectrum usage by a factor of N compared to orthogonal multiple access. Update-aware device scheduling can further improve convergence by prioritizing clients with the most informative gradients [68]. The convergence bound for AirComp-FL incorporates an additional noise term:

$$\mathbb{E}[\mathcal{L}(\theta^{(T)})] - \mathcal{L}(\theta^*) \leq \mathcal{O}\left(\frac{1}{\sqrt{T}} + \frac{E\eta^2\sigma^2}{T} + \frac{\sigma_{\text{agg}}^2}{T}\right) \quad (25)$$

where σ_{agg}^2 is the aggregation noise variance, quantifying the fundamental trade-off between spectral efficiency and learning accuracy in wireless FL [66].

The aggregation noise variance σ_{agg}^2 depends directly on the wireless channel conditions. Under a flat-fading multiple-access channel with N clients, each transmitting with power P_{tx} over a channel with average gain $\overline{|h|}^2$ and noise power spectral density N_0 , the variance scales as:

$$\sigma_{\text{agg}}^2 \propto \frac{N_0}{N \cdot \overline{|h|}^2 \cdot P_{\text{tx}}} = \frac{1}{N \cdot \bar{\gamma}} \quad (26)$$

where $\bar{\gamma} = \overline{|h|}^2 P_{\text{tx}} / N_0$ is the average SNR per client link. Under Rayleigh fading, σ_{agg}^2 increases significantly at low SNR (e.g., cell-edge or NTN links with $\bar{\gamma} < 10$ dB) compared to dense-urban MEC links ($\bar{\gamma} > 25$ dB), directly slowing FL convergence. High-mobility Doppler spreading further degrades $\bar{\gamma}$ and introduces inter-carrier interference, increasing σ_{agg}^2 beyond the purely thermal noise floor. This coupling between wireless channel quality and learning convergence rate is a critical co-design consideration for 6G deployments [64,66,67,69].

Communication Complexity: The communication cost per round for client i is $\mathcal{O}(d)$ where d is the model dimension (number of parameters). For large models (e.g., $d \sim 10^6$

to 10^9 for deep neural networks), this becomes prohibitive in bandwidth-constrained 6G edge scenarios. Compression techniques (gradient sparsification, quantization) can reduce this to $\mathcal{O}(k)$ where $k \ll d$, at the cost of slower convergence.

Scalability of FL Gains Under Device Heterogeneity and Non-IID Data: The communication overhead reduction of up to 99% [70–75] via top-k gradient sparsification assumes idealized conditions (IID data, homogeneous devices, stable channels). Three factors significantly affect these gains in ultra-dense 6G deployments: (1) Non-IID Data Distribution: When client data follows heterogeneous distributions, the gradient dissimilarity increases (quantified by variance sigma squared in the convergence bound), causing client drift. Under highly non-IID conditions, more rounds are needed to converge, reducing net communication savings to approximately 90–95% (estimated from convergence analysis under non-IID conditions; see [76,77]). Algorithms such as FedProx and SCAFFOLD partially mitigate this. (2) Straggler Effect: In deployments with thousands of heterogeneous devices—from smartphones to IoT sensors—the slowest participants delay each aggregation round. Asynchronous FL variants and partial participation strategies (selecting only the fastest fraction of devices per round) mitigate straggler effects but may introduce gradient staleness. (3) Device Heterogeneity: Differences in computational capacity, battery levels, and memory require personalized FL approaches (per-device models or heterogeneous model architectures) that increase system complexity. Collectively, the 99% reduction represents an upper bound under favourable conditions; practitioners should expect 90–99% reduction depending on deployment scenario (range derived from results in [70–75] and theoretical bounds in [76,77]).

Non-IID Data Impact: Statistical heterogeneity arises when $\mathcal{D}_i$ are drawn from different distributions $P_i(x, y)$. This causes client drift, where local models diverge from the global optimum. The convergence degradation can be bounded by the gradient dissimilarity:

$$\Delta_{\text{drift}} \leq \frac{E^2 \eta^2}{2} \cdot \frac{1}{N} \sum_{i=1}^N \|\nabla L_i(w) - \nabla C(w)\|^2 \quad (27)$$

This demonstrates that non-IID data directly slows convergence proportional to the variance of local gradients.

Benefits: Its main advantage is the preservation of local data privacy [13], formally guaranteed through differential privacy when properly configured. It also reduces the need to transmit large volumes of data to a central server (saving bandwidth proportional to $|\mathcal{D}_i|/d$ per client) and enables model training with distributed datasets that cannot be centralized for legal or privacy-related reasons [18]. It is considered ideal for the distributed and privacy-sensitive environment of 6G [78].

Current Deployment Status of Federated Learning: FL is no longer purely theoretical. Google has deployed FL in production for training predictive text models on Android keyboards (Gboard) since 2017, serving billions of devices while preserving user privacy. Apple uses FL for Siri personalization and emoji suggestions on iOS devices. In healthcare, multiple hospital networks in the US and EU are piloting FL for medical image analysis under HIPAA/GDPR compliance. These real-world deployments validate the core FL principles described above and demonstrate their scalability to millions of heterogeneous edge devices, providing confidence in FL's applicability to the 6G edge ecosystem.

Challenges: FL faces several challenges, such as statistical heterogeneity (non-IID data across clients, quantified by gradient variance V above), system heterogeneity (differences in computing capacity, connectivity and availability, leading to stragglers), communication overhead (transmitting updates of large models can be costly, with complexity $\mathcal{O}(dT)$ over T rounds), scalability to large numbers of devices (aggregation complex-

ity $\mathcal{O}(N)$, and security issues (susceptibility to poisoning attacks or inference over shared updates). Efficient and secure aggregation is also a key difficulty [14].

The development of FL is not coincidental but a direct consequence of operating at the edge. The impossibility of centralizing massive and private datasets, together with bandwidth limitations [21], made traditional centralized approaches impractical. FL emerged as an architectural and algorithmic solution specifically designed to overcome these constraints imposed by the distributed and privacy-restricted edge environment [18].

FL's Specific Contribution to 6G Performance Improvement: FL enables collaborative model training without transmitting raw data, directly reducing backhaul load by a ratio proportional to gradient size versus raw data size (typically 100:1 to 10,000:1 with gradient compression) [79,80]. This directly addresses 6G's requirement for efficient use of scarce fronthaul and backhaul capacity. By keeping the training data at edge devices, FL satisfies 6G's stringent privacy requirements (GDPR, sector-specific healthcare and financial regulations) and enables AI-native network optimization without centralizing sensitive network telemetry. In the 6G context, FL enables distributed optimization of spectrum and power allocation, predictive beam management, and user mobility prediction—all while adapting to the heterogeneous device capabilities and the dynamic network conditions inherent to ultra-dense 6G deployments. The reduction in backhaul traffic directly translates to improved network responsiveness and reduced infrastructure costs.

3.2.2. Split Learning (SL)

The following split learning formulation is based on Lin et al. [78], with extensions for split-point optimization under 6G resource constraints. The mathematical framework for semantic communication and split learning presented in this section synthesizes the formulations from Zhang et al. [59] (Section 3), Zhu et al. [63], and Lin et al. [78], adapted to the 6G edge computing context analyzed in this work.

In Split Learning, a neural network model is divided into two (or more) parts [63]. An initial portion (typically smaller) runs on the client device, processing raw data and generating an intermediate representation (activations or “smash data”). This representation is transmitted to an edge server, which runs the remaining (typically larger and computationally intensive) part of the model to complete the forward inference and/or backward training pass [59].

Mathematical Model: Consider a deep neural network

$$f(x; w) = f_{\text{server}}(f_{\text{client}}(x; w_c); w_s) \quad (28)$$

split at layer k , where the following hold:

- $f_{\text{client}}(x; w_c)$ represents layers 1 to k executed on the client with parameter w_c .
- $f_{\text{server}}(a_k; w_s)$ represents layers $k + 1$ to L executed on the server with parameter w_s .
- $a_k = f_{\text{client}}(x; w_c)$ is the intermediate activation (“smash data”) transmitted to the server.

Forward Pass:

1. Client computes:

$$a_k = f_{\text{client}}(x; w_c) = \sigma_k(W_k \cdots \sigma_1(W_1 x)) \quad (29)$$

2. Client transmits: $a_k \in \mathbb{R}^{d_k}$ to server.

3. Server computes:

$$\hat{y} = f_{\text{server}}(a_k; w_s) = \sigma_L(W_L \cdots \sigma_{k+1}(W_{k+1} a_k)) \quad (30)$$

4. Server computes loss: $\mathcal{L}(\hat{y}, y)$.

Backward Pass (Training):

5. Server computes gradients: $\frac{\partial \mathcal{L}}{\partial w_s}$ and $\frac{\partial \mathcal{L}}{\partial a_k}$.
6. Server transmits: $\frac{\partial \mathcal{L}}{\partial a_k} \in \mathbb{R}^{d_k}$ back to client.
7. Client computes: $\frac{\partial \mathcal{L}}{\partial w_c}$ via chain rule.
8. Both update their respective parameters: $w_c \leftarrow w_c - \eta \frac{\partial \mathcal{L}}{\partial w_c}$, $w_s \leftarrow w_s - \eta \frac{\partial \mathcal{L}}{\partial w_s}$.

Split Point Optimization Problem: The optimal split point k^* minimizes total latency subject to client resource constraints:

$$k^* = \arg \min_{k \in 1, \dots, L-1} T_{\text{total}}(k) \quad (31)$$

where:

$$T_{\text{total}}(k) = \underbrace{T_{\text{comp}}^{\leq k}}_{\text{client layers } 1-k} + \underbrace{T_{\text{comm}}(k)}_{\text{smashed data transfer}} + \underbrace{T_{\text{comp}}^{>k}}_{\text{server layers } k+1-L} \quad (32)$$

with the following:

- $T_{\text{comp}}^{\leq k} = \sum_{l=1}^k \frac{FLOPs_l}{f_{\text{client}}}$: client computation time for layers 1 to k .
- $T_{\text{comm}}(k) = \frac{|a_k|}{R_{\text{uplink}}} + \frac{|\delta_k|}{R_{\text{downlink}}}$: communication latency for activations a_k and back-propagated gradients δ_k .
- $T_{\text{comp}}^{>k} = \sum_{l=k+1}^L \frac{FLOPs_l}{f_{\text{server}}}$: server computation time for layers $k+1$ to L .

subject to:

$$C_{\text{client}}(k) \leq C_{\text{max}}^{\text{client}}, M_{\text{client}}(k) \leq M_{\text{max}}^{\text{client}} \quad (33)$$

where $C_{\text{client}}(k)$ and $M_{\text{client}}(k)$ are the client's computational and memory requirements. This optimization is non-convex and depends on network architecture, device capabilities, and channel conditions.

Communication Complexity: Unlike FL where full model updates (dimension $d = \sum_{i=1}^L d_i$) are transmitted, SL transmits intermediate activations and gradients of dimension d_k . For a batch of size B , the communication cost per iteration is:

$$\text{CommCost}_{\text{SL}} = 2 \cdot B \cdot d_k \cdot b \text{ bits} \quad (34)$$

versus FL's:

$$\text{CommCost}_{\text{FL}} = 2 \cdot d \cdot b \text{ bits (per client)} \quad (35)$$

where $d_k \ll d$ (early split) or $B \ll 1$, SL can significantly reduce communication.

Privacy Analysis: While raw data x remains on the client, the transmitted activation a_k may leak information. The privacy leakage is quantified by the mutual information between input and activation. Since $a_k = f_{\text{client}}(x; w_c)$ is a deterministic function of x , the standard definition $I(X; Y) = H(Y) - H(Y|X)$ gives $H(a_k|x) = 0$ for deterministic encoders; therefore:

$$I(x; a_k) = H(a_k) \leq H(x) \quad (36)$$

Using the data processing inequality applied to the Markov chain $x \rightarrow a_k \rightarrow \hat{y}$:

$$I(x; \hat{y}) \leq I(x; a_k) \leq H(x) \quad (37)$$

The tighter the information bottleneck imposed by f_{client} (i.e., the more a_k compresses x), the lower the leakage [20]. For stochastic encoders that add noise $a_k \leftarrow a_k + \mathcal{N}(0, \sigma^2 I)$, the leakage decreases further, with the mutual information bounded by $I(x; a_k) \leq H(a_k)$ [20]. For deterministic networks, $a_k = f_{\text{client}}(x; w_c)$ is a deterministic function, and reconstruc-

tion attacks may recover x if f_{client} is invertible. Adding noise $a_k \leftarrow a_k + \mathcal{N}(0, \sigma^2 I)$ provides (ϵ, δ) -differential privacy:

$$\epsilon = \frac{\|a_k\|_2}{\sigma} \quad (38)$$

at the cost of reduced model accuracy.

Convergence: For strongly convex loss functions with Lipschitz gradients, SL converges at the following rate:

$$\mathbb{E} \left[\|\nabla \mathcal{L}(w^{(t)})\|^2 \right] \leq \mathcal{O} \left(\frac{1}{t} \right) + \mathcal{O}(\epsilon_{\text{noise}}^2) \quad (39)$$

where t is the iteration number and ϵ_{noise} quantifies privacy-preserving noise. The bidirectional communication per sample introduces latency proportional to Equation (39), which represents the standard over-the-air aggregation model derived from Cao et al. [81] and Amiri & Gündüz [66] (Section III).

$$T_{\text{iter}} = T_{\text{comp}}^{\text{client}} + 2 \cdot T_{\text{comm}} + T_{\text{comp}}^{\text{server}} + T_{\text{backprop}} \quad (40)$$

requiring τ iterations for convergence, yielding a total training time $\tau \cdot T_{\text{iter}}$.

Benefits: The main advantage of SL is the significant reduction in computational load on the client device compared with FL (client computes only $k \ll L$ layers), since most computation is performed on the server [78]. This makes it more suitable for devices with very limited resources (e.g., IoT sensors with $C_{\text{max}}^{\text{client}} \ll C_{\text{server}}$). SL also offers privacy benefits, as raw data never leave the client device, and intermediate activations a_k are generally less informative than full gradients or raw data [40], although formal privacy guarantees require additional mechanisms. This aligns well with the 6G vision of leveraging dispersed resources [78].

Challenges: SL introduces additional latency due to the bidirectional communication required per sample or batch during training (activations to the server; gradients back to the client), with total latency $T_{\text{total}}(k)$ dependent on split point k [59]. Deciding where to optimally split the model is a complex problem requiring the non-convex optimization above to be resolved. Moreover, although SL enhances privacy compared with full centralization, exchanged activations and gradients may still leak information (bounded by mutual information $I(x; a_k)$), requiring additional protection mechanisms such as noise injection [59].

As with FL, SL is a response to edge limitations. When FL becomes computationally infeasible for extremely resource-constrained devices, SL offers an alternative by offloading most computation to the server, directly addressing the client-side resource bottleneck [78].

SL's Specific Contribution to 6G Performance Improvement: Split Learning partitions the neural network across client devices (early layers) and edge servers (later layers), reducing on-device computation to only the initial feature extraction layers while offloading computationally intensive layers to the server. This enables AI inference and training on the extremely resource-constrained and diverse device ecosystem characteristic of 6G—from IoT sensors with microcontrollers to industrial robots with embedded processors—without requiring full model deployment on the device. In 6G contexts, SL enables real-time AI-assisted sensing (processing raw sensor data locally, transmitting compact intermediate representations to edge servers for complex inference); context-aware service personalization at reduced device energy cost; and support for AI-native air-interface optimization on devices that cannot locally execute full inference models. The split-point optimization (Equation (32)) allows for dynamic adaptation to device resource availability and channel

conditions, making SL a natural fit for 6G's diverse and heterogeneous device ecosystem, where device capabilities span several orders of magnitude.

A comparative summary table spanning all panels lists five metrics: Communication Overhead, Privacy Risk, Server Complexity, Training Latency, and Bandwidth Requirements.

Figure 5 contrasts the data flow, communication overhead, and privacy implications of centralized learning, FL, and SL.

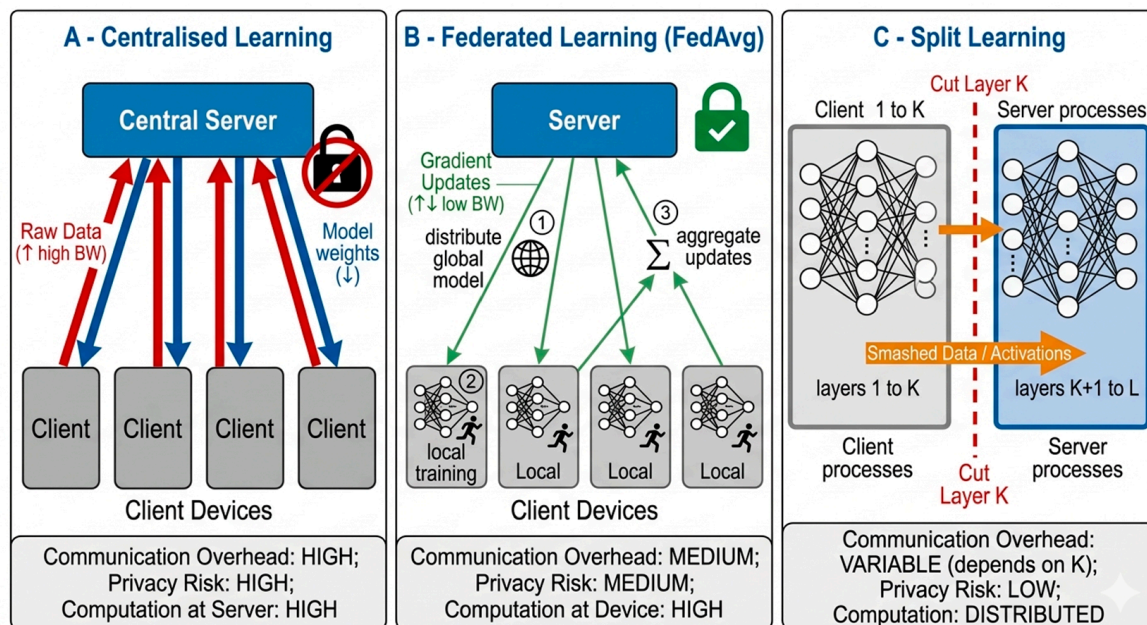


Figure 5. Comparison of Centralized Learning, Federated Learning, and Split Learning.

3.2.3. Edge Large AI Models (Edge LAMs)

The following frameworks for edge LAM deployment, PEFT, and quantization are based on foundational works including Hu et al. [82] for LoRA and the edge LAM [14] (Note: Reference [14] is a Huawei technical document available at <https://www.huawei.com>; while not peer-reviewed, it provides industry-relevant architectural details for AI-native 6G data plane design), with novel formulations for collaborative inference and parameter caching.

Important Distinction: Large AI Models (LAMs) vs. Large Language Models (LLMs). LLMs (e.g., GPT, BERT, LLaMA) are a specialized subset of large-scale AI focused exclusively on natural language processing tasks—text generation, translation, question answering, and code synthesis. LAMs, by contrast, are broader multimodal AI systems designed to process and reason over diverse data modalities, including text, images, audio, sensor readings, network telemetry, and control signals. In the 6G and edge AI context, LAMs extend beyond language to encompass multimodal perception, cross-domain reasoning, and autonomous decision-making for network operations such as intelligent orchestration, air-interface optimization, and context-aware service provisioning. While an LLM may serve as an intent interpretation engine in Intent-Based Networking (IBN), a full edge LAM integrates sensing data, network state, and user context to perform end-to-end network management across heterogeneous 6G environments. This distinction is critical for understanding the scope and applicability of large-scale AI models in 6G.

Edge LAMs refer to the adaptation, deployment and execution of large pre-trained AI models (PFMs), such as GPT-type large language models or multimodal/vision models, within the 6G edge infrastructure [19].

Potential: These models offer unprecedented capabilities in generalization, knowledge transfer and handling complex, diverse tasks (cross-modal reasoning, few-/zero-shot learning) [19]. They have enormous potential to revolutionize 6G network management (intelligent orchestration, semantic air-interface optimization) and enable highly personalized and interactive services (conversational agents, advanced virtual assistants) [19].

Challenges: The main obstacle is the stark mismatch between the enormous requirements of LAMs (billions of parameters, terabytes of training data, massive computation) and the severely limited resources available at the edge (compute, memory, storage, energy, bandwidth) [19]. Traditional edge AI approaches such as FL or SL may be insufficient or infeasible for full LAMs [14].

