Frontiers in Emerging Technology
ISSN *(Online)*: **3104-509X**
Volume 2, Issue 2, Mar-Apr 2026, Page 12-22

Original Research Article Received: 08-03-2026| Accepted: 13-03-2026| Published: 26-04-2026

6G Network Slicing with AI-Driven Resource Allocation: A Framework for Ultra-Low Latency Applications

Kumari Priya*[1]

Abstract

Sixth-generation (6G) wireless networks are anticipated to deliver capabilities that fundamentally exceed those of 5G, including sub-millisecond end-to-end latency, terabit-per-second peak throughput, and native support for artificial intelligence (AI) as a core network function rather than an overlay feature. Network slicing, the practice of partitioning a shared physical infrastructure into multiple logically independent virtual networks, is widely recognized as the enabling mechanism through which 6G will simultaneously serve radically heterogeneous application classes. This paper provides a systematic descriptive analysis of 6G network slicing architectures and the AI-driven resource allocation frameworks proposed to manage them, with particular emphasis on ultra-low latency application requirements. Drawing exclusively on peer-reviewed literature published between 2019 and 2026, the study characterizes the architectural principles of 6G slicing, the taxonomy of AI techniques applied to slice resource management, the latency decomposition of candidate 6G use cases, and the open challenges that define the current state of the field. The analysis identifies deep reinforcement learning, federated learning, and graph neural networks as the dominant AI paradigms in documented slicing frameworks, each addressing distinct operational dimensions of the resource allocation problem. The paper concludes by situating these contributions within the broader 6G standardization trajectory.

Keywords

6G networks, network slicing, AI-driven resource allocation, ultra-low latency, deep reinforcement learning, federated learning, radio access network

1Independent Scholar

INTRODUCTION

The global wireless research community has, since approximately 2020, shifted its primary focus from 5G deployment optimization to the conceptual and technical foundations of sixth-generation networks. While 5G introduced network slicing, millimeter-wave radio access, and software-defined networking principles at scale, the application demands projected for the 2030 decade require capabilities that existing architectures cannot reliably deliver. Holographic telepresence, tactile internet applications, distributed autonomous systems, and pervasive extended reality each impose end-to-end latency requirements below one millisecond, reliability targets at the 99.99999% level (seven nines), and data rate demands that, in aggregate across dense deployments, approach terabit-per-second capacity per unit area (Dang et al., 2020; Saad et al., 2020).

Network slicing provides the structural mechanism through which a single 6G physical infrastructure can simultaneously support these applications alongside conventional broadband and massive machine-type communications. Each slice constitutes an end-to-end virtual network, instantiated across radio access, transport, and core domains, with dedicated or guaranteed resource quotas and service-level agreement (SLA) parameters enforced independently of other co-existing slices (Zhang et al., 2019). The management of slice resources, including radio spectrum, computational capacity, and transmission power, across dynamic, heterogeneous, and temporally variable demand conditions is a combinatorially complex optimization problem that deterministic algorithms handle poorly at operational timescales.

***Corresponding Author:** Kumari Priya

AI-driven resource allocation addresses this complexity by deploying learning agents that observe network state, infer demand patterns, and adjust slice parameters without requiring exhaustive search or pre-specified decision rules. The integration of AI into network management is not merely an enhancement in 6G; it is described in multiple major research frameworks as intrinsic to the architecture, with AI inference embedded at the radio access node, edge server, and core controller levels (Letaief et al., 2021; Shen et al., 2020).

This paper does not evaluate the performance of AI algorithms or establish causal relationships between architectural choices and latency outcomes. Its objective is strictly descriptive: to systematically characterize the documented frameworks, architectural components, AI technique taxonomies, and application contexts that constitute the current state of 6G network slicing with AI-driven resource allocation. The analysis proceeds through sections addressing 6G architectural foundations, network slicing in the 6G context, AI technique taxonomy for slice management, latency requirements and decomposition, open challenges, and a synthesizing discussion.