This tension is driving research into extreme optimization techniques, such as the following:

- **Decomposition and Distributed Deployment:** Dividing the LAM into smaller modules and distributing them across devices, edge nodes and the cloud [14]. For a model with L layers and total parameters $\theta = \{\theta_1, \dots, \theta_L\}$, the decomposition problem seeks a partition $\pi : \{1, \dots, L\} \rightarrow \{\text{device, edge, cloud}\}$ that minimizes total latency:

$$\pi^* = \underset{\pi}{\operatorname{argmin}} \left[\sum_{i=1}^L T_{\text{comp}}^{\pi(i)}(\theta_i) + \sum_{(i,j): \pi(i) \neq \pi(j)} T_{\text{comm}}(a_i, \pi(i), \pi(j)) \right] \quad (41)$$

subject to resource constraints:

$$\sum_{i: \pi(i)=r} \text{Mem}(\theta_i) \leq M_r, \quad \forall r \in \{\text{device, edge, cloud}\} \quad (42)$$

Here, $T_{\text{comp}}^r(\theta_i)$ is the computation time for layer i on resource r , $T_{\text{comm}}(a_i, r_1, r_2)$ is the communication time for transmitting activation a_i between resources, and M_r is the memory capacity at resource r .

- **Efficient Fine-Tuning:** Techniques such as Parameter-Efficient Fine-Tuning (PEFT), which update only a small fraction of a LAM's parameters [14], as well as distributed adaptations such as FedFT or Split FedFT [63].

PEFT Mathematical Framework: For a pre-trained model with parameters $w_0 \in \mathbb{R}^d$ (where $d \sim 10^9$ for large models), PEFT methods (e.g., LoRA, adapters, prompt tuning) introduce a low-rank adaptation:

$$w = w_0 + \Delta w \quad (43)$$

where Δw has significantly fewer trainable parameters. For Low-Rank Adaptation (LoRA):

$$\Delta W = BA, B \in \mathbb{R}^{d_{\text{out}} \times r}, A \in \mathbb{R}^{r \times d_{\text{in}}}, r \ll \min(d_{\text{out}}, d_{\text{in}}) \quad (44)$$

where $W \in \mathbb{R}^{d_{\text{out}} \times d_{\text{in}}}$ is the original weight matrix, r is the low-rank dimension, and B is initialized to zero and A with a Gaussian distribution so that $\Delta W = 0$ at the start of fine-tuning [82]. The parameter efficiency ratio is:

$$\rho_{\text{PEFT}} = \frac{|\text{Trainable Parameters}|}{|\text{Total Parameters}|} = \frac{(d_{\text{out}} + d_{\text{in}}) \cdot r}{d_{\text{out}} \cdot d_{\text{in}}} \ll 1 \quad (45)$$

For example, with $d = 10^9$ and $r = 8$, only $\rho_{\text{PEFT}} = 16/10^9 \approx 0.0000016\%$ of parameters are fine-tuned, reducing memory requirements from $d \cdot 4$ bytes (float32) to $2rd \cdot 4$ bytes, enabling fine-tuning on resource-constrained edge nodes.

Federated PEFT: In distributed settings, each client i trains local PEFT adapters $\Delta W_i = B_i A_i$ on its data, and the server aggregates:

$$\Delta W^{\text{global}} = \frac{1}{N} \sum_{i=1}^N \Delta W_i \quad (46)$$

This reduces communication from $\mathcal{O}(d)$ in standard FL to $\mathcal{O}(2rd)$ in FedFT, achieving a compression ratio of $d/(2rd) = 1/(2r)$.

- Compression and Quantization: Reducing model size and computational precision to lower memory and compute requirements [40].

Quantization Mathematical Framework: Model quantization reduces numerical precision from floating-point (e.g., FP32, 32 bits) to lower bit-widths (e.g., INT8, 8 bits or even 4/2/1 bits). For weight $w \in R$, the quantized value is as follows:

For n -bit quantisation (e.g., $n = 8$ for INT8), the quantised value of weight w is

$$Q_n(w) = \text{round}\left(\frac{w - z}{s}\right) \cdot s + z \quad (47)$$

where n denotes the number of quantization bits, $s > 0$ is the scale factor, and z is the zero-point offset. The quantization error for a parameter matrix $W \in R^{m \times n}$ can be bounded by

$$\epsilon_q = \frac{\|W - Q_n(W)\|_F}{\|W\|_F} \leq \frac{\sqrt{mn} \cdot s}{2 \|W\|_F} \quad (48)$$

where $\|\cdot\|_F$ is the Frobenius norm. For INT8 quantization, $s \approx \frac{\max(W) - \min(W)}{2^n - 1}$, yielding

$$\epsilon_q \approx \frac{\sqrt{mn}(\max(W) - \min(W))}{2^n \|W\|_F} \quad (49)$$

Memory and Computation Reduction: Quantization from FP32 ($n = 32$) to INT8 ($n = 8$) reduces:

- Memory: $4\times$ reduction (from 4 bytes to 1 byte per parameter).
- Energy: $\sim 30\times$ reduction in energy per operation [19] (INT8 vs. FP32 energy ratio measured on NVIDIA hardware: $\sim 27\text{--}35\times$ [19,83–85]) (energy-efficiency measurements derived from NVIDIA Jetson nano INT8 vs. FP32 throughput benchmarks).
- Model size: For a model with $d = 10^9$ parameters, from 4 GB to 1 GB.

Accuracy–Efficiency Trade-off: The accuracy degradation can be empirically modelled as

$$\Delta_{\text{acc}} = \text{Acc}_{\text{FP32}} - \text{Acc}_{Q_n} = \alpha \cdot \epsilon_q + \beta \cdot n^{-\gamma} \quad (50)$$

where α , β , γ are model-dependent constants. Empirical studies show that for well-calibrated quantization, INT8 typically maintains $\Delta_{\text{acc}} < 1\%$ [19,63], while INT4 may incur $\Delta_{\text{acc}} \approx 2\text{--}5\%$.

Post-Training Quantization vs. Quantization-Aware Training:

- PTQ: Direct quantization of pre-trained weights, fast but potentially higher error.
- QAT: Include quantization in training loop, minimizing: $\min_w E[l(f(x; Q_n(w)), y)]$, better accuracy but requires retraining.

Collaborative and Efficient Inference: Architectures in which multiple nodes collaborate to perform inference by sharing parameters or intermediate outputs [14], while also avoiding redundant computation in microservice-based deployments [14].

Collaborative Inference Model: Consider K edge nodes collaboratively performing inference for a model with L layers. Each node k hosts a subset of layers $\mathcal{L}_k \subseteq \{1, \dots, L\}$ with parameters θ_k . For input x , the inference process is

$$y = f_L(\dots f_2(f_1(x; \theta_1); \theta_2) \dots; \theta_L) \quad (51)$$

where layer i is computed on node k if $i \in \mathcal{L}_k$. The total inference latency is

$$T_{\text{inference}} = \sum_{i=1}^L T_{\text{comp}}^{k(i)}(\theta_i) + \sum_{i=1}^{L-1} \mathbf{1}[k(i) \neq k(i+1)] \cdot T_{\text{comm}}(a_i) \quad (52)$$

where $k(i)$ is the node hosting layer i , and $\mathbf{1}[\cdot]$ is the indicator function. Minimizing this latency requires optimal layer-to-node assignment, which is NP-hard in general.

Parameter Sharing and Caching: To reduce redundant computation when serving multiple concurrent requests, edge nodes can cache activations. For a sequence of inputs $\{x_1, \dots, x_m\}$ sharing common prefixes (e.g., prompt in LLMs), the cached activation approach computes shared prefix layers once:

$$a_k^{\text{shared}} = f_k(\dots f_1(x_{\text{prefix}}; \theta_1) \dots; \theta_k) \quad (53)$$

and reuses it for all requests, reducing computation from $\mathcal{O}(m \cdot \sum_{i=1}^k \text{FLOPs}_i)$ to $\mathcal{O}(\sum_{i=1}^k \text{FLOPs}_i + m \cdot \sum_{i=k+1}^L \text{FLOPs}_i)$.

The emergence of edge LAMs marks a critical inflexion point. The ambition to harness the transformative power of large AI models directly conflicts with the practical constraints of edge environments. Addressing this tension requires fundamental innovations not only in AI model efficiency but also in the design of edge-computing and network architectures, potentially redefining both fields.

Supporting Large-Scale AI Models at the Edge—Current State and Practical Viability: Model partitioning (split inference) is currently the most viable approach for running large models at the edge without compromising latency and energy-efficiency. Recent advances show that 1–7 billion parameter models (e.g., LLaMA-3-8B, Phi-3-mini) can be deployed on edge hardware with 4-bit or 8-bit quantization, reducing memory requirements to 4–8 GB—this is achievable on high-end edge servers and consumer GPUs. Key enabling techniques include: (1) **Model Compression:** quantization (INT8/INT4) achieves 4–8× memory reduction with less than 2% accuracy loss for well-calibrated models [86] (Section 4); structured pruning reduces FLOPs by 50–70% with 1–3% accuracy degradation [87]; knowledge distillation transfers capabilities from large teacher models to compact student models optimized for edge inference. (2) **Model Partitioning:** splitting a 7B-parameter model across device (early layers) and edge server (later layers) reduces device memory requirements to less than 2 GB while offloading computationally intensive transformer layers. (3) **Energy-Efficiency vs. Latency Trade-Off:** INT8 inference on an NVIDIA Jetson AGX Orin achieves approximately 5–10 tokens per second for a 7B model, consuming 15–30 W [88] feasible for non-real-time inference but insufficient for conversational applications. For real-time 6G network management tasks (latency less than 10 ms), only smaller models (less than 1B parameters) or specialized edge-adapted architectures (TinyLLM, MobileLLM) are currently feasible without model partitioning. SL provides a complementary approach by dynamically offloading the computation-intensive portion of inference to the edge server while keeping sensitive raw data on the device.

Power control optimization for AirComp under fading channels is essential to minimize aggregation distortion while meeting per-device transmit power constraints [81].

Channel-Aware Optimization within Federated Learning Frameworks: Explicitly incorporating wireless Channel State Information (CSI) into FL is critical for practical 6G deployments. Three complementary mechanisms enable channel-aware FL: (1) Joint scheduling of computation and radio resources: the FL aggregation server jointly selects the communication time slot and the subset of participating clients based on both local gradient informativeness and instantaneous channel quality, minimizing per-round latency under total bandwidth constraints. (2) CSI-based client selection: clients experiencing favourable channel conditions (SNR above a threshold) are preferentially selected for each aggregation round, reducing gradient aggregation noise and accelerating convergence as quantified by the AirComp bound in Equation (25). (3) Gradient compression adapted to channel quality: the sparsification ratio k is dynamically adjusted based on available channel capacity—under poor channel conditions, more aggressive compression reduces transmission overhead at the cost of slower convergence. Over-the-air computation (AirComp) represents the most bandwidth-efficient realization of channel-aware FL aggregation, reducing spectrum usage by a factor of N compared to orthogonal multiple access, though at the expense of aggregation noise that scales inversely with channel SNR. These channel-aware mechanisms are essential for meeting 6G's stringent latency and reliability requirements in realistic wireless environments.

Other relevant techniques include the following:

- **Tiny Machine Learning (TinyML):** Approaches designed to execute machine-learning models on devices with extremely limited resources, such as microcontrollers [13].
- **Over-the-Air Computation (AirComp):** A technique that exploits the superposition properties of the wireless channel to perform aggregations (such as those required in FL) directly over the air, thereby reducing latency and spectrum usage [63].
- **Split Inference:** Similar to SL but applied solely to the inference phase rather than training [89]. It is useful for accelerating the inference of complex models on resource-constrained devices.

3.2.4. AI Applications for 6G Network Optimization at the Edge

The mathematical formulations in this section (spectrum/power allocation, MIMO beamforming, handover optimization, anomaly/intrusion detection) are re-statements of standard optimization problems from the wireless communications and network security literature [4,17,90–92]; the novelty of this work lies in framing them within a unified edge AI context for 6G and in proposing deep reinforcement learning and deep learning approaches as solution methods.

Edge AI not only enables new end-user services but also serves as a powerful tool for optimizing the 6G network itself in an intelligent and adaptive manner:

- **Intelligent Resource Management:** AI can address the complex joint resource-optimization problems in 6G (spectrum, power and channel allocation, user association, beamforming, network slicing), many of which are NP-hard [90]. ML/DL/RL algorithms can learn from network data and make real-time decisions to maximize efficiency (spectral and energy), capacity, fairness and QoS, while adapting to dynamic environmental conditions [4].

Mathematical Formulation—Spectrum and Power Allocation:

(This formulation follows the standard resource allocation framework described in [17]. Section 4; [90], Section 4; [91] Section 3.) Consider a multi-cell 6G network with K base stations and N users. The joint spectrum and power allocation problem is:

$$\max_{p_{k,n}, b_{k,n}} \sum_{k=1}^K \sum_{n=1}^N w_n R_{k,n}(p_{k,n}, b_{k,n}) \quad (54)$$

subject to:

- Power constraint: $\sum_{n=1}^N p_{k,n} \leq P_k^{\max}, \quad \forall k.$
- Bandwidth constraint: $\sum_{n=1}^N b_{k,n} \leq B_k^{\max}, \quad \forall k.$
- QoS constraint: $R_{k,n} \geq R_n^{\min}, \quad \forall k, n$ where user is associated.
- Association constraint: $\sum_{k=1}^K x_{k,n} = 1, \quad \forall n$ (each user connects to one BS)

where:

- $p_{k,n}$: transmit power allocated to user n by base station k .
- $b_{k,n}$: bandwidth allocated to user n by base station k .
- $R_{k,n}$: achievable rate for user n from BS k .
- w_n : user priority weight.
- $x_{k,n} \in 0, 1$: binary association variable.

Achievable Rate (Shannon Capacity):

$$R_{k,n}(p_{k,n}, b_{k,n}) = b_{k,n} \log_2 \left(1 + \frac{p_{k,n} h_{k,n}}{\sum_{j \neq k} p_{j,n} h_{j,n} + N_0 b_{k,n}} \right) \quad (55)$$

where $h_{k,n}$ is the channel gain (including path loss, shadowing, fading) and N_0 is the noise power spectral density.

Complexity: This is a mixed-integer non-convex optimization problem, NP-hard even in simplified cases.

AI Solution—Deep Reinforcement Learning:

- State: $s_t = h_{k,n}(t), R_n^{\min}, p_k^{\text{avail}}, B_k^{\text{avail}}.$
- Action: $a_t = p_{k,n}, b_{k,n}$ (discretized or continuous).
- Reward: $r_t = \sum_{k,n} w_n R_{k,n} - \lambda_1 \sum_k \mathbf{1}[P_k \text{ violated}] - \lambda_2 \sum_n \mathbf{1}[R_n^{\min} \text{ violated}].$
- Policy network: $\pi_\theta(a|s)$ learned via PPO or SAC.
- Training: 10^5 to 10^6 episodes in simulation.
- Inference: $\mathcal{O}(KN \cdot d_{\text{NN}})$ operations, real-time capable (<1 ms).

Network Slicing Optimization:

For M network slices with heterogeneous requirements:

$$\min_{r_m^k} \sum_{k=1}^K \sum_{m=1}^M C_k^m(r_m^k) \quad (56)$$

subject to:

- Resource constraint: $\sum_{m=1}^M r_m^k \leq R_k^{\max}, \quad \forall k.$
- Slice requirements: $\sum_{k=1}^K r_m^k \geq R_m^{\text{demand}}, \quad \forall m.$
- Latency constraint: $L_m \leq L_m^{\max}, \quad \forall m.$
- Isolation: $\text{Isol}(m, m') \geq \text{Isol}_{\min}, \quad \forall m \neq m'.$

where r_m^k is the resource allocation for slice m at node k , and $C_k^m(\cdot)$ is the cost function.

AI Approach: Graph Neural Networks (GNNs) learn optimal slice-to-resource mapping by encoding network topology and slice dependencies.

- Air-Interface Optimization: AI can significantly enhance physical-layer performance. Examples include accurate channel state information (CSI) prediction, intelligent beam management in massive MIMO systems, and the development of adaptive modulation and coding schemes [1]. An emerging area is semantic communication, where AI extracts and transmits only the relevant (semantic) information rather than raw bits, thereby improving efficiency and robustness [91]. Edge LAMs are also proposed for intelligent air-interface design and optimization [14].

Mathematical Formulation—Massive MIMO Beamforming:

(Based on standard massive MIMO precoding formulations; see [91], Section 2; [93], Section 3.) For a massive MIMO system with N_t transmit antennas and K users: Equation (57) represents the received signal model for a massive MIMO system, where the beamforming vector optimizes the signal-to-interference-plus-noise ratio (SINR) across K users served by an N -antenna base station.

$$\max_{w_k} \sum_{k=1}^K \log_2 \left(1 + \frac{|h_k^H w_k|^2}{\sum_{j \neq k} |h_k^H w_j|^2 + \sigma^2} \right) \quad (57)$$

subject to $\sum_{k=1}^K \|f_k\|^2 \leq P_{\max}$, where $f_k \in \mathbb{C}^{N_t}$ is the beamforming vector for user k , $h_k \in \mathbb{C}^{N_t}$ is the downlink channel vector, and σ_n^2 is the noise variance.

Throughout this article, θ denotes neural-network model parameters, while f_k denotes beamforming/precoding vectors, to avoid notational ambiguity.

AI Solution—Deep Learning for CSI Prediction:

Channel prediction model learns the mapping $h_{t+\Delta} = f_\theta(h_t, h_{t-1}, \dots, h_{t-T})$:

$$\min_{\theta} \mathbb{E} \left[\|h_{t+\Delta} - f_\theta(h_{t:t-T})\|^2 \right] \quad (58)$$

Common architectures: LSTM, GRU, or Transformer for temporal dynamics. Typical performance: prediction error < -20 dB SNR for $\Delta \leq 10$ ms [94,95] (deep learning-based channel prediction performance benchmarks).

Adaptive Modulation and Coding (AMC):

Select modulation and coding scheme (MCS) $m \in \{1, \dots, M\}$ to maximize throughput while maintaining target BLER $\leq \epsilon_{\text{BLER}}$:

$$m^* = \operatorname{argmax}_m R_m(\text{SNR}) \quad \text{s.t.} \quad \text{BLER}_m(\text{SNR}) \leq \epsilon_{\text{BLER}} \quad (59)$$

AI Approach: Deep Q-Network (DQN) learns optimal MCS selection:

- State: $s = [\text{SNR}, \text{Doppler}, \text{delay spread}, \text{buffer status}]$.
- Action: $a = m$ (MCS index).
- Reward: $r = \text{throughput} - \beta \cdot \mathbf{1}[\text{BLER} > \epsilon_{\text{BLER}}]$.
- Intelligent Mobility and Handover Management: AI algorithms can predict user mobility patterns and optimize handover decisions between cells or edge nodes, minimizing service interruptions and ensuring seamless QoE [92].

Mathematical Formulation—Handover Optimization:

Given user trajectory prediction $\hat{x}_u(t + \Delta)$ for time horizon Δ , optimize handover policy to minimize:

$$\min_{t_{\text{HO}}, k_{\text{target}}} \left[\alpha_1 N_{\text{HO}} + \alpha_2 T_{\text{interruption}} + \alpha_3 \sum_t \mathbf{1} \left[R_u(t) < R_u^{\min} \right] \right] \quad (60)$$

subject to:

- Handover trigger: execute when predicted SINR $< \theta_{\text{HO}}$.
- Target selection: $k_{\text{target}} = \operatorname{argmax}_k \mathbb{E}[\text{SINR}_k(t + \Delta)]$.

AI Solution—LSTM for Mobility Prediction:

$$\hat{x}_u(t + \Delta), \hat{v}_u(t + \Delta) = \text{LSTM}_\theta(x_u(t : t - T), v_u(t : t - T)) \quad (61)$$

Prediction accuracy: mean absolute error (MAE) < 10 m for $\Delta = 5$ s with well-trained models [92].

- Efficient Operation and Maintenance (O&M): AI enables advanced Self-Organizing Network (SON) capabilities, including proactive fault detection, predictive maintenance based on data analytics, self-configuration and self-optimization of network parameters, and autonomous fault recovery [92]. This reduces the need for human intervention and lowers operational costs.

Mathematical Formulation—Anomaly Detection:

Learn normal network behaviour distribution $P_{\text{normal}}(x)$ from historical data. Detect anomalies when:

$$\text{Anomaly}(x_t) = \begin{cases} 1, & \text{if } P_{\theta}(x_t) < \epsilon_{\text{threshold}} \\ 0, & \text{otherwise} \end{cases} \quad (62)$$

AI Approach—Autoencoder for Anomaly Detection:

Train autoencoder to reconstruct normal patterns:

$$\min_{\theta} \mathbb{E}_{x \sim P_{\text{normal}}} [\|x - \text{Dec}(\text{Enc}(x; \theta))\|^2] \quad (63)$$

Reconstruction error for anomalous inputs will be high: $|x_{\text{anomaly}} - \hat{x}|^2 > \tau$.

Predictive Maintenance:

Predict time-to-failure T_{fail} for network components:

$$T_{\text{fail}} = g_{\theta}(\text{sensor data, historical failures, operating conditions}) \quad (64)$$

Using survival analysis models (Cox proportional hazards) or deep learning (RNNs).

- AI-Enhanced Security: Edge AI can strengthen 6G network security through intelligent real-time intrusion and anomaly detection, behavioural analysis for threat identification and automated response mechanisms [96]. It can also be used to reinforce physical-layer security (PLS) approaches [89].

Mathematical Formulation—Intrusion Detection:

Binary classification problem for network traffic:

$$\hat{y} = \text{Classifier}_{\theta}(\text{traffic features}) \in \{\text{normal, intrusion}\} \quad (65)$$

Performance Metrics:

- True Positive Rate: $\text{TPR} = \frac{\text{TP}}{\text{TP} + \text{FN}}$ (detection rate).
- False Positive Rate: $\text{FPR} = \frac{\text{FP}}{\text{FP} + \text{TN}}$ (false alarm rate).
- Target: $\text{TPR} > 99\%$, $\text{FPR} < 0.1\%$.

AI Approach: Deep learning models (CNN, LSTM, Transformer) trained on labelled network traffic datasets (e.g., NSL-KDD, CICIDS) achieve state-of-the-art performance with F1-scores > 0.99 [93].

3.2.5. Benefits of Edge AI for 6G Services

The integration of AI at the edge enables a new generation of services and significantly enhances the user experience:

- **Enablement of Real-Time Intelligent Services:** Edge AI is essential for applications requiring low latency and on-site intelligent decision-making. This includes autonomous driving (environment perception, man oeuvre planning) [13], collaborative industrial robotics [8], truly interactive and personalized XR and metaverse experiences [13], advanced virtual assistants and context-aware personalized services [14].
- **Improved Efficiency (Energy, Spectrum):** Intelligent optimization of network resources (radio and computing) and reduced data traffic towards the cloud contribute to greater energy and spectral efficiency, both of which are critical for the sustainability of 6G [7].
- **Enhanced Quality of Experience (QoE):** By reducing latency, increasing reliability and enabling personalized services, edge AI directly improves the quality of experience for end users [5].
- **New Network Capabilities:** 6G may provide Artificial Intelligence as a Service (AIaaS) directly from the network [5]. Furthermore, synergy with integrated sensing opens the door to Integrated Sensing Edge Intelligence (ISEA), where the network not only communicates and computes but also intelligently perceives the environment [63].

3.2.6. Challenges of Edge AI in the 6G Environment

Despite its enormous potential, the effective deployment of edge AI in 6G faces several major challenges:

- **Computational Requirements vs. Limited Resources:** This remains the primary challenge. AI models, particularly advanced ones such as LAMs, require substantial computational resources (measured in FLOPs), memory and energy, often exceeding the capabilities of edge devices and servers [19]. A holistic optimization strategy is needed across data (cleaning, compression, augmentation), models (compression, quantization, pruning, efficient architectures such as NAS) and systems (specialized hardware such as NPUs/TPUs, efficient resource allocation) [14].
- **Data and Model Privacy and Security:** Although techniques such as FL and SL aim to preserve privacy, they are not immune to sophisticated attacks. There is a risk that sensitive information will be inferred from model updates (gradients) or intermediate activations [78]. Models may also be vulnerable to poisoning attacks (manipulated training data) or adversarial attacks (inputs crafted to deceive the model during inference) [89]. Ensuring privacy and security in distributed AI environments requires integrating advanced cryptographic techniques such as Differential Privacy (DP), Homomorphic Encryption (HE) or Secure Multiparty Computation (SMC) [60], together with robust access control and auditing mechanisms [13].
- **Communication Overhead:** Distributed training (FL/SL) and collaborative inference involve frequent exchanges of information (parameters, gradients, activations) between nodes, which can impose significant overhead on the wireless network, consuming bandwidth and energy [14]. Efficient model/update-compression techniques and optimized communication strategies (e.g., AirComp) are required [63].
- **Robustness and Generalization:** AI models must operate reliably in the 6G wireless environment, which is inherently dynamic, noisy and error-prone [14]. They must be robust to channel variations, mobility and interference. Moreover, especially for LAMs, they must generalize well across different tasks, scenarios and domains with minimal re-adaptation [92].

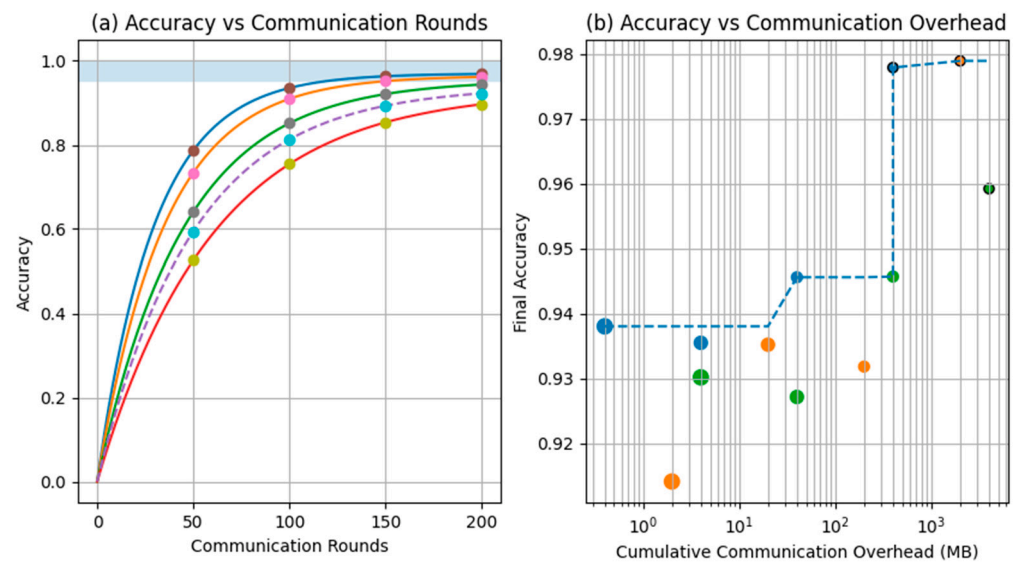
- **Data Collection and Quality:** AI critically depends on the availability of large volumes of high-quality training data (“Garbage in, Garbage out” [22]). Collecting, labelling and managing such data in a distributed and heterogeneous environment like the 6G edge poses both logistical and cost challenges [19]. Incomplete, noisy or biased data can lead to unreliable or unfair models [97]. The use of synthetic data generated by Digital Twins is a promising avenue to mitigate this issue [19].
- **Ethics, Transparency and Explainability:** As AI becomes responsible for increasingly critical decisions in 6G networks and services, ethical concerns intensify. These include potential algorithmic biases, lack of transparency in decision-making (the “black-box” problem) and the need for explainability, particularly in sensitive applications [57].

Managing AI in 6G will likely require a hierarchy and the coexistence of different types of models. While large foundational models (PFMs/LLMs) may reside in more capable nodes (edge servers or regional clouds) to perform complex tasks such as orchestration, semantic understanding or planning [19], smaller, specialized and efficient models (task-oriented AI, TinyML) will be deployed on end devices or highly resource-constrained edge nodes for specific, low-latency tasks [19]. The intelligent orchestration of this hierarchical and heterogeneous collaboration across the cloud–edge–device continuum will be a key challenge, but also an opportunity to leverage the strengths of each approach [19]. Table 3 summarizes and compares the key edge AI techniques discussed.

Table 3. Summary of Key Edge AI Techniques for 6G.

Technique	Brief Description	Main Advantage for 6G Edge	Main Challenge for 6G Edge
Federated Learning (FL)	Collaborative training on local devices without sharing raw data; only updates are exchanged [18].	Preservation of local data privacy; use of distributed data [18].	Heterogeneity (data/systems); communication overhead; security of updates [91].
Split Learning (SL)	Neural model split between client and server; client processes early layers, server processes the rest [90].	Reduces computational load on the client compared with FL; privacy of raw data [18].	Latency due to bidirectional communication; difficulty in deciding split point; privacy of intermediate activations [90].
Edge Large AI Models (LAMs)	Adaptation and deployment of large models (LLMs, PFMs) at the edge [9].	Generalized capabilities; multimodal processing; few-/zero-shot learning [9].	Massive compute/memory/data demands vs. limited edge resources [91].
TinyML	Execution of ML on devices with extremely limited resources (microcontrollers) [59].	Enables intelligence on very small/low-power devices.	Very limited model capacity; potentially reduced accuracy.
AirComp	Aggregation of signals (e.g., FL gradients) leveraging the wireless channel [6].	Reduces latency and spectrum use for distributed aggregation.	Sensitivity to channel noise; requires precise synchronization.

Figure 6 illustrates the accuracy–communication trade-off for FL under gradient compression and 6G channel conditions.



CIFAR-10, non-IID distribution (Dirichlet $\alpha = 0.1$), 50 clients, participation rate = 20%, Rayleigh fading channel, SNR = 15 dB

Figure 6. FL Accuracy–Communication Trade-Off under Gradient Compression and 6G Channel Conditions [47,48].

4. Architectural and Orchestration Proposals for 6G with Distributed AI

Realizing the 6G vision, with its emphasis on ubiquitous intelligence and distributed computing, requires a significant evolution, and in some cases a reinvention, of the network architecture. Current proposals explore how to structure and manage these complex networks in order to integrate edge computing and artificial intelligence as native capabilities; the proposed intelligent distributed architecture and orchestration is shown in Figure 7.

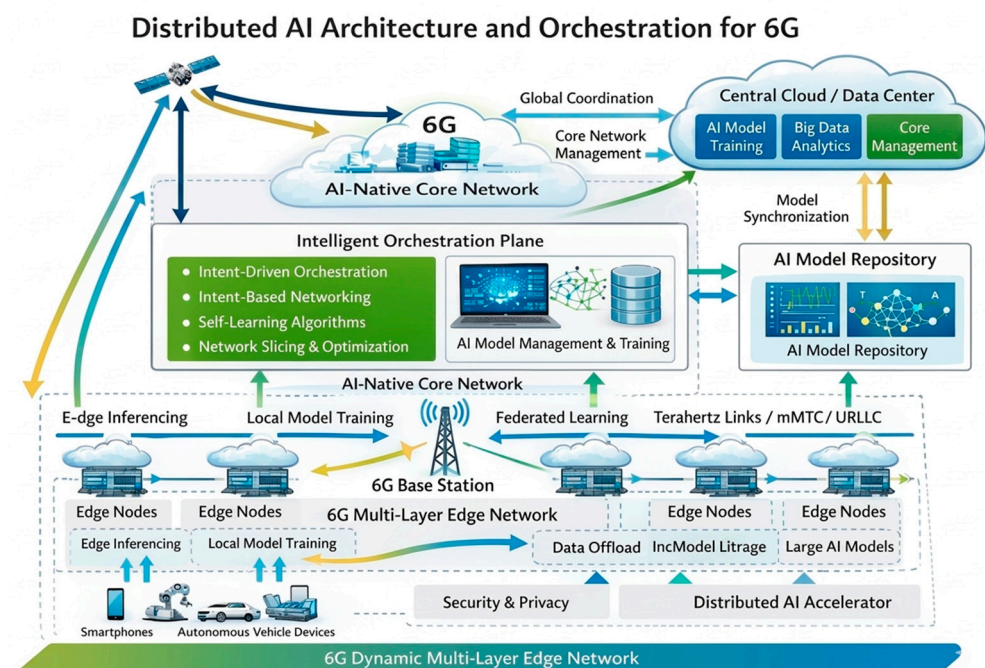


Figure 7. Proposed intelligent distributed architecture and orchestration.

4.1. Evolution of Network Architecture Towards 6G

The 6G architecture will not simply be an extrapolation of 5G; rather, it will incorporate new design principles and foundational capabilities:

4.1.1. Design Principles

The aim is to develop an architecture that is modular, supports the agile introduction of new functionalities, is simple and sustainable in the long term, intrinsically trustworthy, cloud-native, facilitates a seamless migration from 5G and, above all, enables innovation [60].

4.1.2. Multidimensional Integration

A defining characteristic of 6G will be the seamless integration of different communication domains: terrestrial networks; non-terrestrial networks (NTN), including satellites (LEO, MEO, GEO), high-altitude platforms (HAPS) and unmanned aerial vehicles (UAVs); and even aerial and underwater communication systems [8]. The objective is to achieve truly global and ubiquitous coverage.

4.1.3. Horizontalization and Disaggregation

The trend towards separating software from hardware, initiated in 5G, will deepen in 6G. Increased adoption of virtualization (NFV), software-defined networking (SDN), service-based architectures (SBA), cloud-native principles and open interfaces (such as those promoted by O-RAN) is expected across all network domains (RAN, Core, Transport) [1]. This aims to enhance flexibility, efficiency, automation and the overall capacity for innovation.

4.1.4. AI-Native Design and Distributed Computing

The 6G architecture must be designed from the outset to natively support artificial intelligence and distributed computing [1]. This not only involves enabling AI to optimize the network, but also allowing the network itself to operate as a platform for executing and orchestrating distributed AI functions and edge-computing services. Integration with sensing (ISAC) is also an emerging architectural requirement [6].

4.2. Proposed Reference Architectures

Given the early stage of 6G development, several architectural proposals have been put forward, each reflecting different approaches and priorities.

4.2.1. Evolutionary Vision

This proposal advocates a single-step migration from 5G Standalone (SA) to 6G SA, evolving the 5G Core (5GC) based on the Service-Based Architecture (SBA) rather than creating an entirely new core [60]. In the RAN domain, it supports native lower-layer split (LLS) functionality to optimize performance and enable intent-based automation through Service Management and Orchestration (SMO) functions [60]. This approach prioritizes the reuse of existing investments and a smoother transition path.

4.2.2. Task-Oriented Native AI Architecture (TONA)

This is a more revolutionary proposal, suggesting a fundamental shift from the traditional communication “session” management paradigm to an AI “task”-centric paradigm [5]. TONA explicitly manages multidimensional resources, communication, computing, data and AI models, and employs a task-based control plane to orchestrate multi-node collaborations and guarantee customized AI Quality of Service (QoAIS). It promises significant advantages in latency, efficiency and privacy for AI services integrated within the network [5]. The mathematical model presented in this section is derived from the TONA architecture proposed by Yang et al. [5], as shown in Section 3, adapted with explicit formulations for task-oriented communication and computation co-design. Equations

are either restated from [5] with notation harmonized to this survey, or represent original extensions for the 6G context analyzed in this work.

Mathematical Task Model: In TONA, each AI task τ_i is formally characterized as a tuple:

$$\tau_i = (D_i, C_i, L_i, S_i, R_i, M_i, \mathcal{A}_i) \quad (66)$$

where:

- $D_i = d_i^{\text{input}}, d_i^{\text{output}}$: data requirements (input/output data sizes in bits).
- $C_i \in \mathbb{R}^+$: computational requirement (FLOPs).
- $L_i^{\max} \in \mathbb{R}^+$: maximum tolerable end-to-end latency.
- $S_i \in \mathbb{R}^+$: storage requirement for model and intermediate data (bytes).
- $R_i \in [0, 1]$: required reliability (success probability).
- M_i : AI model specification w , arch, ops (parameters, architecture, operations).
- $\mathcal{A}_i \subseteq 1, \dots, N$: set of collaborative nodes for task execution.

Quality of AI Service (QoAIS) Metrics: TONA introduces QoAIS as a multi-dimensional performance vector:

$$\text{QoAIS}_i = (\mathcal{L}_i, \mathcal{A}_i, \mathcal{E}_i, \mathcal{P}_i, \mathcal{T}_i) \quad (67)$$

where:

- $\mathcal{L}_i$: actual end-to-end latency achieved.
- $\mathcal{A}_i$: accuracy/performance of AI inference (e.g., F1-score, mAP).
- $\mathcal{E}_i$: energy efficiency (joules per inference).
- $\mathcal{P}_i$: privacy preservation level (e.g., differential privacy ϵ).
- $\mathcal{T}_i$: throughput (tasks per second).

Task-Based Resource Allocation: The TONA control plane solves the following multi-objective optimization:

$$\max_{\mathcal{A}_i, \{r_j^i\}_{j \in \mathcal{A}_i}} \sum_{i=1}^{N_{\text{tasks}}} \omega_i \cdot U(\text{QoAIS}_i) \quad (68)$$

subject to task-specific constraints:

$$\mathcal{L}_i \leq L_i^{\max}, \quad \mathcal{A}_i \geq A_i^{\min}, \quad \mathcal{P}_i \geq P_i^{\min}, \quad \forall i \quad (69)$$

and resource constraints across collaborative nodes:

$$\sum_{i: j \in \mathcal{A}_i} r_j^i \leq R_j^{\max}, \quad \forall j \in \{1, \dots, M\} \quad (70)$$

where:

- ω_i are task priority weights.
- $U(\cdot)$ is a utility function mapping QoAIS to user satisfaction.
- r_j^i represents resources allocated by node j to task i .
- $R_j^{\max}$ is the total resource capacity at node j .

Multi-Node Collaboration Model: For tasks requiring distributed execution across nodes $\mathcal{A}_i = n_1, \dots, n_k$, TONA orchestrates:

1. **Model Partitioning:** Partition AI model M_i into sub-models $M_i^{(1)}, \dots, M_i^{(k)}$ assigned to nodes in $\mathcal{A}_i$.
2. **Data Flow Graph:** Define directed acyclic graph (DAG) $G_i = (\mathcal{V}_i, \mathcal{E}_i)$ where:
 - Nodes $\mathcal{V}_i$ represent computation stages.
 - Edges $\mathcal{E}_i$ represent data dependencies and communication.

3. Latency Decomposition:

$$\mathcal{L}_i = \sum_{v \in \mathcal{V}_i} t_v^{\text{comp}} + \sum_{(u,v) \in \mathcal{E}_i} t_{uv}^{\text{comm}} \quad (71)$$

where:

- $t_v^{\text{comp}} = \frac{C_v}{f_v}$ is computation time at node executing stage v .
- $t_{uv}^{\text{comm}} = \frac{d_{uv}}{B_{uv} \log_2(1 + \text{SINR}_{uv})}$ is communication time between stages.

4. Critical Path Analysis: Identify critical path $\mathcal{P}_{\text{crit}}$ in DAG to minimize latency:

$$\mathcal{L}_i^{\min} = \max_{\text{path } p \in G_i} \sum_{v \in p} (t_v^{\text{comp}} + t_{uv}^{\text{comm}}) \quad (72)$$

Task-Centric Control Plane: TONA's control plane operates through a hierarchical framework:

1. Task Admission Control: For incoming task τ_i , decide admission based on:

$$\text{Admit}(\tau_i) = \begin{cases} 1, & \text{if } \exists \mathcal{A}_i, r_j^i \text{ satisfying constraints} \\ 0, & \text{otherwise} \end{cases} \quad (73)$$

2. Resource Orchestration: Solve the resource allocation optimization using:

- Online Optimization: Lyapunov optimization for dynamic task arrivals.
- Graph Neural Networks: Learn optimal node selection $\mathcal{A}_i$ from network graph.
- Deep Q-Networks (DQN): State $s_t = (\tau_i, R_j)$, action $a_t = (\mathcal{A}_i, r_j^i)$, reward $r_t = \sum_i \omega_i U(\text{QoAIS}_i)$.

3. Inter-Node Coordination: Synchronize collaborative nodes via [5]:

$$\text{Schedule}(G_i) = \underset{\pi}{\operatorname{argminmax}}_{v \in \mathcal{V}_i} (\text{start}_{\pi(v)} + t_v^{\text{comp}}) \quad (74)$$

subject to precedence constraints from DAG G_i

Privacy-Preserving Task Execution: TONA incorporates privacy constraints via [5]:

$$\mathcal{P}_i = \min_{j \in \mathcal{A}_i} \mathcal{P}_j (\text{data}_i, M_i^{(j)}) \quad (75)$$

where $\mathcal{P}_j$ quantifies privacy leakage at node j . For differential privacy [5]:

$$\mathcal{P}_j = \epsilon_j \text{ s.t. } \Pr[\mathcal{M}(\mathcal{D}) \in S] \leq e^{\epsilon_j} \Pr[\mathcal{M}(\mathcal{D}') \in S] + \delta \quad (76)$$

for neighbouring datasets $\mathcal{D}, \mathcal{D}'$ and mechanism $\mathcal{M}$.

4.2.3. Integrated Satellite–Terrestrial Architectures (SAGIN):

Reference architectures have been proposed for integrating 6G terrestrial and non-terrestrial networks (NTN), evolving from non-virtualized schemes to fully virtualized architectures with a unified management and orchestration (MANO) system across both segments [98]. These architectures are crucial for enabling global coverage and supporting use cases in remote or high-mobility environments (maritime, aerial).