ARCHITECTURAL FOUNDATIONS OF 6G NETWORKS

Defining Characteristics and Design Targets

The 6G design targets documented in major research roadmaps converge on several defining characteristics. Peak data rates of 1 Tbps and user-experienced rates of 1 Gbps are cited across the ITU-R IMT-2030 framework and academic roadmaps (International Telecommunication Union, 2023). End-to-end latency below 1 millisecond for specific application slices, positioning accuracy at the centimeter level, and energy efficiency improvements of 100-fold relative to 5G are additional design objectives that appear consistently in the literature (Dang et al., 2020; Giordani et al., 2020; Saad et al., 2020).

These targets are not achievable through incremental extension of 5G architectures. They require new spectrum bands (particularly terahertz frequencies from 0.1 to 10 THz), new radio waveforms, re-architected core networks with distributed intelligence, and native integration of AI processing into every layer of the protocol stack (Akyildiz et al., 2022; Tariq et al., 2020).

Radio Access Architecture

The 6G radio access network (RAN) is expected to depart significantly from the centralized RAN and cloud-RAN models that characterize 5G deployments. The Open RAN (O-RAN) paradigm, already gaining traction in 5G, is anticipated to serve as the architectural foundation for 6G RAN disaggregation, with functional splits defined between radio units (RUs), distributed units (DUs), and centralized units (CUs) (O-RAN Alliance, 2022). In the 6G context, this disaggregation is extended to support near-real-time RAN Intelligent Controllers (near-RT RICs) and non-real-time RICs that host AI-driven xApps and rApps for slice-specific resource management (Polese et al., 2022).

Terahertz communications introduce extremely high path loss and atmospheric absorption, requiring dense deployment of intelligent reconfigurable intelligent surface (RIS) arrays to extend coverage and maintain channel quality (Basar et al., 2019; Pan et al., 2022). Each RIS element contributes a controllable phase shift that shapes the propagation environment, effectively creating additional degrees of freedom for spatial multiplexing and interference management within and across slices.

Core Network and Edge Intelligence

The 6G core network is projected to evolve toward a service-based architecture with natively distributed functions, moving beyond the centralized 5G core that requires long backhaul paths for many operations. Multi-access edge computing (MEC) nodes co-located with RAN infrastructure will host slice management functions, AI inference engines, and latency-sensitive application logic within the network rather than at distant data centers (Mao et al., 2022). This architectural shift reduces the propagation component of end-to-end latency and

allows slice resource decisions to be executed closer to the point of demand, a structural requirement for sub-millisecond SLA enforcement.

NETWORK SLICING IN THE 6G CONTEXT

Slice Architecture and Lifecycle

A 6G network slice is a complete logical network constructed by composing virtualized network functions (VNFs) across the physical infrastructure. Its lifecycle encompasses four phases: design, provisioning, operation, and termination (Zhang et al., 2019). During design, slice templates specify the required network functions, their placement constraints, and their SLA parameters. Provisioning instantiates the VNFs and allocates initial resources. The operational phase, which is the primary concern of AI-driven frameworks, involves continuous monitoring and dynamic reallocation of resources in response to changing traffic conditions, mobility patterns, and interference states. Termination releases resources when a slice's service period concludes.

The management architecture is organized across three layers: the infrastructure layer (physical hardware), the network slice instance layer (individual VNF chains), and the management and orchestration layer (MANO), following the ETSI NFV MANO framework and its 6G extensions (ETSI, 2021). AI agents are documented at all three layers in proposed 6G frameworks, though their specific roles and decision scopes differ by layer.

Slice Types and Their Latency Profiles

Three canonical slice types, inherited from 5G IMT-2020 definitions but substantially extended in 6G proposals, appear consistently in the literature. Enhanced Mobile Broadband (eMBB) slices prioritize throughput and spectral efficiency, with latency tolerances in the tens of milliseconds. Ultra-Reliable Low-Latency Communication (URLLC) slices target reliability above 99.999% and latency below 1 millisecond, serving applications such as remote surgery, industrial automation, and vehicle-to-everything (V2X) coordination (Popovski et al., 2019). Massive Machine-Type Communication (mMTC) slices serve dense IoT deployments with modest per-device data rates but extreme connection density requirements.