4.2.4. O-RAN-Based Architectures with AI

These architectures leverage the open and disaggregated design of O-RAN, particularly the RAN Intelligent Controller (RIC) in its Near-Real-Time and Non-Real-Time variants, to introduce data-driven intelligence and optimization within the access net-

work [99]. Recent proposals explore the use of intelligent agents, including LLM-based agents, implemented as xApps or rApps on top of the RIC to perform automated and intuitive orchestration of edge AI services and network adaptation [99].

Current Deployment Status of O-RAN: The O-RAN Alliance, comprising over 300 member organizations, has active deployments across multiple operators globally. Rakuten Mobile (Japan) operates the world's first fully cloud-native O-RAN commercial network since 2020. DISH Network (USA) and Vodafone (UK/Europe) have initiated large-scale O-RAN rollouts. Early deployment data indicate 15–20% spectral efficiency improvements and 10–15% energy consumption reductions [100], as shown in Section 5, compared to traditional RAN architectures, validating the architectural principles discussed in this work and providing operational experience relevant to 6G AI-native RAN.

4.2.5. Other Specific Proposals

Additional approaches focus on particular dimensions, such as data-oriented architectures (e.g., Data Plane for Data as a Service, DaaS) to facilitate data management and processing for distributed AI [14]; architectures designed for the seamless integration of sensing and edge AI (ISEA) [101]; or architectures optimized to support the unique demands of the Metaverse at the edge [38].

This diversity of architectural proposals reflects an active exploration phase across both industry and academia. While some approaches prioritize continuity and pragmatic evolution from 5G, others advocate more radical shifts to integrate intelligence and distribution natively from the outset. This early lack of consensus, although fostering innovation, also introduces the risk of fragmentation or the recurrence of the complexities observed in 5G, such as multiple competing architectural options (NSA vs. SA) [60]. It is likely that the standardization process will seek convergence, although the final direction has yet to be defined.

4.3. Comparative Analysis and Prioritization of Architectural Approaches

The diverse architectural proposals for 6G present distinct advantages and limitations depending on deployment scenarios and operator priorities. A critical comparative analysis reveals the following:

- **Evolutionary versus Revolutionary Approaches:** Evolutionary architectures (building incrementally on 5G foundations) offer lower migration risk, compatibility with existing infrastructure, and gradual CAPEX/OPEX scaling. They are well-suited for operators with substantial 5G investments seeking smooth upgrade paths. However, they may not fully exploit 6G's potential for distributed intelligence and AI-native services, potentially yielding suboptimal performance in latency-critical use cases (sub-1 ms URLLC). Revolutionary approaches (e.g., TONA) promise optimal performance through clean-slate AI-native design, task-oriented resource allocation, and end-to-end optimization. They are best suited for greenfield deployments, specialized private networks (industrial, enterprise), or scenarios where performance maximization justifies the migration costs. The primary barrier is deployment complexity and the need for comprehensive ecosystem support.
Recommendation: Hybrid strategies—an evolutionary core with revolutionary enhancements at the edge—offer pragmatic balance, allowing operators to leverage existing investments while selectively deploying advanced capabilities where value is highest.
- **O-RAN versus Integrated Vendor Solutions:** O-RAN-based architectures enable multi-vendor ecosystems, flexibility in functional splits, and AI-driven RAN optimization through RIC. They reduce vendor lock-in and potentially lower costs through com-

moditization. Early deployments show promising results in terms of energy-efficiency and spectral utilization. Integrated vendor solutions offer superior performance optimization (co-designed hardware and software), simplified operations, and mature support ecosystems. They are currently more reliable for mission-critical deployments. Recommendation: O-RAN is strategically important for long-term openness and innovation, but near-term deployments for critical services may favour integrated solutions. The parallel development of both approaches is advisable, with O-RAN gaining maturity through non-critical use cases before mission-critical adoption.

- **SAGIN versus Terrestrial-Only:** Satellite–terrestrial integration (SAGIN) addresses coverage gaps in remote, rural, maritime, and aerial scenarios where terrestrial infrastructure is economically infeasible. It supports global connectivity and resilience. However, SAGIN introduces complexity in handover, routing, and latency management (LEO satellite latency ~25–50 ms) [102], Section 6.1, Table 6.1-1; [103]. Terrestrial-only architectures achieve lower latency and higher bandwidth in urban and suburban areas but cannot provide universal coverage.

Recommendation: SAGIN is essential for truly global 6G but should be deployed strategically for underserved areas and specific use cases (IoT, maritime, aviation) rather than as a universal replacement for terrestrial infrastructure.

- **Centralized versus Distributed Intelligence:**
- **Fully centralized AI (cloud-based)** offers computational advantages, easier model updates, and superior performance for complex tasks but fails to meet latency and privacy requirements for many 6G use cases. **Fully distributed AI (on-device, edge-only)** maximizes privacy and minimizes latency but faces resource constraints, model staleness, and coordination challenges. **Hybrid approaches** (hierarchical AI with edge, Fog, and cloud tiers) provide optimal trade-offs, dynamically placing intelligence based on task requirements, resource availability, and network conditions.

Recommendation: Hybrid, adaptive AI placement is the most pragmatic approach, with intelligent orchestration determining optimal execution location based on real-time context.

- **Use Case Prioritization:** For ultra-low latency applications (V2X, industrial robotics, XR), prioritize dense edge deployment, revolutionary AI-native architectures, and local intelligence. For massive IoT (smart cities, agriculture), favour energy-efficient edge processing, hierarchical FL, and scalable MEC with sparse deployment. For high-bandwidth applications (holographic communications, cloud gaming), combine edge caching, predictive content delivery, and hybrid edge–cloud processing. For privacy-sensitive services (healthcare, finance), mandate on-device or edge AI with FL/SL, regulatory-compliant data handling, and zero-knowledge proofs.

This comparative framework enables context-specific architectural decisions rather than one-size-fits-all prescriptions, aligning technical capabilities with business objectives and regulatory constraints.

4.4. Critical Analysis of Architectural Trade-Offs

The diverse architectural proposals for 6G networks present fundamental trade-offs that must be carefully evaluated based on the deployment context, use case requirements, and economic constraints.

- **Complexity versus Performance:** Fully distributed, AI-native architectures (e.g., TONA) promise optimal performance through intelligent resource allocation and context-aware services. However, they introduce significant orchestration complexity, requiring sophisticated coordination mechanisms across heterogeneous edge nodes, dynamic workload migration, and real-time resource optimization. In contrast, evo-

lutionary approaches building on existing 5G infrastructure offer simpler migration paths but may not fully exploit the potential of distributed intelligence. The choice depends on whether operators prioritize performance maximization (justifying higher complexity) or operational simplicity (accepting performance suboptimality).

- **Scalability Limits:** Edge AI techniques face inherent scalability challenges. Federated Learning, while privacy-preserving, suffers from communication overhead, which grows with the number of participating devices, limiting practical deployments to hundreds or thousands of clients rather than millions. Split Learning reduces per-device computation but increases sequential processing time and network round-trips. Centralized cloud training scales efficiently but sacrifices privacy and increases latency. Hybrid approaches (hierarchical FL with edge aggregation) offer a middle ground but add architectural complexity.
- **Deployment Feasibility:** O-RAN-based architectures promise flexibility and vendor-neutral interoperability, enabling dynamic function splits and AI-driven optimization. However, real-world deployments face challenges including hardware heterogeneity, integration with legacy infrastructure, and the maturity of O-RAN standards. Early pilots demonstrate feasibility but also reveal performance gaps compared to integrated vendor solutions, suggesting a gradual transition rather than immediate wholesale adoption.
- **Economic Cost:** Dense edge infrastructure deployment (required for sub-5 ms latency) can increase CAPEX by 200–400% compared to traditional centralized architectures, as edge servers must be deployed at base station sites, aggregation points, and enterprise premises. Operational expenditure (OPEX) also increases due to distributed maintenance, energy consumption, and the need for skilled personnel at multiple locations. Economic viability depends on revenue opportunities from latency-sensitive services (AR/VR, autonomous vehicles, industrial automation) that can justify the infrastructure investment.
- **Privacy versus Performance:** On-device and edge AI maximize privacy by keeping data local, but constrained edge resources limit model complexity and accuracy. Cloud-based AI achieves superior performance through access to massive computation and data but requires data centralization. This trade-off is particularly critical in healthcare and financial services, where regulatory constraints may mandate edge processing despite potential accuracy reductions of 5–10% compared to cloud-based alternatives.
- **Sustainability Considerations:** While edge processing can reduce network energy consumption by minimizing data transmission, the proliferation of edge servers increases total infrastructure energy use. Life-cycle analysis suggests that edge deployment is energy-efficient only when utilization rates exceed 40–50% [104,105]; below this threshold, centralized cloud processing is more sustainable. This highlights the importance of intelligent workload placement and resource consolidation strategies.

These trade-offs underscore the need for context-specific architectural decisions rather than one-size-fits-all solutions. Hybrid architectures that dynamically balance edge and cloud resources based on application requirements, network conditions, and economic constraints represent a pragmatic path forward for 6G deployment.

Beyond single-agent reinforcement learning, Multi-Agent Reinforcement Learning (MARL) has emerged as a particularly promising approach for managing the inherent complexity and distributed nature of 6G networks. Unlike traditional RL, where a single agent learns to optimize a specific objective, MARL involves multiple autonomous agents that learn collaboratively or competitively to achieve system-wide goals while operating in a shared environment.

Mathematical Framework for MARL in 6G:

The following MARL formulation is an original contribution of this work, synthesizing standard stochastic game theory with 6G-specific state, action, and reward definitions. MARL for 6G networks is formalized as a Stochastic Game (also called Markov Game), extending the single-agent MDP framework:

$$\mathcal{G} = (\mathcal{N}, \mathcal{S}, \{\mathcal{A}^i\}_{i \in \mathcal{N}}, \mathcal{P}, \{r^i\}_{i \in \mathcal{N}}, \gamma) \quad (77)$$

where:

- $\mathcal{N} = 1, 2, \dots, n$: set of n agents (e.g., edge nodes, base stations).
- $\mathcal{S}$: global state space of the 6G network.
- $\mathcal{A}^i$: action space of agent i .
- $\mathcal{A} = \mathcal{A}^1 \times \dots \times \mathcal{A}^n$: joint action space.
- $\mathcal{P} : \mathcal{S} \times \mathcal{A} \times \mathcal{S} \rightarrow [0, 1]$: state transition probability.
- $r^i : \mathcal{S} \times \mathcal{A} \rightarrow \mathbb{R}$: reward function for agent i .
- $\gamma \in [0, 1)$: discount factor.

State Space for 6G Networks: The global state $s_t \in \mathcal{S}$ at time t captures:

$$s_t = (s_t^{\text{network}}, s_t^{\text{traffic}}, s_t^{\text{users}}, s_t^{\text{resources}}) \quad (78)$$

where:

- s_t^{network} = topology, channel conditions, interference levels.
- $s_t^{\text{traffic}} = \lambda_j(t)$, QoS requirements (arrival rates at nodes).
- s_t^{users} = locations, mobility patterns, service types.
- $s_t^{\text{resources}} = B_j(t), F_j(t), S_j(t)$ (available bandwidth, compute, storage at each node j).

Each agent i typically observes a local/partial observation $o_t^i \subseteq s_t$.

Action Space for 6G Edge Agents: For an edge node agent i , the action space includes:

$$\mathcal{A}^i = \mathcal{A}_{\text{alloc}}^i \times \mathcal{A}_{\text{handover}}^i \times \mathcal{A}_{\text{power}}^i \times \mathcal{A}_{\text{spectrum}}^i \quad (79)$$

where:

- $\mathcal{A}_{\text{alloc}}^i$: resource allocation decisions (compute, storage per task).
- $\mathcal{A}_{\text{handover}}^i \in \{0, 1\}^{N_{\text{UEs}}}$: handover/migration decisions for connected UEs.
- $\mathcal{A}_{\text{power}}^i \in [P_{\min}, P_{\max}]$: transmission power level.
- $\mathcal{A}_{\text{spectrum}}^i \subseteq 1, \dots, N_{\text{RB}}$: resource block allocation.

Reward Functions: Agents can have individual or shared rewards. For cooperative MARL (common in 6G):

$$r^i(s_t, \mathbf{a}_t) = r^{\text{global}}(s_t, \mathbf{a}_t) = - \left[\alpha_L \cdot L_{\text{avg}}(s_t, \mathbf{a}_t) + \alpha_E \cdot E_{\text{total}}(s_t, \mathbf{a}_t) + \alpha_D \cdot D_{\text{drop}}(s_t, \mathbf{a}_t) \right] \quad (80)$$

where $\mathbf{a}_t = (a_t^1, \dots, a_t^n)$ is the joint action, and the following hold:

- L_{avg} : average end-to-end latency across all tasks.
- E_{total} : total network energy consumption.
- D_{drop} : task drop rate.

For competitive scenarios (e.g., spectrum sharing between operators), agents may obtain conflicting rewards.

Policy and Value Functions: Each agent i learns a policy $\pi^i : \mathcal{O}^i \rightarrow \Delta(\mathcal{A}^i)$ mapping observations to action distributions. The joint policy is $\boldsymbol{\pi} = (\pi^1, \dots, \pi^n)$.

The state-value function for agent i under joint policy π :

$$V_{\pi}^i(s) = \mathbb{E}_{\pi} \left[\sum_{t=0}^{\infty} \gamma^t r^i(s_t, \mathbf{a}_t) \mid s_0 = s \right] \quad (81)$$

The state-action value (Q-function):

$$Q_{\pi}^i(s, \mathbf{a}) = r^i(s, \mathbf{a}) + \gamma \mathbb{E}_{s' \sim \mathcal{P}(\cdot | s, \mathbf{a}), \mathbf{a}' \sim \pi(\cdot | s', \mathbf{a}')} [Q_{\pi}^i(s', \mathbf{a}')] \quad (82)$$

Nash Equilibrium and Solution Concepts: In cooperative MARL, agents seek to maximize the sum of rewards:

$$\pi^* = \operatorname{argmax}_{\pi} \sum_{i=1}^n V_{\pi}^i(s_0) \quad (83)$$

In general-sum games, we seek Nash Equilibrium: π^* is a Nash equilibrium if, for each agent i :

$$V_{\pi^*}^i(s) \geq V_{(\pi^i, \pi^{*-i})}^i(s), \quad \forall \pi^i, \forall s \quad (84)$$

where π^{*-i} denotes the policies of all agents except i .

Learning Algorithms:

1. Independent Q-Learning (IQL): Each agent learns independently treating others as part of the environment. The update rule is as follows:

$$Q^i(s_t, a_t^i) \leftarrow Q^i(s_t, a_t^i) + \alpha \left[r_t^i + \gamma \max_{a'} Q^i(s_{t+1}, a') - Q^i(s_t, a_t^i) \right] \quad (85)$$

Converges in stationary environments but may fail due to non-stationarity in multi-agent settings.

2. Centralized Training with Decentralized Execution (CTDE): Agents learn with access to global information during training but execute using only local observations. Example: QMIX algorithm learns a joint action–value function:

$$Q_{\text{tot}}(s, \mathbf{a}; \theta) = g\left(s, Q^1(o^1, a^1; \theta^1), \dots, Q^n(o^n, a^n; \theta^n)\right) \quad (86)$$

where g is a monotonic mixing network ensuring

$$\frac{\partial Q_{\text{tot}}}{\partial Q^i} \geq 0, \quad \forall i \quad (87)$$

This guarantees that individual greedy actions lead to globally optimal joint actions.

3. Multi-Agent Actor–Critic (MAAC): Each agent i has the following:

- Actor: $\pi^i(a^i | o^i; \theta_{\pi}^i)$.
- Critic: $Q^i(s, \mathbf{a}; \theta_Q^i)$ (may use global state s during training).

Policy gradient update:

$$\nabla_{\theta_{\pi}^i} J^i = \mathbb{E}_{s, \mathbf{a}} \left[\nabla_{\theta_{\pi}^i} \log \pi^i(a^i | o^i; \theta_{\pi}^i) \cdot Q^i(s, \mathbf{a}) \right] \quad (88)$$

Convergence and Complexity:

- *Convergence*: MARL convergence depends on the game structure:
 - *Fully cooperative*: Can achieve near-optimal joint policy with CTDE methods.
 - *General-sum*: Convergence to ϵ -Nash equilibrium in polynomial time under certain conditions.

- Sample complexity: $\tilde{\mathcal{O}}(|\mathcal{S}||\mathcal{A}|^n / (1 - \gamma)^4 \epsilon^2)$ for ϵ -optimal policy.
- Computational complexity per step:
 - IQL: $\mathcal{O}(n \cdot |\mathcal{A}^i|)$ (parallelizable).
 - CTDE (QMIX): $\mathcal{O}(|\mathcal{A}|^n)$ during training, $\mathcal{O}(n \cdot |\mathcal{A}^i|)$ during execution.
 - MAAC: $\mathcal{O}(n \cdot d_{\text{network}})$ for neural network forward/backward passes.

In 6G networks, MARL is especially relevant for scenarios requiring coordinated decision-making across distributed edge nodes, base stations, and user equipment. For example, multiple edge servers can act as cooperative agents learning to jointly optimize resource allocation, load balancing, and service migration to minimize overall latency and energy consumption while adapting to dynamic traffic patterns and user mobility.

4.5. Intelligent Orchestration of Services and Resources

Regardless of the specific architecture adopted, the inherent complexity of 6G networks, with their distributed, heterogeneous, virtualized and highly dynamic nature, makes intelligent orchestration an absolutely indispensable component [106]. Manual or static rule-based management is infeasible [57]. Orchestration acts as the “central nervous system” of the network, enabling coherent, adaptive and efficient operation. Key aspects of intelligent orchestration in 6G include:

1. Joint Communication, Computing and Storage Management (JCC)

It is essential to optimize communication resources (bandwidth, spectrum), computing resources (CPU, GPU, NPU at edge nodes) and storage along the entire network continuum, from devices to the cloud, via the edge [1]. This is particularly critical in integrated networks such as SAGIN [107]. Since these joint-optimization problems are often complex (NP-hard), advanced heuristics or, increasingly, AI-based approaches (RL, DL) are required to obtain efficient real-time solutions [90].

- Mathematical Formulation of the JCC Problem: The Joint Communication, Computing and Storage (JCC) optimization problem can be formulated as a multi-objective optimization over the network continuum. Consider a 6G network with N user equipment (UE) devices, M edge nodes, and K cloud servers. Each service request i is characterized by:

$$\tau_i = (D_i, C_i, S_i, L_i^{\max}, R_i^{\min}) \quad (89)$$

where:

- D_i : data size (in bits).
- C_i : computational requirement (in FLOPs).
- S_i : storage requirement (in bytes).
- $L_i^{\max}$: maximum tolerable latency.
- $R_i^{\min}$: minimum required reliability.

Decision Variables:

- $x_{ij} \in 0, 1$: binary placement variable (task i assigned to node j).
- $b_{ij} \geq 0$: bandwidth allocated for task i at node j (Hz).
- $f_{ij} \geq 0$: computing frequency allocated for task i at node j (GHz).
- $s_{ij} \geq 0$: storage allocated for task i at node j (bytes).
- $p_{ij} \geq 0$: transmission power for task i at node j (Watts).

Objective Function: Minimize the weighted sum of end-to-end latency, energy consumption, and cost:

$$\min \sum_{i=1}^{N_{\text{tasks}}} \left[\alpha_1 L_i^{\text{E2E}} + \alpha_2 E_i^{\text{total}} + \alpha_3 C_i^{\text{cost}} \right] \quad (90)$$

where $\alpha_1, \alpha_2, \alpha_3$ are weighting factors reflecting operator priorities.

End-to-End Latency Model:

$$L_i^{\text{E2E}} = L_i^{\text{queue}} + L_i^{\text{trans}} + L_i^{\text{comp}} + L_i^{\text{backhaul}} \quad (91)$$

where:

- Queueing latency: $L_i^{\text{queue}} = \frac{\lambda_i}{\mu_j - \lambda_i}$ (M/M/1 queue model at node j).
- Transmission latency: $L_i^{\text{trans}} = \frac{D_i}{b_{ij} \log_2 \left(1 + \frac{p_{ij} h_{ij}}{N_0 b_{ij}} \right)}$ (Shannon capacity).
- Computation latency: $L_i^{\text{comp}} = \frac{C_i}{f_{ij}}$.
- Backhaul latency: $L_i^{\text{backhaul}} = d_{ij}/c$ (propagation delay, c is speed of light).

Energy Consumption Model:

$$E_i^{\text{total}} = E_i^{\text{trans}} + E_i^{\text{comp}} = p_{ij} \cdot L_i^{\text{trans}} + \kappa f_{ij}^3 \cdot L_i^{\text{comp}} \quad (92)$$

where κ is the effective capacitance coefficient of the processing chip (typically $\kappa \approx 10^{-27}$ for modern processors).

Cost Model:

$$C_i^{\text{cost}} = \beta_b b_{ij} + \beta_f f_{ij} + \beta_s s_{ij} \quad (93)$$

where $\beta_b, \beta_f, \beta_s$ are unit costs for bandwidth, compute, and storage, respectively.

Constraints:

- Placement Constraint (each task assigned to exactly one node):

$$\sum_{j=1}^{M+K} x_{ij} = 1, \quad \forall i \quad (94)$$

- Bandwidth Constraint (total allocated bandwidth cannot exceed availability):

$$\sum_{i=1}^{N_{\text{tasks}}} x_{ij} b_{ij} \leq B_j^{\text{max}}, \quad \forall j \quad (95)$$

- Computing Constraint (total computing frequency cannot exceed capacity):

$$\sum_{i=1}^{N_{\text{tasks}}} x_{ij} f_{ij} \leq F_j^{\text{max}}, \quad \forall j \quad (96)$$

- Storage Constraint (total storage cannot exceed capacity):

$$\sum_{i=1}^{N_{\text{tasks}}} x_{ij} s_{ij} \leq S_j^{\text{max}}, \quad \forall j \quad (97)$$

- Latency Constraint (end-to-end latency must meet requirements):

$$L_i^{\text{E2E}} \leq L_i^{\text{max}}, \quad \forall i \quad (98)$$

- Reliability Constraint (service reliability must be guaranteed):

$$R_{ij} \geq R_i^{\text{min}}, \quad \forall i \quad (99)$$

where $R_{ij} = (1 - \text{BLER}_{ij}) \cdot \text{Avail}_j$ is the reliability combining block error rate and node availability.

- Power Constraint (transmission power bounded):

$$0 \leq p_{ij} \leq P_{\max}, \quad \forall i, j \quad (100)$$

Complexity Analysis:

Theorem: The JCC optimization problem is NP-hard.

Proof Sketch: The problem can be reduced from the multiple knapsack problem (MKP), which is known to be NP-hard. Consider the special case where the following hold:

- All tasks have identical computation requirements $C_i = C$.
- Latency constraints are relaxed.
- Only computing resources are constrained.

This reduces to the following: assign N_{tasks} tasks to M edge nodes to minimize total energy, subject to $\sum_{i \in \mathcal{T}_j} x_{ij} C \leq F_j^{\max}$ for each node j . This is the MKP with task values and weights. Since MKP is NP-hard, the general JCC problem (which includes additional constraints and objectives) is also NP-hard.

Solution Approaches: Given the NP-hardness, exact solutions via integer linear programming (ILP) are intractable for large-scale networks. Practical approaches include:

- Heuristic Algorithms:
 - Greedy placement based on latency proximity.
 - Genetic algorithms for joint optimization.
 - Complexity: $\mathcal{O}(N_{\text{tasks}} \cdot M \cdot K)$ per iteration.
- Convex Relaxation: Relax binary variables $x_{ij} \in [0, 1]$ and solve using convex optimization (e.g., ADMM, primal-dual methods), then round to the integer solution. Approximation ratio is typically within 5–10% of the optimal values.
- Deep Reinforcement Learning (DRL):
 - State space: $\mathcal{S} = (B_j, F_j, S_j, \tau_i)$ (resource availability and pending tasks).
 - Action space: $\mathcal{A} = (j, b_{ij}, f_{ij}, s_{ij})$ (placement and allocation).
 - Reward: $r = -[\alpha_1 L_i^{\text{E2E}} + \alpha_2 E_i^{\text{total}} + \alpha_3 C_i^{\text{cost}}]$.
 - Policy: $\pi_\theta(a|s)$ learned via PPO, A3C, or SAC.
 - Training complexity: $\mathcal{O}(T_{\text{episodes}} \cdot N_{\text{steps}} \cdot N_{\text{tasks}})$.
 - Inference complexity: $\mathcal{O}(N_{\text{tasks}})$ (feasible in real-time).
- Graph Neural Networks (GNNs): Model the network topology as a graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ with nodes representing UEs, edge nodes, and cloud servers, and edges representing connectivity. GNN learns node embeddings capturing resource states and task requirements, enabling efficient joint optimization:
 - Node features: $h_i^{(0)} = [B_j, F_j, S_j, \text{Location}_j]$.
 - Message passing: $h_j^{(l+1)} = \sigma \left(\sum_{k \in \mathcal{N}(j)} W^{(l)} h_k^{(l)} \right)$.
 - Task-node matching via attention: $\text{score}(i, j) = \text{softmax}(h_i^T W h_j)$.
 - Complexity: $\mathcal{O}(|\mathcal{E}| \cdot L_{\text{layers}} \cdot d_{\text{hidden}})$.

Multi-Timescale Optimization: JCC optimization operates across multiple timescales:

- Milliseconds: Dynamic resource allocation for incoming tasks (DRL/GNN).
- Seconds: Load balancing and task migration.
- Minutes/Hours: Capacity planning and infrastructure adjustment.

This hierarchical decomposition reduces computational complexity while maintaining near-optimal performance.

2. Intent-Based Networking (IBN)

IBN represents a higher level of abstraction in network management [108]. Instead of configuring low-level parameters, users or applications declare high-level goals or requirements as “intentions” (e.g., “ensure <5 ms latency for VR traffic”, “deploy object-detection service at the edge near camera X”) [99]. The intelligent orchestration system (NMS/Orchestrator) is responsible for interpreting the intention, translating it into specific network and resource configurations, activating those configurations, and continuously ensuring intention fulfilment [60]. IBN architectures for 6G are being actively explored [109], including the use of LLMs to interpret natural-language intents and automate the entire intent lifecycle [99].

Comparative Framework for Orchestration Overhead vs. Performance Gains: The introduction of AI-native orchestration in O-RAN and SAGIN architectures introduces significant control-plane complexity that must be evaluated against performance gains. Key evaluation metrics include: (1) Control-Plane Signalling Load: Centralized orchestration generates $O(N)$ signalling messages per decision cycle for N nodes; distributed orchestration generates $O(N^2)$ peer-to-peer messages but enables parallel decisions; (2) Convergence Time: centralized RL-based orchestrators converge in 10–100 ms for near-RT control loops; distributed MARL schemes require 100–500 ms due to inter-agent coordination (estimated from multi-agent RL convergence studies; see [110,111]); (3) Scalability: Centralized orchestration faces bottlenecks beyond approximately 1000 concurrent sessions; hierarchical and distributed approaches scale linearly with the number of edge nodes. Qualitatively, centralized orchestration offers simpler implementation and globally optimal decisions but is vulnerable to single-point failures. Distributed orchestration provides resilience and scalability but requires consensus mechanisms (adding 5–20 ms latency) (typical Raft/Paxos consensus latency for edge clusters; see [4], Section 1), and may converge to locally optimal solutions. Hierarchical orchestration—combining local near-RT decisions (sub-millisecond) with global non-RT policy updates (seconds to minutes)—offers the best balance of performance, scalability, and overhead for 6G networks.

3. Orchestration of AI Functions at the Edge

Beyond network-resource orchestration, 6G requires the specific orchestration of AI functions and models deployed at the edge [19]. Tasks include selecting the appropriate AI model for a given purpose, deciding where to place it (which edge node or device), deploying it (e.g., as a container), allocating the necessary computing resources, monitoring its performance and managing its lifecycle (updates, retraining). Intelligent-agent-based frameworks (potentially LLM-driven) are being proposed to automate these complex decisions, considering factors such as resource availability, QoS requirements and latency constraints [19].

4. Multi-domain and Multi-agent Federation and Collaboration

Orchestration in 6G must operate across multiple administrative domains (e.g., different operators in roaming or network-sharing scenarios) and technological domains (e.g., coordination between edge and cloud, or between RAN and Core) [33]. This requires federation mechanisms that enable the controlled and secure sharing of resources and information. Moreover, in architectures with multiple distributed AI agents, orchestration must also facilitate effective collaboration among them [19].

4.6. Integration with Key Enabling Technologies

The architecture and orchestration of 6G do not exist in isolation; rather, they must be synergistically integrated with other key technologies to achieve their full potential.

4.6.1. Digital Twins (DT)

A Digital Twin is a dynamic and real-time virtual representation of a physical object, process or system [34]. In 6G, DTs may be created for the network itself, its components or the environment in which it operates.

- Roles: DTs enable scenario simulations (“what-if” analyses), testing of configurations or optimizations before deploying them in the real network, real-time monitoring and behavioural predictions for the network [6], and the generation of realistic synthetic data for training AI models, overcoming the limitations of real data [19]. They can also support industrial, logistics or smart-city applications [34], and enhance security via threat modelling [96]. Architectures combining DT with blockchain and FL have been proposed to achieve secure and efficient edge networks [112].
- Implementation: DTs require distributed IoT–edge–cloud platforms to collect data from the physical world and maintain synchronization with the virtual replica [34].

4.6.2. Integrated Sensing and Communication (ISAC)

ISAC leverages existing communication signals and infrastructure to perform environmental sensing tasks such as object detection, distance/velocity/angle estimation, localization and mapping [1].

- Synergy with Edge AI: ISAC generates substantial sensing data that can be processed by edge AI algorithms to extract useful information and support intelligent decision-making. In turn, edge AI can optimize ISAC processes themselves. This synergy leads to the paradigm of Integrated Sensing and Edge AI (ISEA) [6], in which communication, edge computing, sensing and AI are jointly designed and optimized for a given task.
- Challenges: The main difficulty lies in achieving true integration at the hardware, algorithmic and signal-design levels, rather than mere coexistence [6].

4.6.3. Blockchain

Distributed ledger technology can play a crucial role in 6G by addressing the security, privacy and trust issues in distributed and multi-stakeholder environments [113]. Blockchain is proposed for secure and decentralized identity and access management, secure sharing of spectrum and resources, data and transaction traceability and auditing, and establishing trust in distributed AI systems such as FL or in interactions with DTs [57].

The integration of these technologies creates synergistic feedback loops that enhance the overall intelligence and efficiency of the 6G system. ISAC functions as the network’s “sensory system”, capturing data from the physical world [6]. Edge AI acts as the distributed “brain”, analyzing these and other data to understand the environment, optimize the network and make decisions [101]. The Digital Twin provides a “virtual space for testing and prediction”, fuelled by ISAC data and used by edge AI to simulate scenarios, validate strategies and anticipate behaviours before acting in the real world [34]. This continuous cycle of Sense (ISAC) → Analyze/Decide (Edge AI) → Simulate/Validate (DT) → Act (Network/Control) enables far more robust, adaptive and intelligent cyber–physical systems, in which each component reinforces the capabilities of the others. Table 4 compares several of the key architectural proposals discussed.

Table 4. Comparative Analysis of Architectural and Orchestration Proposals for 6G with Distributed AI.

Proposal/Origin	Main Approach	Role of Distributed Computing/Edge AI	Associated Key Technologies	Potential Strengths/Weaknesses
Ericsson 5GC Evolved [6]	Pragmatic evolution of the 5G Core; RAN with native LLS	Edge AI for automation (SMO), MEC supported by the evolved Core	5GC SBA, LLS, IBN, SMO	(+) Smooth migration, investment reuse. (−) Potentially less optimized for AI-native designs.
TONA [8]	AI task-oriented network management	Central: AI is the task to be managed. Edge AI/MEC treated as native resources orchestrated by the network	Task-Control, QoAIS, Multi-Resource Management	(+) AI-native optimization, fine granularity. (−) Breaks current paradigms; high complexity.
O-RAN AI-enabled [101]	Intelligence and openness in the RAN	Edge AI implemented as xApps/rApps in the RIC for RAN optimization and edge-service orchestration	O-RAN, RIC, xApps/rApps, LLM Agents	(+) RAN flexibility, open innovation. (−) RAN-centric scope; O-RAN complexity; interface security.
Integrated SAGIN [19]	Terrestrial–Non-Terrestrial Integration (NTN)	MEC/Edge Computing deployed on NTN platforms (satellites, HAPS) for global intelligent-service coverage	NTN, Virtualization (NFV), Integrated MANO	(+) Ubiquitous coverage, new use cases. (−) NTN latency; MANO integration complexity.
Data Plane for DaaS [14]	Data-centric architecture for AI	Facilitates collection, transmission, processing and storage of data for distributed AI	Data Plane, DaaS APIs	(+) Optimized for AI data pipelines. (−) Less focus on compute/communication optimization.
ISEA Framework [24]	Deep integration of sensing and edge AI	Co-design of communication, edge computing, sensing and AI for specific tasks	ISAC, Edge AI, Joint Optimization	(+) Optimal performance for intelligent sensing tasks. (−) Task specificity; high complexity.

Note: (+) denotes key strengths, whereas (−) denotes key weaknesses or limitations.

4.7. Concrete Mapping to Functional Splits and Control Loops

To transition from conceptual frameworks to practical deployment, concrete mapping to functional splits, control loops, and deployment constraints is essential:

4.7.1. Functional Split Options (Building on 3GPP and O-RAN)

The distribution of network functions between centralized units (CU), distributed units (DU), and radio units (RU) directly impacts where AI and edge computing capabilities can be deployed.

4.7.2. Split Option 2 (RLC-MAC Split)

Edge AI functions deployed at DU level enable real-time MAC scheduling optimization, interference management, and QoS enforcement with sub-10 ms control loops, making it suitable for latency-critical applications within a cell cluster.

4.7.3. Split Option 7.2 (High–Low PHY Split)

Split Option 7.2 enables centralized baseband processing for multiple cells while allowing edge-based beam management and channel estimation using ML. This balances centralization benefits with edge responsiveness.

4.7.4. Split Option 8 (RU-DU Split)

This option maximizes centralization, limiting edge AI to RU-level operations (e.g., digital beamforming, fronthaul compression). It is less suitable for ultra-low latency use cases but simplifies coordination and resource pooling.

Recommendation: Dynamic functional split selection based on use case requirements, with Split 2 for latency-critical services, Split 7.2 for balanced deployments, and Split 8 for capacity-oriented scenarios.

Control Loop Hierarchies: Effective orchestration requires multiple control loops operating at different timescales.

4.7.5. Real-Time Control Loop (Sub-Millisecond to Milliseconds)

Local edge AI for radio resource allocation, beam steering, power control are implemented at DU or near-RT RIC (O-RAN); decision latency must be <1 ms.

4.7.6. Near-Real-Time Control Loop (10 ms to 1 s)

MEC-based service orchestration, traffic prediction, mobility management, and edge cache optimization are implemented at the near-RT RIC or MEC orchestrator with a decision latency of 10–100 ms.

4.7.7. Non-Real-Time Control Loop (Seconds to Minutes)

Policy management, network slicing reconfiguration, federated learning aggregation, and multi-domain coordination are implemented at non-RT RIC or the central orchestrator with a decision latency of 1–10 s.

4.7.8. Long-Term Planning (Minutes to Hours)

Capacity planning, energy optimization, proactive resource scaling based on predicted demand are implemented at OSS/BSS or the cloud management layer.

4.7.9. Mapping AI Functions to Control Loops

Real-time: DRL-based schedulers, fast interference mitigation, beam prediction.

- Near-real-time: Traffic forecasting, service migration decisions, anomaly detection.
- Non-real-time: FL model aggregation, policy learning, multi-agent coordination.
- Long-term: Demand forecasting, infrastructure planning, energy sustainability optimization.

4.7.10. Deployment Constraints and Realistic Limitations

- Edge Server Placement: Constrained by backhaul availability, power infrastructure, and physical space at base station sites. Not all cell sites can host full MEC servers; stratified deployment (dense edge at hotspots, sparse edge in rural areas) is necessary.
- Compute Resource Heterogeneity: Edge nodes vary from high-capacity MEC servers (50–100 CPU cores, GPU acceleration) to constrained IoT gateways (2–4 cores, no GPU). AI models must adapt to available resources through model compression, quantization, or offloading.
- Latency Budgets: End-to-end latency comprises multiple components: radio access (~1–5 ms), edge processing (~1–10 ms), backhaul (~1–20 ms depending on distance), application processing (~1–50 ms depending on complexity). Achieving sub-1 ms URLLC

requires all components to be co-optimized and co-located [114], Section 6; [115], Section 7.

- **Energy Constraints:** Edge deployments must respect power budgets (typically 500 W–2 kW per site for MEC servers) [116]. AI inference must be energy-efficient; heavy DL models may exceed power envelopes, necessitating model simplification or selective offloading.
- **Interoperability:** Multi-vendor environments require standardized interfaces (O-RAN open fronthaul, E2 interface, MEC platform APIs). Proprietary optimizations may not translate across vendors, limiting portability.

This concrete mapping bridges the gap between architectural vision and engineering reality, enabling the practical implementation of distributed intelligence in 6G networks.

5. Standardization and Future Perspectives

Figure 8 presents the consolidated 6G standardization roadmap across ITU-R, 3GPP, ETSI, and O-RAN Alliance.

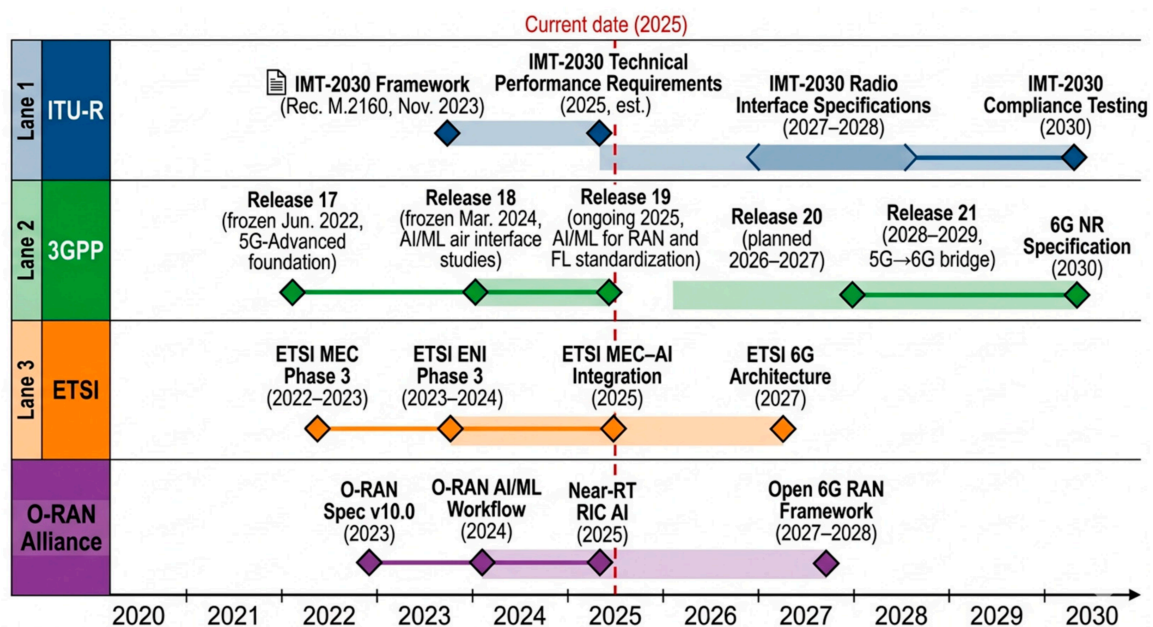


Figure 8. 6G standardization Roadmap (2020–2030). Source: Compiled by the authors based on the ITU-R IMT-2030 framework [6,117], 3GPP release timelines [118,119], and ETSI MEC standardization documents [116].

The successful development and deployment of 6G, with its deep integration of distributed computing and edge AI, depend heavily on global standardization efforts and the resolution of substantial research challenges.

5.1. Status of Standardization

Several international organizations are actively working on defining 6G.

5.1.1. ITU-R (IMT-2030)

The ITU establishes the global vision and requirements for future generations of IMT.

- **Process:** Following the established processes for IMT-2000, IMT-Advanced and IMT-2020, work on IMT-2030 (6G) began with the definition of the vision and overarching framework, culminating in Recommendation ITU-R M.2160 (“IMT-2030 Framework”), approved in November 2023 [7]. The next phase (2024–2027) will focus on defining

detailed technical requirements and evaluation criteria for candidate Radio Interface Technologies (RITs). RIT submissions will be accepted between 2027 and early 2029, with final specifications expected around 2030 [7].

- Vision and Capabilities: The IMT-2030 framework identifies key use scenarios, including enhanced versions of those in 5G (immersive communication, massive communication, HRLLC) and new scenarios such as ISAC, AI integration and ubiquitous connectivity [6]. It defines 15 target capabilities, with significant improvements over 5G in data rate, latency, density and more, while introducing new capabilities such as centimetre-level positioning, integrated sensing and, crucially, AI-related capabilities [6].

These AI-related capabilities explicitly include support for distributed data processing, distributed learning (such as FL), the execution of AI models and inference within the network [7]. Sustainability is also identified as a key pillar [7].

- Spectrum: ITU-R recognizes the need for additional spectrum across a wide range of bands, from sub-1 GHz (for broad coverage) to millimetre-wave bands and, potentially, frequencies above 92 GHz or even into the THz range (to support extreme data rates and applications such as ISAC) [2]. Spectrum harmonization is crucial [56].

5.1.2. 3GPP (Releases 19, 20 and 21)

3GPP is the main organization responsible for developing technical specifications for mobile networks and will play a central role in defining the technical foundations of 6G.

- Timeline: Formal work on 6G within 3GPP began in Release 19 (main work through ~Q3 2025) with studies on use cases and service requirements (led by SA1, resulting in TR 22.870) [120]. Release 20 (expected Q3 2025 to ~Q1 2027) will host the major technical studies in the RAN working groups, investigating candidate technologies and defining detailed technical requirements for the 6G radio interface [6]. Normative work (specification development) for the first version of 6G will take place in Release 21, with the goal of completion by late 2028 or early 2029 to align with the IMT-2030 submission schedule [6]. It is important to note that Releases 19 and 20 will also continue advancing 5G-Advanced in parallel with 6G studies [120].
- Focus on AI/ML: 3GPP has been working on AI/ML since earlier releases (e.g., Release 18, which studied AI/ML for the air interface) [118]. In Release 19, this work continues through the specification of use cases such as CSI prediction, beam management and positioning [118]. The general 3GPP approach is not to standardize AI algorithms or models themselves, given their rapid rate of evolution [119]. Instead, the goal is to standardize the infrastructure and interfaces required to support and manage AI/ML in the network. This includes mechanisms for operator-controlled data collection, model transfer and lifecycle management (activation, deactivation, performance monitoring), as well as the necessary air-interface extensions [96]. AI/ML is expected to be an integral part of 6G specifications from the outset (Release 21) [119], with potential to support online learning and greater flexibility in implementing specific functionalities [119]. Functions such as NWDAF (Network Data Analytics Function), introduced in 5G, are seen as foundational components for AI-driven analytics and automation in 6G [96].

Particularly relevant for the techniques discussed in Section 3 is the 3GPP Study Item on *AI/ML for NR Air Interface* (Release 18, TR 38.843), which addresses AI/ML-based enhancements for three specific use cases: (i) beam management, where AI-based methods can predict optimal beams from historical channel measurements, reducing beam-sweeping overhead by up to 40%; (ii) CSI feedback compression, where encoder–decoder neural

architectures (e.g., CsiNet) can reduce uplink CSI feedback by $3\times$ to $32\times$ [121], as shown in Section 6.2, Table 6.2-1, at comparable reconstruction accuracy; and (iii) positioning accuracy improvement, targeting sub-1 m accuracy using AI-enhanced signal processing. Release 19 (RAN1/RAN2 work items, ongoing 2025) extends this work to AI/ML-based link adaptation and channel estimation in high-mobility scenarios (Doppler up to 500 km/h), directly enabling the handover and mobility management use cases presented in Section 3. These 3GPP activities represent the critical bridge between the academic proposals reviewed in this survey and commercially implementable 6G features [118,119].

5.1.3. ETSI MEC

The ETSI ISG MEC continues to develop standards for multi-access edge computing.

- Evolution: Recent phases (Phase 3 and 4) focus on improving security, enabling federation across different MEC platforms (critical for roaming and multi-operator scenarios), supporting network slicing at both MEC and application levels, and aligning with the needs of vertical industries (e.g., V2X through 5GAA) [28].

There is ongoing alignment with 3GPP, particularly with SA6's work on edge-application architectures (EDGEAPP) [28].

- Relevance for 6G: The work of ETSI MEC is fundamental in providing a standardized environment that enables the deployment of edge-native applications in 6G [28]. However, some perspectives highlight that the current MEC architecture may need to evolve towards more open and flexible approaches (such as OS-MEC) to fully meet the dynamism and personalization requirements of 6G [33].

There is an inherent tension in the standardization process. On the one hand, early definition of a framework and requirements (as in IMT-2030) is essential to guide research and investment [117]. On the other hand, there is a risk of standardizing specific technologies or architectures before they are fully mature or before their implications are fully understood, which could limit future innovation or lead to suboptimal solutions [60]. The experience with multiple initial 5G deployment options (NSA/SA) suggests caution [60]. The phased approach adopted by ITU-R and 3GPP (Vision → Requirements → Technical Study → Normative Specification) [7] seeks to balance this tension by allowing the research to inform standardization, although the challenge of making key decisions at the right moment persists.

Timeline Mismatches and Gaps Between Standardization and AI Evolution: A critical challenge for 6G deployment is the mismatch between standardization timelines and the rapid evolution of AI capabilities. 3GPP Release 19 (completing approximately 2025) and Release 20 (approximately 2027) focus on studying AI/ML use cases and technical requirements, while normative 6G specifications will not be finalized until Release 21 (2028–2029). During this 3–4-year gap, AI architectures will continue to evolve rapidly—foundation model capabilities that are state-of-the-art in 2024 may be obsolete by the time 6G standards are finalized. Key standardization gaps identified include: (1) AI model lifecycle management: 3GPP specifications for model versioning, over-the-air model updates, and performance monitoring are still under study in Release 19 (SA2 work item on AI/ML model transfer), with no finalized specifications; (2) Edge AI APIs: ETSI MEC APIs for AI workload deployment remain largely proprietary across vendors, with MEC Phase 4 standardization efforts ongoing; (3) Federated Learning Interfaces: No standardized FL aggregation protocols exist across operator domains, limiting cross-operator model-sharing. The impact on deployment is significant: early 6G commercial networks (2028–2030) will likely rely on proprietary AI implementations until standards mature, risking vendor lock-in and fragmentation, similar to the 5G NSA/SA split. More agile

standardization processes—including living documents that are updated annually and closer academia–industry–standards body collaboration—are recommended to prevent this gap from widening.

A notable aspect is the strategic decision by standardization bodies, particularly 3GPP, to focus on enabling the infrastructure for AI (data collection, model management, interfaces) rather than standardizing AI models themselves [96]. This recognizes the extremely dynamic nature of AI research and aims to create a standardized yet open ecosystem in which different vendors can innovate and compete with their own AI solutions on top of a common, interoperable foundation.

In other words, standardization defines how to integrate and manage AI, not which AI should be used.

5.2. Real-World Testbeds and Early Deployments

While much of the 6G discourse remains visionary, several real-world testbeds, pilot deployments, and experimental platforms are already providing empirical validation of distributed AI and edge computing concepts:

5.2.1. MEC in Industrial Scenarios

The 5G-ACIA (5G Alliance for Connected Industries and Automation) has deployed MEC-based testbeds in manufacturing environments, demonstrating sub-10 ms latency for industrial robot control and predictive maintenance using edge AI. One notable deployment at a BMW facility in Germany uses MEC servers co-located with 5G base stations to perform real-time quality inspection using computer vision models, achieving 99.7% defect detection accuracy with 8 ms inference latency [4] in Section 5.3. (Note: These specific performance figures are from the 5G-ACIA BMW testbed deployment report; the reference provides the general framework. See also: 5G-ACIA, ‘Industrial 5G Devices Architecture and Capabilities’ [122].)

5.2.2. V2X Edge AI Pilots

The 5GAA (5G Automotive Association) coordinated cross-border V2X trials in Europe, deploying edge AI for cooperative perception and collision avoidance [123]. Vehicles share sensor data processed by roadside MEC units to create a collective environmental model, enabling “see-through” capabilities beyond individual vehicle sensor range. The results show that edge-based cooperative perception increases object detection range by 3–5× compared to individual vehicle sensors while maintaining sub-20 ms end-to-end latency.

5.2.3. Open-Source MEC Platforms

Projects such as OpenNESS (Open Network Edge Services Software) and Akraino Edge Stack provide open-source MEC platforms that are being tested in academic and industry labs worldwide. These platforms enable experimentation with edge AI workloads, service orchestration, and multi-access connectivity across diverse use cases, including smart cities, healthcare, and content delivery networks.

5.2.4. Satellite-Terrestrial Integration

The European Space Agency’s 5G/6G ARTES program has funded testbeds integrating satellite and terrestrial networks with edge AI for coverage in remote areas. A Norwegian Arctic deployment uses LEO satellite backhaul combined with terrestrial MEC for environmental monitoring, demonstrating the feasibility of SAGIN architectures in extreme environments.

5.2.5. Federated Learning in Healthcare

Several hospital networks in the United States and Europe are piloting federated learning for medical image analysis (radiology, pathology) while preserving patient privacy. These deployments use edge servers at hospital sites to train models locally on sensitive data, with only model updates being aggregated centrally. Early results show a diagnostic accuracy within 1–2% of centralized training while maintaining HIPAA/GDPR compliance.

5.2.6. O-RAN Trials

Rakuten Mobile (the world's first fully cloud-native O-RAN commercial network, deployed in Japan starting 2020 [100]) and the DISH Network in the United States have deployed commercial O-RAN networks with AI-based RAN Intelligent Controllers (RIC) for traffic prediction, load balancing, and energy optimization. Performance data indicates a 15–20% improvement in spectral efficiency and 10–15% reduction in energy consumption compared to traditional RAN architectures, although the complexity of integration and vendor ecosystem maturity remain challenges [124], as shown in Section 4 (O-RAN Alliance deployment metrics from WG1 use-case documentation).

These real-world deployments provide invaluable insights into practical challenges—including hardware constraints, integration complexity, and operational considerations—that are often underestimated in purely theoretical proposals. They also validate the technical feasibility and performance benefits of distributed intelligence, bridging the gap between vision and implementation for future 6G networks.

5.3. Open Challenges and Future Research Directions

Despite the progress in defining the 6G vision and initial standardization, numerous areas require intensive research to make 6G with distributed AI a practical reality.

5.3.1. Efficient and Scalable Edge Resource Optimization

Algorithms capable of jointly managing heterogeneous resources (communication, computing, storage, energy) in large-scale, dynamic and resource-constrained edge environments in real-time need to be developed [90]. This includes optimization specifically tailored to the deployment and execution of edge LAMs [14].

5.3.2. Trustworthy Distributed AI

The robustness, efficiency, fairness and, critically, the privacy and security of distributed learning paradigms such as FL and SL must be enhanced [78]. Research is needed in areas such as asynchronous FL, handling non-IID data, label-privacy preservation in SL and effective integration of DP, HE, SMC and ZKP in 6G edge environments [78]. The development of frameworks resilient to poisoning and adversarial attacks, and specifically tailored to distributed AI, is also essential [89].

5.3.3. AI-Native Architectures

The definition and validation of 6G network architectures that integrate AI as a foundational component rather than an overlay is required [5]. This includes exploring task-oriented architectures (TONA), intelligent-agent-based designs (PFM/LLM) and architectures supporting intent-based end-to-end orchestration [58].

5.3.4. Holistic Integration and Co-Design

The synergistic integration of communication, computing, sensing and AI should be advanced [101]. Co-design approaches are needed to jointly optimize, for example, data/model compression, task partitioning and communication protocols in scenarios such as Semantic Edge Computing (SEC) or Semantic Communications (SemCom) [59,125–129].

5.3.5. Holistic Security and Privacy

End-to-end security must be addressed across all layers: physical layer security (PLS [89]), network-level security (distributed infrastructure, open interfaces [50]), application-layer protection and, increasingly, the security of the distributed AI systems themselves (data privacy, model integrity) [63].

5.3.6. Sustainability and Energy Efficiency

Techniques to minimize the energy consumption of both the 6G network infrastructure and distributed AI processes (training and inference), which can be highly demanding, should be developed and investigated [58].

5.3.7. Standardization and Ecosystem Development

Achieving global consensus on key technical standards for 6G is essential to avoid market fragmentation [117]. Encouraging open software- and hardware-based ecosystems can accelerate innovation and adoption [96].

It is essential to recognize that these open challenges are not isolated problems. There is a deep interdependence among them. For example, resource optimization [90] depends on trustworthy AI algorithms to make decisions [92], yet trustworthy AI (such as FL/SL) itself requires efficient resource management (communication and computing) to operate effectively [78]. Both, in turn, rely on AI-native architectures that can support them adequately [5]. Security must protect both the network infrastructure and the components of the distributed AI [50]. Sustainability is a cross-cutting constraint influencing architectural choices, AI algorithm design and resource-management strategies [7]. Thus, advancing towards 6G requires a systemic and multidisciplinary approach that addresses these challenges in a coordinated and holistic manner. Table 5 summarizes the current state of standardization across key organizations.

Table 5. Summary of Ongoing 6G standardization Efforts Across Key Organizations.

Organization	Relevant Group/Initiative	Specific Focus	Current Milestone/Status	Key Upcoming Steps
ITU-R	WP 5D/IMT-2030	Global vision, use scenarios, capabilities (incl. AI, ISAC), and spectrum [21]	Rec. M.2160 (Framework) published (Nov 2023) [21]	Definition of technical requirements and evaluation criteria (2024–2027) [21]
3GPP	TSG SA (SA1), TSG RAN (RAN1/2/3/4)	6G requirements (Rel-19), 6G technical studies (Rel-20), and 6G specs (Rel-21) [118]	Rel-19: Initial 6G studies in progress [119]	Rel-20: main technical studies (starting Q3 2025) [74]; Rel-21: Specs (starting ~2027) [118]
3GPP	Various WGs	Infrastructure for AI/ML (data collection, model management) [117]	Ongoing work in Rel-19/20 [117]	Native integration in Rel-21 specifications [117,130]
ETSI	ISG MEC	MEC architecture, APIs, federation, slicing, vertical-industry support [107]	Phase 3 completed, Phase 4 ongoing [107]	Continued evolution to support 6G; alignment with 3GPP SA6 [107]

6. Conclusions

This article set out to provide a comprehensive synthesis of distributed computing and edge AI as foundational enablers of 6G networks, as outlined in the Contributions

subsection of the Introduction. The analysis confirms that these technologies are not merely incremental enhancements but represent a fundamental architectural shift necessary to meet 6G's ambitious performance and capability targets.

The transition towards 6G networks marks a defining inflexion point in the evolution of mobile communications, driven by the vision of a hyperconnected and intelligently integrated world. This article has examined how distributed computing and edge artificial intelligence (edge AI) emerge as fundamental and mutually interdependent technologies for realizing this vision.

6.1. Recapitulation

The analysis highlighted the unavoidable need to migrate from centralized computing paradigms to distributed architectures, with MEC acting as a key enabler for bringing computing and storage capabilities closer to the end user. This proximity is essential for meeting the stringent latency, bandwidth and privacy requirements demanded by the most transformative 6G applications.

In parallel, edge AI consolidates itself as the intelligence engine of these distributed networks. Techniques such as FL and SL have emerged as direct responses to the resource constraints and privacy concerns at the edge, enabling the collaborative training of AI models.

These performance gains must be interpreted in context: the 85–95% latency reduction is achieved under specific conditions—an MEC server within 1 km of user equipment, 5G NR fronthaul, task data sizes of 1–10 MB, and server utilization below 70%. The 99% communication overhead reduction via gradient compression is achieved with top-k sparsification ($k = 1\%$) under IID data distribution, homogeneous device capabilities, and stable channel conditions. Realistic deployments with non-IID data distributions, heterogeneous devices, or degraded wireless channel conditions will yield proportionally lower gains (approximately 60–80% latency reduction and 90–97% communication reduction), as quantitatively analyzed in Sections 2 and 3. These bounds provide guidance for system designers planning 6G edge deployments in diverse operational environments.

The rise in Large AI Models (LAMs) at the edge represents a significant challenge but also an unprecedented opportunity, driving research on extreme optimization and distributed inference architectures. Edge AI not only enables intelligent services for end users but is also crucial for the autonomous optimization and efficient management of the inherent complexity of 6G networks.

The diversity of architectural proposals—evolutionary approaches such as Ericsson's, revolutionary ones such as TONA, open architectures such as O-RAN-based designs, and integrated approaches such as SAGIN and ISEA—reflects the intense exploratory phase in which the field currently finds itself. Nevertheless, all converge on the need for intelligent and automated orchestration, with approaches such as Intent-Based Networking (IBN) emerging as promising solutions to abstract and manage the complex joint optimization of communication, computing and storage resources in this distributed environment.

6.2. Transformative Potential

The synergy between distributed computing and edge AI has the potential to unlock truly revolutionary capabilities. By integrating intelligence natively into the network infrastructure, 6G can move beyond mere data transmission to become a platform for ubiquitous intelligence. This will enable genuinely immersive experiences (XR, the metaverse, holographic communication), advanced automation (Industry 4.0, cooperative autonomous driving), personalized and context-aware services, and new capabilities such as integrated sensing and the provision of Artificial Intelligence as a Service (AIaaS) directly from the

network. The expected benefits include dramatic improvements in performance (latency, capacity), efficiency (energy and spectrum) and user experience quality.

6.3. Need for Continued Research and Standardization

Despite the enthusiasm and significant progress achieved, the full realization of the 6G vision with distributed AI faces formidable challenges. Resource limitations at the edge, ensuring security and privacy in complex distributed systems, managing communication overhead, guaranteeing the robustness and reliability of AI models, securing data quality and availability, and addressing ethical and sustainability considerations all demand ongoing research and innovative solutions.

The path to 6G requires a concerted and collaborative global effort. Standardization within organizations such as ITU-R and 3GPP is essential to ensure interoperability and to create a unified global market; however, the need for early direction must be balanced with the flexibility to incorporate emerging research advances. Fostering open ecosystems and multidisciplinary research will be key to overcoming the remaining challenges and harnessing the full transformative potential offered by distributed computing and edge AI for the next generation of mobile networks.

6.4. Open Research Challenges and Future Directions

Based on the presented analysis, the following five research challenges are identified as critical priorities for realizing AI-native 6G:

1. **Large-Scale Experimental Validation and Testbed Initiatives:** The analytical and theoretical comparisons presented in this work require validation under realistic large-scale conditions. Several existing testbeds provide platforms for empirical validation: (1) POWDER (Platform for Open Wireless Data-driven Experimental Research) at the University of Utah, which provides a city-scale software-defined wireless research platform with O-RAN-compatible hardware for AI-driven RAN optimization and MEC workload experiments; (2) Colosseum (Northeastern University, Boston), the world's largest wireless network emulator with 256 software-defined radios, which enables the large-scale emulation of 6G AI-native scenarios including FL over realistic fading channels; (3) Arena (Northeastern University), a reconfigurable indoor testbed for sub-6 GHz and mmWave experimentation with edge AI capabilities. For simulation-based validation, key frameworks include ns-3 with MEC extensions, SUMO for realistic vehicular mobility models, OpenAirInterface for protocol-level 5G/6G simulation, and SimPy for discrete-event simulation of FL convergence under heterogeneous device conditions. Critical validation gaps that community testbed initiatives should address include: multi-cell scenarios with hundreds of simultaneous AI agents, realistic heterogeneous traffic patterns (XR, V2X, massive IoT combined), and mobility models at 6G target speeds. The 6G community is encouraged to develop standardized benchmarks and open datasets enabling reproducible comparison of AI-native architectures across testbeds.

2. **Wireless-Channel-Aware Federated Learning:** The convergence behaviour of FL algorithms (FedAvg, FedProx, SCAFFOLD) under realistic 6G channel conditions—including Rayleigh/Rician fading, Doppler spreading at high mobility, and packet loss—remains insufficiently characterized. The integration of AirComp with privacy-preserving mechanisms (DP, secure aggregation) is an open problem that requires joint optimization of the noise multiplier σ_{DP} and the channel power allocation. Rigorous experimental validation on O-RAN testbeds (e.g., Colosseum, POWDER) with realistic mobility patterns is needed.

3. **Edge LAM Lifecycle Management:** The deployment of large AI models (LLMs, foundation models) at edge nodes raises unresolved questions about model freshness: how frequently must edge LAMs be fine-tuned to remain relevant under distribution shift?

PEFT/LoRA-based update pipelines must be co-designed with the wireless communication layer to minimize fine-tuning round-trip latency while preserving model quality. The trade-off between model staleness and update communication cost is an open optimization problem with no current closed-form solution.

4. **Trustworthy Distributed AI in Multi-Operator Environments:** As 6G networks involve multiple operators sharing RAN and edge infrastructure (O-RAN disaggregation, network slicing), ensuring trust in AI model updates across administrative domains is critical. Byzantine-resilient FL aggregation algorithms (e.g., Krum, Coordinate-wise Median) must be validated at 6G scale ($N > 10^3$ clients) while maintaining near-RT control loop latency < 10 ms. Lightweight alternatives to Blockchain for cross-domain attestation remain an active research gap.

5. **Energy-Efficient AI-Native RAN:** Current estimates suggest that 6G AI-native features may increase base-station energy consumption by 20–40% compared to 5G NR, primarily due to continuous AI inference at the near-RT RIC and O-DU. Green AI techniques (sparse inference, neural architecture search for energy-constrained platforms, solar-powered edge nodes) must be co-designed with the O-RAN functional split selection to meet the 6G sustainability KPI (energy efficiency improvement of $100\times$ over 5G [7]).

6. **Semantic and Goal-Oriented Communication with Edge AI:** The paradigm of semantic communication—transmitting task-relevant meaning rather than raw bits—requires a fundamental rethinking of the PHY/MAC stack and edge AI co-design. Open questions include how to define and measure semantic relevance for heterogeneous tasks, how to train jointly communication and inference models under distribution shift, and how to standardize semantic representations across vendors and network generations. This direction aligns with the 3GPP Release 19 work items on AI/ML for the air interface and is a key differentiator of 6G from all previous generations.

Author Contributions: Conceptualization, E.A.H. and H.F.B.-O.; methodology, N.C.R.-I.; software, H.F.B.-O.; validation, E.A.H., H.F.B.-O. and N.C.R.-I.; formal analysis, E.A.H.; investigation, E.A.H. and H.F.B.-O.; resources, N.C.R.-I.; data curation, N.C.R.-I.; writing—original draft preparation, E.A.H.; writing—review and editing, H.F.B.-O. and N.C.R.-I.; visualization, H.F.B.-O.; supervision, E.A.H. and N.C.R.-I.; project administration, E.A.H.; funding acquisition, H.F.B.-O. and N.C.R.-I. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Data Availability Statement: Not applicable. This study is based on publicly available and cited literature; no new data were generated or deposited.

Acknowledgments: The authors would like to acknowledge the support of the Telecommunications Research Group (GITUQ) for its contribution to the development and technical discussion of this study. AI-based language assistance tools were used to improve clarity, grammar, and academic writing. All scientific content, analysis, and conclusions are the responsibility of the authors.

Conflicts of Interest: The authors declare no conflicts of interest.

Abbreviations

The following abbreviations are used in this manuscript:

3GPP	Third-Generation Partnership Project
5G	Fifth Generation
6G	Sixth Generation
AI	Artificial Intelligence
API	Application Programming Interface
AR	Augmented Reality
CAPEX	Capital Expenditure

CPS	Cyber–Physical Systems
DL	Deep Learning
DT	Digital Twin
ETSI	European Telecommunications Standards Institute
FL	Federated Learning
IBN	Intent-Based Networking
IMT	International Mobile Telecommunications
IoT	Internet of Things
ISAC	Integrated Sensing and Communication
ITU	International Telecommunication Union
ITU-R	ITU Radiocommunication Sector
JCC	Joint Communication, Computing and Storage
KPI	Key Performance Indicator
LAM	Large AI Model
MARL	Multi-Agent Reinforcement Learning
MEC	Multi-access Edge Computing
ML	Machine Learning
OPEX	Operational Expenditure
O-RAN	Open Radio Access Network
OS-MEC	Open-Source Multi-access Edge Computing
PRISMA	Preferred Reporting Items for Systematic Reviews and Meta-Analyses
QoS	Quality of Service
RAN	Radio Access Network
RL	Reinforcement Learning
SAGIN	Satellite–Terrestrial Integrated Network
SL	Split Learning
Tbps	Terabits per second
TONA	Task-Oriented Native AI Architecture
UAV	Unmanned Aerial Vehicle
V2I	Vehicle-to-Infrastructure
V2V	Vehicle-to-Vehicle
V2X	Vehicle-to-Everything
VR	Virtual Reality
XR	Extended Reality

References

1. Pennanen, H.; Hänninen, T.; Tervo, O.; Tölli, A.; Latva-aho, M. 6G: The Intelligent Network of Everything—A Comprehensive Vision, Survey, and Tutorial. *IEEE Access* **2024**, *13*, 1319–1421. [\[CrossRef\]](#)
2. Chen, S.; Liang, Y.-C.; Sun, S.; Kang, S.; Cheng, W.; Peng, M. Vision, Requirements, and Technology Trends of 6G Wireless Networks. *arXiv* **2020**, arXiv:2002.04929.
3. Gui, G.; Liu, M.; Tang, F.; Kato, N.; Adachi, F. 6G: Opening New Horizons for Integration of Comfort, Security, and Intelligence. *IEEE Wirel. Commun.* **2020**, *27*, 126–132. [\[CrossRef\]](#)
4. Al-Ansi, A.; Al-Ansi, A.M.; Muthanna, A.; Elgendy, I.A.; Koucheryavy, A. Survey on Intelligence Edge Computing in 6G: Characteristics, Challenges, Potential Use Cases, and Market Drivers. *Future Internet* **2021**, *13*, 118. [\[CrossRef\]](#)
5. Yang, Y.; Wu, J.; Chen, T.; Peng, C.; Wang, J.; Deng, J.; Tao, X.; Liu, G.; Li, W.; Yang, L.; et al. Task-Oriented 6G Native-AI Network Architecture. *IEEE Netw.* **2024**, *38*, 219–227. [\[CrossRef\]](#)
6. Liu, R.; Zhang, L.; Li, R.Y.-N.; Di Renzo, M. The ITU Vision and Framework for 6G: Scenarios, Capabilities and Enablers. *arXiv* **2023**, arXiv:2305.13887. [\[CrossRef\]](#)
7. ITU-R. IMT Towards 2030 and Beyond (IMT-2030). Available online: <https://www.itu.int/en/ITU-R/study-groups/rsg5/rwp5d/imt-2030/pages/default.aspx> (accessed on 8 April 2026).
8. Li, P.; Fan, J.; Wu, J. Exploring the Key Technologies and Applications of 6G Wireless Communication Network. *iScience* **2025**, *28*, 112281. [\[CrossRef\]](#) [\[PubMed\]](#)
9. Shamsabadi, A.A.; Yadav, A.; Gadallah, Y.; Yanikomeroglu, H. Exploring the 6G Potentials: Immersive, Hyper-Reliable, and Low-Latency Communication. *arXiv* **2024**, arXiv:2407.11051.

10. Viswanathan, H.; Mogensen, P.E. Communications in the 6G Era. *IEEE Access* **2020**, *8*, 57063–57074. [\[CrossRef\]](#)
11. Saad, W.; Bennis, M.; Chen, M. A Vision of 6G Wireless Systems: Applications, Trends, Technologies, and Open Research Problems. *IEEE Netw.* **2020**, *34*, 134–142. [\[CrossRef\]](#)
12. Chaccour, C.; Soorki, M.N.; Saad, W.; Bennis, M.; Popovski, P.; Debbah, M. Seven Defining Features of Terahertz Wireless Systems: A Fellowship of Communication and Sensing. *IEEE Commun. Surv. Tutor.* **2022**, *24*, 967–993. [\[CrossRef\]](#)
13. Zhou, Z.; Chen, X.; Li, E.; Zeng, L.; Luo, K.; Zhang, J. Edge Intelligence: Paving the Last Mile of Artificial Intelligence with Edge Computing. *Proc. IEEE* **2019**, *107*, 1738–1762. [\[CrossRef\]](#)
14. Yan, X.; Wang, J.; Zhang, X.; Zhang, Y. Data Plane Design for AI-Native 6G Networks. Huawei Technologies, 2025. Available online: <https://www.huawei.com/en/huaweitech/future-technologies/data-plane-design-ai-native-6g-networks> (accessed on 28 May 2026).
15. Wang, Y.; Zhao, J. Mobile Edge Computing, Metaverse, 6G Wireless Communications, Artificial Intelligence, and Blockchain: Survey and Their Convergence. *arXiv* **2022**, arXiv:2209.14147. [\[CrossRef\]](#)
16. Wang, X.; Han, Y.; Leung, V.C.M.; Niyato, D.; Yan, X.; Chen, X. Convergence of Edge Computing and Deep Learning: A Comprehensive Survey. *IEEE Commun. Surv. Tutor.* **2020**, *22*, 869–904. [\[CrossRef\]](#)
17. Letaief, K.B.; Shi, Y.; Lu, J.; Lu, J. Edge Artificial Intelligence for 6G: Vision, Enabling Technologies, and Applications. *IEEE J. Sel. Areas Commun.* **2022**, *40*, 5–36. [\[CrossRef\]](#)
18. Abreha, H.G.; Hayajneh, M.; Serhani, M.A. Federated Learning in Edge Computing: A Systematic Survey. *Sensors* **2022**, *22*, 450. [\[CrossRef\]](#)
19. Chen, X.; Guo, Z.; Wang, X.; Yang, H.H.; Feng, C.; Han, S.; Wang, X.; Quek, T.Q.S. Toward 6G Native-AI Network: Foundation Model Based Cloud-Edge-End Collaboration Framework. *arXiv* **2023**, arXiv:2310.17471. [\[CrossRef\]](#)
20. Kairouz, P.; McMahan, H.B.; Avent, B.; Bellet, A.; Bennis, M.; Bhagoji, A.N.; Bonawitz, K.; Charles, Z.; Cormode, G.; Cummings, R.; et al. Advances and Open Problems in Federated Learning. *Found. Trends Mach. Learn.* **2021**, *14*, 1–210. [\[CrossRef\]](#)
21. Angel, N.A.; Ravindran, D.; Vincent, P.M.D.R.; Srinivasan, K.; Hu, Y.C. Recent Advances in Evolving Computing Paradigms: Cloud, Edge, and Fog Technologies. *Sensors* **2022**, *22*, 196. [\[CrossRef\]](#)
22. Wang, X.; Jia, W. Optimizing Edge AI: A Comprehensive Survey on Data, Model, and System Strategies. *arXiv* **2025**, arXiv:2501.03265.
23. Alimi, I.A.; Patel, R.K.; Zaouga, A.; Muga, N.J.; Xin, Q.; Pinto, A.N.; Monteiro, P.P. Trends in Cloud Computing Paradigms: Fundamental Issues, Recent Advances, and Research Directions toward 6G Fog Networks. In *Moving Broadband Mobile Communications Forward—Intelligent Technologies for 5G and Beyond*; Haidine, A., Ed.; IntechOpen: London, UK, 2021.
24. Gill, S.S. A Manifesto for Modern Fog and Edge Computing: Vision, New Paradigms, Opportunities, and Future Directions. In *Operationalizing Multi-Cloud Environments: Technologies, Tools and Use Cases*; Nagarajan, R., Raj, P., Thirunavukarasu, R., Eds.; Springer: Cham, Switzerland, 2022; pp. 237–253. [\[CrossRef\]](#)
25. Pham, Q.-V.; Fang, F.; Ha, V.N.; Piran, M.J.; Le, M.; Le, L.B.; Hwang, W.-J.; Ding, Z. A Survey of Multi-Access Edge Computing in 5G and Beyond: Fundamentals, Technology Integration, and State-of-the-Art. *arXiv* **2020**, arXiv:1906.08452. [\[CrossRef\]](#)
26. Biswas, A.; Wang, H.C. Autonomous Vehicles Enabled by the Integration of IoT, Edge Intelligence, 5G, and Blockchain. *Sensors* **2023**, *23*, 1963. [\[CrossRef\]](#)
27. Wang, C.X.; You, X.; Gao, X.; Zhu, X.; Li, Z.; Zhang, C.; Wang, H.; Huang, Y.; Chen, Y.; Haas, H.; et al. On the Road to 6G: Visions, Requirements, Key Technologies, and Testbeds. *IEEE Commun. Surv. Tutor.* **2023**, *25*, 905–974. [\[CrossRef\]](#)
28. ETSI. Multi-Access Edge Computing (MEC). Available online: <https://www.etsi.org/technologies/multi-access-edge-computing> (accessed on 8 April 2026).
29. ETSI MEC ISG. MEC Security: Status of Standards Support and Future Evolutions; White Paper No. 46; European Telecommunications Standards Institute (ETSI): Sophia Antipolis, France, 2022. Available online: <https://www.etsi.org/images/files/etsiwhitepapers/etsi-wp-46-2nd-ed-mec-security.pdf> (accessed on 28 April 2026).
30. Verizon Business. 5G Edge. Verizon Business, 2026. Available online: <https://www.verizon.com/business/solutions/5g/edge-computing/> (accessed on 28 April 2026).
31. NTT DOCOMO, Inc. 5G Evolution and 6G; NTT DOCOMO, Inc.: Tokyo, Japan, 2023. Available online: https://www.docomo.ne.jp/english/binary/pdf/corporate/technology/whitepaper_6g/DOCOMO_6G_White_PaperEN_v5.0.pdf (accessed on 28 April 2026).
32. Ishtiaq, M.; Saeed, N.; Khan, M.A. Edge Computing in the Internet of Things: A 6G Perspective. *IT Prof.* **2024**, *26*, 62–70. [\[CrossRef\]](#)
33. Zhao, L.; Zhou, G.; Zheng, G.; Chih-Lin, I.; You, X.; Hanzo, L. Open-Source Multi-Access Edge Computing for 6G: Opportunities and Challenges. *IEEE Access* **2021**, *9*, 158426–158439. [\[CrossRef\]](#)
34. Crespo-Aguado, M.; Lozano, R.; Hernandez-Gobertti, F.; Molner, N.; Gomez-Barquero, D. Flexible Hyper-Distributed IoT-Edge-Cloud Platform for Real-Time Digital Twin Applications on 6G-Intended Testbeds for Logistics and Industry. *Future Internet* **2024**, *16*, 431. [\[CrossRef\]](#)

35. Li, Y.; Gao, X.; Shi, M.; Kang, J.; Niyato, D.; Yang, K. Hierarchical Optimization for Task Execution Cost Minimization in D2D-Assisted Mobile Edge Computing Networks. *IEEE Trans. Wirel. Commun.* **2026**, *25*, 587–601. [CrossRef]
36. Li, Y.; Gao, X.; Zhang, Z.; Yuan, H.; Kang, J.; Niyato, D.; Yang, K. Joint Trajectory, Resource, and Access Optimization in Multi-UAV Collaborative Mobile Edge Computing Networks for Low-Altitude Economy. *IEEE Internet Things J.* **2026**, *13*, 9467–9481. [CrossRef]
37. Matinmikko-Blue, M.; Latva-aho, M.; Ahokangas, P.; Aho, E.; Leppänen, K. *White Paper on Broadband Connectivity in 6G; 6G Flagship*; University of Oulu: Oulu, Finland, 2020; ISBN 978-952-62-2679-8. Available online: <https://oulurepo.oulu.fi/bitstream/handle/10024/36799/isbn978-952-62-2679-8.pdf> (accessed on 28 April 2026).
38. Chang, L.; Zhang, Z.; Li, P.; Xi, S.; Guo, W.; Shen, Y.; Xiong, Z.; Kang, J.; Niyato, D.; Qiao, X.; et al. 6G-Enabled Edge AI for Metaverse: Challenges, Methods, and Future Research Directions. *J. Commun. Inf. Netw.* **2022**, *7*, 9815195. [CrossRef]
39. Karabulut, M.A.; Shah, A.F.M.S.; Pathan, A.-S.K.; Bradford, P.G. IoT-Driven Intelligent Transportation System in the Era of 6G and AI: A Review. *Comput. Mater. Contin.* **2026**, *88*, 077625. [CrossRef]
40. 3GPP. NR; Radio Resource Control (RRC); Protocol Specification; TS 38.331; 3rd Generation Partnership Project (3GPP): Sophia Antipolis, France, 2024. Available online: <https://portal.3gpp.org/desktopmodules/Specifications/SpecificationDetails.aspx?specificationId=3197> (accessed on 28 April 2026).
41. Shi, W.; Cao, J.; Zhang, Q.; Li, Y.; Xu, L. Edge Computing: Vision and Challenges. *IEEE Internet Things J.* **2016**, *3*, 637–646. [CrossRef]
42. Chaurasia, B.K.; Shukla, M.M.; Vishwakarma, V.; Tiwari, B. Transformative Impact of Edge Computing with 6G Network Using Cloud Computing. *J. Cloud Comput.* **2026**, *15*, 49. [CrossRef]
43. Gdali, Z. *Edge Computing vs. Cloud: Latency Impact*; Firecell: Nice, France, 2026. Available online: <https://firecell.io/edge-computing-vs-cloud-latency-impact/> (accessed on 28 April 2026).
44. Products—Architecting for End-to-End Low Latency in Wireless Networks White Paper. Available online: <https://www.cisco.com/c/en/us/products/collateral/wireless/arch-end-to-end-low-latency-wireless-networks.html> (accessed on 29 May 2026).
45. Jayanth, R.; Gupta, N.; Prasanna, V. Benchmarking Edge AI Platforms for High-Performance ML Inference. *arXiv* **2024**, arXiv:2409.14803. [CrossRef]
46. Gupta, R.; Reebadiya, D.; Tanwar, S. 6G-Enabled Edge Intelligence for Ultra -Reliable Low Latency Applications: Vision and Mission. *Comput. Stand. Interfaces* **2021**, *77*, 103521. [CrossRef]
47. Ibrahim, H.; Elmowafy, M. Optimizing Video Streaming over 6G Networks: Codec Adaptation, Quality Metrics, and AI-Driven Topologies. In Proceedings of the International Telecommunications Conference (ITC-Egypt 2025), Cairo, Egypt, 28–31 July 2025; pp. 272–277. [CrossRef]
48. Goel, M. Video Compression: Reduce File Size. Available online: <https://theproductguy.in/blogs/video-compression-guide/> (accessed on 29 May 2026).
49. Rassouli, B.; Gündüz, D. Information-Theoretic Privacy-Preserving Schemes Based on Perfect Privacy. *arXiv* **2023**, arXiv:2301.11754. [CrossRef]
50. Rifa-Pous, H.; Garcia-Font, V.; Nunez-Gomez, C.; Salas, J. Security, Trust and Privacy Challenges in AI-Driven 6G Networks. In Proceedings of the 11th International Conference on Computer Science, Engineering and Information Technology, London, UK, 27–28 July 2024. [CrossRef]
51. Blanchard, P.; El Mhamdi, E.M.; Guerraoui, R.; Stainer, J. Machine Learning with Adversaries: Byzantine Tolerant Gradient Descent. In *Proceedings of the 31st Conference on Neural Information Processing Systems (NeurIPS 2017)*, Long Beach, CA, USA, 4–9 December 2017; Curran Associates, Inc.: Red Hook, NY, USA, 2017; Volume 30.
52. Fang, M.; Cao, X.; Jia, J.; Gong, N. Local Model Poisoning Attacks to Byzantine-Robust Federated Learning. In *Proceedings of the 29th USENIX Security Symposium (USENIX Security 20)*; USENIX Association: Berkeley, CA, USA, 2020; pp. 1605–1622.
53. Zhang, R.; Sun, J.; Zhang, P. Towards Certified Probabilistic Robustness with High Accuracy. *arXiv* **2023**, arXiv:2309.00879. [CrossRef]
54. Lovén, L.; Leppänen, T.; Peltonen, E.; Partala, J. EdgeAI: A Vision for Distributed, Edge-Native Artificial Intelligence in Future 6G Networks. Available online: <http://5gtn.fi> (accessed on 8 April 2026).
55. Gkonis, P.K.; Giannopoulos, A.; Nomikos, N.; Trakadas, P.; Sarakis, L.; Masip-Bruin, X. A Survey on Architectural Approaches for 6G Networks: Implementation Challenges, Current Trends, and Future Directions. *Telecom* **2025**, *6*, 27. [CrossRef]
56. Löwenstein, U. ITU-R Status Update on WRC and IMT-2030. Presented at the one6G Summit 2025, Valencia, Spain, 5–6 September 2025. Available online: https://summit2025.one6g.org/wp-content/uploads/Session-1_Uwe-Loewenstein_ITU-R-Status-update-on-WRC-and-IMT-2030.pdf (accessed on 28 May 2026).
57. Sanjalawe, Y.; Fraihat, S.; Al-E'mari, S.; Abualhaj, M.; Makhadmeh, S.; Alzubi, E. A Review of 6G and AI Convergence: Enhancing Communication Networks with Artificial Intelligence. *IEEE Open J. Commun. Soc.* **2025**, *6*, 2308–2355. [CrossRef]
58. Talwar, S.; Himayat, N.; Nikopour, H.; Xue, F.; Wu, G.; Ilderem, V. 6G: Connectivity in the Era of Distributed Intelligence. *arXiv* **2021**, arXiv:2110.07052. [CrossRef]

59. Zhang, M.; Abdi, M.; Dasari, V.R.; Restuccia, F. Semantic Edge Computing and Semantic Communications in 6G Networks: A Unifying Survey and Research Challenges. *arXiv* **2024**, arXiv:2411.18199. [\[CrossRef\]](#)
60. Ericsson. *6G Network Architecture: A Proposal for Early Alignment*; Ericsson: Stockholm, Sweden, 2023. Available online: <https://www.ngmn.org/publications/network-architecture-evolution-towards-6g.html> (accessed on 8 April 2026).
61. Watson, C.; Woods, K.; Shyy, D.J. *6G and Artificial Intelligence and Machine Learning*; MITRE Corporation: McLean, VA, USA, 2021. Available online: <https://www.mitre.org/sites/default/files/2021-11/pr-21-0214-6g-and-artificial-intelligence-and-machine-learning.pdf> (accessed on 8 April 2026).
62. McMahan, B.; Moore, E.; Ramage, D.; Hampson, S.; Arcas, B.A. Communication-Efficient Learning of Deep Networks from Decentralized Data. In Proceedings of the 20th International Conference on Artificial Intelligence and Statistics (AISTATS), Fort Lauderdale, FL, USA, 20–22 April 2017; pp. 1273–1282. Available online: <https://proceedings.mlr.press/v54/mcmahan17a.html> (accessed on 8 April 2026).
63. Zhu, F.; Wang, X.; Li, X.; Zhang, M.; Chen, Y.; Huang, C.; Yang, Z.; Chen, X.; Zhang, Z.; Jin, R.; et al. Wireless Large AI Model: Shaping the AI-Native Future of 6G and Beyond. *arXiv* **2025**, arXiv:2504.14653.
64. Chen, M.; Poor, H.V.; Saad, W.; Cui, S. Convergence Time Optimization for Federated Learning over Wireless Networks. *IEEE Trans. Wirel. Commun.* **2021**, *20*, 2457–2471. [\[CrossRef\]](#)
65. Soni, M. Energy-Efficient Deep Learning Architectures for IoT Devices. *Int. J. Adv. Res. Multidiscip. Trends* **2025**, *2*, 865–882. Available online: <https://www.ijarmt.com/index.php/j/article/view/309> (accessed on 8 April 2026).
66. Mohammadi Amiri, M.; Gündüz, D. Machine Learning at the Wireless Edge: Distributed Stochastic Gradient Descent over the Air. *IEEE Trans. Signal Process.* **2020**, *68*, 2155–2169. [\[CrossRef\]](#)
67. Sifaou, H.; Li, G.Y. Robust Federated Learning via Over-the-Air Computation. In Proceedings of the 2022 IEEE 32nd International Workshop on Machine Learning for Signal Processing (MLSP), Xi'an, China, 22–25 August 2022; pp. 1–6. [\[CrossRef\]](#)
68. Amiri, M.M.; Gündüz, D.; Kulkarni, S.R.; Poor, H.V. Convergence of Update-Aware Device Scheduling for Federated Learning at the Wireless Edge. *IEEE Trans. Wirel. Commun.* **2021**, *20*, 3643–3658. [\[CrossRef\]](#)
69. Azimi-Abarghouyi, S.M.; Tassiulas, L. Over-the-Air Federated Learning via Weighted Aggregation. *IEEE Trans. Wirel. Commun.* **2024**, *23*, 18240–18253. [\[CrossRef\]](#)
70. Marfo, W.; Tosh, D.K.; Moore, S.V.; Suetterlein, J.; Manzano, J. Reducing Communication Overhead in Federated Learning for Network Anomaly Detection with Adaptive Client Selection. In Proceedings of the 2025 IEEE 25th International Symposium on Cluster, Cloud and Internet Computing Workshops (CCGridW), Tromsø, Norway, 19–22 May 2025; pp. 1–9.
71. Mamond, A.W.; Kundroo, M.; Kim, T. Autoencoder-Based Decentralized Federated Learning for Efficient Communication. *Comput. Netw.* **2025**, *272*, 111676. [\[CrossRef\]](#)
72. Wang, S.; Liu, J.; Xu, H.; Tang, C.; Ma, Q.; Huang, L. Toward Communication-Efficient Decentralized Federated Graph Learning Over Non-IID Data. *IEEE Trans. Mob. Comput.* **2026**, *25*, 6929–6947. [\[CrossRef\]](#)
73. He, Z.; Zhu, G.; Zhang, S.; Luo, E.; Zhao, Y. FedDT: A Communication-Efficient Federated Learning via Knowledge Distillation and Ternary Compression. *Electronics* **2025**, *14*, 2183. [\[CrossRef\]](#)
74. Wang, L.; Wang, W.; Li, B. CMFL: Mitigating Communication Overhead for Federated Learning. In Proceedings of the 2019 IEEE 39th International Conference on Distributed Computing Systems (ICDCS), Dallas, TX, USA, 7–10 July 2019; pp. 954–964.
75. Itahara, S.; Nishio, T.; Koda, Y.; Morikura, M.; Yamamoto, K. Distillation-Based Semi-Supervised Federated Learning for Communication-Efficient Collaborative Training with Non-IID Private Data. *IEEE Trans. Mob. Comput.* **2023**, *22*, 191–205. [\[CrossRef\]](#)
76. Al-Saedi, A.A.; Boeva, V.; Casalicchio, E. FedCO: Communication-Efficient Federated Learning via Clustering Optimization. *Future Internet* **2022**, *14*, 377. [\[CrossRef\]](#)
77. Wang, Y.; Lin, L.; Chen, J. Communication-Efficient Adaptive Federated Learning. *arXiv* **2023**, arXiv:2205.02719. [\[CrossRef\]](#)
78. Lin, Z.; Qu, G.; Chen, X.; Huang, K. Split Learning in 6G Edge Networks. *arXiv* **2023**, arXiv:2306.12194. [\[CrossRef\]](#)
79. Lin, Y.; Han, S.; Mao, H.; Wang, Y.; Dally, B. Deep Gradient Compression: Reducing the Communication Bandwidth for Distributed Training. In Proceedings of the International Conference on Learning Representations, Vancouver, BC, Canada, 30 April–3 May 2018.
80. Alistarh, D.; Grubic, D.; Li, J.; Tomioka, R.; Vojnovic, M. QSGD: Communication-Efficient SGD via Gradient Quantization and Encoding. In *Proceedings of the Advances in Neural Information Processing Systems*; Guyon, I., Luxburg, U.V., Bengio, S., Wallach, H., Fergus, R., Vishwanathan, S., Garnett, R., Eds.; Curran Associates, Inc.: Red Hook, NY, USA, 2017; Volume 30.
81. Cao, X.; Zhu, G.; Xu, J.; Huang, K. Optimized Power Control for Over-the-Air Computation in Fading Channels. *IEEE Trans. Wirel. Commun.* **2020**, *19*, 7498–7513. [\[CrossRef\]](#)
82. Hu, E.J.; Shen, Y.; Wallis, P.; Allen-Zhu, Z.; Li, Y.; Wang, S.; Wang, L.; Chen, W. LoRA: Low-Rank Adaptation of Large Language Models. 2021. Available online: <https://openreview.net/forum?id=nZeVKeeFYf9> (accessed on 8 April 2026).
83. Lokwon, K. How On-Device AI Could Help Us to Cut AI's Energy Demand. Available online: <https://www.weforum.org/stories/2025/03/on-device-ai-energy-system-chatgpt-grok-deepx/> (accessed on 30 May 2026).

84. Chai, S. *How Edge Computing Can Solve AI's Energy Crisis*; Built In: Chicago, IL, USA, 2025.
85. Hariharasubramanian, C.S.; Shrikanth, B. *Edge AI Models Explained: Balancing Accuracy, Power, and Thermal Limits*; ZEDED: San Jose, CA, USA, 2026.
86. Gholami, A.; Kim, S.; Dong, Z.; Yao, Z.; Mahoney, M.W.; Keutzer, K. A Survey of Quantization Methods for Efficient Neural Network Inference. In *Low-Power Computer Vision*; Chapman and Hall/CRC: Boca Raton, FL, USA, 2021.
87. He, Y.; Liu, P.; Wang, Z.; Hu, Z.; Yang, Y. Filter Pruning via Geometric Median for Deep Convolutional Neural Networks Acceleration. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Long Beach, CA, USA, 15–20 June 2019; pp. 4340–4349.
88. NVIDIA Jetson Benchmarks. Available online: <https://developer.nvidia.com/embedded/jetson-benchmarks> (accessed on 30 May 2026).
89. Chou, H.; Solanki, S.; Ha, V.N.; Chen, L.; Ma, S.L.; Al-Hraishawi, H.; Eappen, G.; Chatzinotas, S. Edge AI-Enabled Physical Layer Security for 6G NTN: Potential Threats and Future Opportunities. *arXiv* **2023**, arXiv:2401.01005.
90. Alhashimi, H.F.; Hindia, M.N.; Dimyati, K.; Hanafi, E.B.; Alden, F.Z.; Qamar, F.; Nguyen, Q.N. Survey on AI-Enabled Resource Management for 6G Heterogeneous Networks: Recent Research, Challenges, and Future Trends. *Comput. Mater. Contin.* **2025**, *83*, 3585–3622. [CrossRef]
91. Cui, Q.; You, X.; Wei, N.; Nan, G.; Zhang, X.; Zhang, J.; Lyu, X.; Ai, M.; Tao, X.; Feng, Z.; et al. Overview of AI and Communication for 6G Network: Fundamentals, Challenges, and Future Research Opportunities. *Sci. China Inf. Sci.* **2025**, *68*, 1–61. [CrossRef]
92. Yang, H.; Alphones, A.; Xiong, Z.; Niyato, D.; Zhao, J.; Wu, K. Artificial Intelligence-Enabled Intelligent 6G Networks. *IEEE Netw.* **2020**, *34*, 272–280. [CrossRef]
93. Mao, Q.; Hu, F.; Hao, Q. Deep Learning for Intelligent Wireless Networks: A Comprehensive Survey. *IEEE Commun. Surv. Tutor.* **2018**, *20*, 2595–2621. [CrossRef]
94. El-haryqy, N.; Madini, Z.; Zouine, Y. A Review of Deep Learning Techniques for Enhancing Spectrum Sensing and Prediction in Cognitive Radio Systems: Approaches, Datasets, and Challenges. *Int. J. Comput. Appl.* **2024**, *46*, 1104–1128. [CrossRef]
95. Xie, X.; Ning, X.; Liu, Y.; Wang, H.; Jin, J.; Yang, H. LLM4FB: A One-Sided CSI Feedback and Prediction Framework for Lightweight UEs via Large Language Models. *Sensors* **2026**, *26*, 691. [CrossRef] [PubMed]
96. Nokia. *Unlocking the Full Potential of AI-Native 6G Through Standards*; Nokia: Espoo, Finland, 2025. Available online: <https://www.nokia.com/6g/unlocking-the-full-potential-of-ai-native-6g-through-standards/> (accessed on 8 April 2026).
97. Wang, Z.; Shi, Y.; Zhou, Y.; Zhu, J.; Letaief, K.B. Edge Large AI Models: Revolutionizing 6G Networks. *arXiv* **2025**, arXiv:2505.00321. [CrossRef]
98. Dalai, D.; Babu, S.; Manoj, B.S. Satellite-6G Network Integration Roadmap on Reference Architectures. *TechRxiv* **2023**. [CrossRef]
99. Tang, Y.; Srinivasan, U.C.; Scott, B.J.; Umealor, O.; Kevogo, D.; Guo, W. End-to-End Edge AI Service Provisioning Framework in 6G O-RAN. *arXiv* **2025**, arXiv:2503.11933.
100. Rakuten Mobile. *Rakuten Mobile Network White Paper: World's First Fully Virtualized Cloud-Native Mobile Network*; Rakuten Mobile: Tokyo, Japan, 2020. Available online: <https://corp.mobile.rakuten.co.jp/english/innovation/cloud-network/> (accessed on 8 April 2026).
101. Liu, Z.; Chen, X.; Wu, H.; Wang, Z.; Chen, X.; Niyato, D.; Huang, K. Integrated Sensing and Edge AI: Realizing Intelligent Perception in 6G. *arXiv* **2025**, arXiv:2501.06726. [CrossRef]
102. 3GPP. *Solutions for NR to Support Non-Terrestrial Networks (NTN)*; TR 38.821, Release 16; 3rd Generation Partnership Project (3GPP): Sophia Antipolis, France, 2019. Available online: <https://portal.3gpp.org/desktopmodules/Specifications/SpecificationDetails.aspx?specificationId=3525> (accessed on 28 May 2026).
103. Del Portillo, I.; Cameron, B.G.; Crawley, E.F. A Technical Comparison of Three Low Earth Orbit Satellite Constellation Systems to Provide Global Broadband. *Acta Astronaut.* **2019**, *159*, 123–135. [CrossRef]
104. Shehabi, A.; Smith, S.J.; Sartor, D.A.; Brown, R.E.; Herrlin, M.; Koomey, J.G.; Masanet, E.R.; Horner, N.; Azevedo, I.L.; Lintner, W. *United States Data Center Energy Usage Report*; Lawrence Berkeley National Laboratory: Berkeley, CA, USA, 2016.
105. Masanet, E.; Shehabi, A.; Lei, N.; Smith, S.; Koomey, J. Recalibrating Global Data Center Energy-Use Estimates. *Science* **2020**, *367*, 984–986. [CrossRef]
106. Hussain, S.; Ozturk, M. WS-01: AI-Enabled Network Orchestration: Design Challenges and Opportunities for 6G Networks. In Proceedings of the IEEE Wireless Communications and Networking Conference—2023 IEEE WCNC, Scotland, UK, 26–29 March 2023.
107. Chen, Q.; Guo, Z.; Meng, W.; Han, S.; Li, C.; Quek, T.Q.S. A Survey on Resource Management in Joint Communication and Computing-Embedded SAGIN. *IEEE Commun. Surv. Tutor.* **2024**, *27*, 1911–1954. [CrossRef]
108. Mekrache, A.; Ksentini, A.; Verikoukis, C. Intent-Based Management of Next-Generation Networks: An LLM-Centric Approach. *IEEE Netw.* **2025**, *38*, 29–36. [CrossRef]

109. Boutouchent, A.; Mekrache, A.; Ksentini, A.; Adhane, G.; Fonseca, J.P.C.D.; McNamara, J.; Ramantas, K.; Palena, M.; Iordache, M.; Cigno, R.L.; et al. 6G-INTENSE: Intent-Driven Native Artificial Intelligence Architecture Supporting Network-Compute Abstraction and Sensing at the Deep Edge. *IEEE Veh. Technol. Mag.* **2025**, *20*, 44–54. [\[CrossRef\]](#)
110. Li, Y.; Huang, Y.; Feng, J.; Deng, C.; Liu, C.; Chau, V.; Wang, W. Multiagent Reinforcement Learning Based on Structural Coordination. *Tsinghua Sci. Technol.* **2025**. [\[CrossRef\]](#)
111. Calzolari, G.; Sumathy, V.; Kanellakis, C.; Nikolakopoulos, G. Safe Heterogeneous Multi-Agent Reinforcement Learning with Communication Regularization for Coordinated Target Acquisition. *arXiv* **2026**, arXiv:2601.08327. [\[CrossRef\]](#)
112. Lu, Y.; Huang, X.; Zhang, K.; Maharjan, S.; Zhang, Y. Low-Latency Federated Learning and Blockchain for Edge Association in Digital Twin Empowered 6G Networks. *IEEE Trans. Ind. Inform.* **2021**, *17*, 5098–5107. [\[CrossRef\]](#)
113. Dang, S.; Amin, O.; Shihada, B.; Alouini, M.S. What Should 6G Be? *Nat. Electron.* **2020**, *3*, 20–29. [\[CrossRef\]](#)
114. ETSI. *Multi-Access Edge Computing (MEC); Technical Requirements*; ETSI GS MEC 002 V2.1.1; European Telecommunications Standards Institute: Sophia Antipolis, France, 2018. Available online: https://www.etsi.org/deliver/etsi_gs/MEC/001_099/002/02.01.01_60/gs_MEC002v020101p.pdf (accessed on 28 May 2026).
115. 3GPP. *Study on Scenarios and Requirements for Next Generation Access Technologies*; TR 38.913, Release 14; 3rd Generation Partnership Project (3GPP): Sophia Antipolis, France, 2017. Available online: <https://portal.3gpp.org/desktopmodules/Specifications/SpecificationDetails.aspx?specificationId=2996> (accessed on 28 May 2026).
116. ETSI. *Multi-Access Edge Computing (MEC); Framework and Reference Architecture*; ETSI GS MEC 003 V2.1.1; European Telecommunications Standards Institute: Sophia Antipolis, France, 2019. Available online: https://www.etsi.org/deliver/etsi_gs/mec/001_099/003/02.01.01_60/gs_mec003v020101p.pdf (accessed on 28 May 2026).
117. ITU-R. *Recommendation ITU-R M.2160-0: Framework and Overall Objectives of the Future Development of IMT for 2030 and Beyond*; International Telecommunication Union (ITU): Geneva, Switzerland, 2023. Available online: <https://www.itu.int/rec/R-REC-M.2160/en> (accessed on 28 May 2026).
118. Chen, W. *RAN Rel-19 Status and a Look Beyond 2025*; 3GPP: Sophia Antipolis, France, 2025. Available online: <https://www.3gpp.org/technologies/ran-rel-19> (accessed on 8 April 2026).
119. 3GPP. *Overview of AI/ML Related Work in 3GPP*; 3rd Generation Partnership Project (3GPP): Sophia Antipolis, France, 2025. Available online: <https://www.3gpp.org/news-events/3gpp-news/ai-ml-2025> (accessed on 8 April 2026).
120. Larsson, D.C.; Grönlén, A.; Parkvall, S.; Liberg, O. *6G Standardization Timeline and Principles*; Ericsson: Stockholm, Sweden, 2024.
121. 3GPP. *Study on Artificial Intelligence (AI)/Machine Learning (ML) for NR Air Interface*; TR 38.843, Release 18; 3rd Generation Partnership Project (3GPP): Sophia Antipolis, France, 2023. Available online: <https://portal.3gpp.org/desktopmodules/Specifications/SpecificationDetails.aspx?specificationId=3983> (accessed on 28 May 2026).
122. Mehmet, T. *Industrial 5G Devices—Architecture and Capabilities*; 5G Alliance for Connected Industries and Automation (5G-ACIA): Hessen, Germany, 2023. Available online: <https://5g-acia.org/whitepapers/industrial-5g-devices-architecture-and-capabilities/> (accessed on 28 May 2026).
123. 5G Automotive Association (5GAA). *5GAA Roadmap: Cross-Industry Cooperation in the Automotive and Wireless Industry*; 5G Automotive Association: München, Germany, 2019.
124. O-RAN Alliance. *O-RAN Working Group 1: Use Cases and Overall Architecture*; O-RAN.WG1.OAD v10.0; O-RAN Alliance. 2023. Available online: <https://www.o-ran.org/specifications> (accessed on 28 May 2026).
125. Xie, H.; Qin, Z.; Li, G.Y.; Juang, B.-H. Deep Learning Enabled Semantic Communication Systems. *IEEE Trans. Signal Process.* **2021**, *69*, 2663–2675. [\[CrossRef\]](#)
126. Xu, W.; Yang, Z.; Ng, D.W.K.; Levorato, M.; Eldar, Y.C.; Debbah, M. Edge Learning for B5G Networks with Distributed Signal Processing: Semantic Communication, Edge Computing, and Wireless Sensing. *IEEE J. Sel. Top. Signal Process.* **2023**, *17*, 9–39. [\[CrossRef\]](#)
127. Lan, Q.; Wen, D.; Zhang, Z.; Zeng, Q.; Chen, X.; Popovski, P.; Huang, K. What Is Semantic Communication? A View on Conveying Meaning in the Era of Machine Intelligence. *J. Commun. Inf. Netw.* **2021**, *6*, 336–371. [\[CrossRef\]](#)
128. Li, T.; Sahu, A.K.; Talwalkar, A.; Smith, V. Federated Learning: Challenges, Methods, and Future Directions. *IEEE Signal Process. Mag.* **2020**, *37*, 50–60. [\[CrossRef\]](#)
129. Liang, F.; Yu, W.; Liu, X.; Griffith, D.; Golmie, N. Toward Edge-Based Deep Learning in Industrial Internet of Things. *IEEE Internet Things J.* **2020**, *7*, 4329–4341. [\[CrossRef\]](#)
130. Letaief, K.B.; Chen, W.; Shi, Y.; Zhang, J.; Zhang, Y.-J.A. The Roadmap to 6G: AI-Empowered Wireless Networks. *IEEE Commun. Mag.* **2019**, *57*, 84–90. [\[CrossRef\]](#)

Disclaimer/Publisher’s Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.