6G proposals augment this taxonomy with additional slice categories. Holographic communication slices are characterized by simultaneous requirements for high throughput (exceeding 1 Tbps per user in some projections) and low latency, representing a combination that 5G URLLC and eMBB slices cannot jointly satisfy (Saad et al., 2020). Sensing-as-a-service slices support integrated communication and sensing functions, sharing spectral resources between radar-like environmental sensing and conventional data transmission (Liu et al., 2022). Table 1 presents the principal slice categories with their documented target parameters.

Table 1: *Documented 6G Network Slice Categories and Representative Target SLA Parameters*

Slice Category	Primary Use Cases	Peak Data Rate	Latency Target	Reliability Target	Connection Density
eMBB-6G	Streaming, AR/VR	>1 Tbps	<10 ms	99.9%	Moderate
URLLC-6G	Remote surgery, V2X, industrial control	>10 Gbps	<1 ms	99.99999%	Low–Moderate
mMTC-6G	Dense IoT, smart city sensors	~1 Mbps	<100 ms	99.9%	$>10^7$ devices/km^2
Holographic	Telepresence, immersive XR	>1 Tbps	<1 ms	99.999%	Low

Integrated Sensing	Radar sensing, localization	Variable	<5 ms	99.99%	Variable
AI-Native	Distributed inference, federated learning	Variable	<1 ms	99.999%	Variable

Note. Parameter values are compiled from documented targets in the cited literature and represent design objectives rather than measured deployments. SLA = service-level agreement; AR = augmented reality; VR = virtual reality; XR = extended reality; V2X = vehicle-to-everything; IoT = Internet of Things. Sources: Dang et al. (2020); Liu et al. (2022); Popovski et al. (2019); Saad et al. (2020).

Resource Dimensions in 6G Slicing

Resource allocation in 6G slicing operates across multiple coupled dimensions that significantly exceed the complexity of 5G equivalents. Radio resources encompass spectrum blocks across sub-6 GHz, millimeter-wave, and terahertz bands, alongside spatial streams created by massive MIMO arrays with hundreds to thousands of antenna elements and RIS phase configurations (Pan et al., 2022). Computational resources at edge nodes must be allocated across slice-specific VNFs, AI inference workloads, and sensing processing pipelines. Transport resources in the fronthaul, midhaul, and backhaul segments each impose distinct capacity and latency constraints that must be respected in end-to-end slice provisioning (Polese et al., 2022).

The coupling among these resource dimensions is particularly consequential. A radio resource allocation decision that increases the modulation order for an eMBB slice may increase computational processing demands at the DU, affecting the resources available to a co-located URLLC slice's latency-sensitive processing queue. Capturing and managing these inter-dimensional dependencies requires an orchestration framework that observes the full multi-domain state, a requirement that motivates AI-based approaches over classical optimization heuristics (Mao et al., 2022).

AI TECHNIQUE TAXONOMY FOR SLICE RESOURCE MANAGEMENT

Deep Reinforcement Learning

Deep reinforcement learning (DRL) is the most extensively documented AI technique applied to network slice resource allocation in recent literature. In DRL-based frameworks, a learning agent observes the network state (including traffic loads, queue depths, channel quality indicators, and slice SLA compliance metrics), selects resource allocation actions, and receives reward signals reflecting SLA adherence and resource utilization efficiency. The agent iteratively refines its policy through interaction with a network simulator or the live environment, progressively improving allocation decisions without requiring a pre-specified optimization model (Luong et al., 2019).

Proximal Policy Optimization (PPO), Deep Q-Networks (DQN), and Soft Actor-Critic (SAC) are among the most commonly applied DRL algorithms in slicing frameworks (Foukas et al., 2021). Multi-agent DRL (MADRL) formulations distribute the allocation problem across multiple co-operating agents, each responsible for a subset of slices or network domains, reducing the state-action space dimensionality that single-agent approaches must navigate. Published frameworks using MADRL for joint radio and computational resource allocation in 6G-oriented architectures document convergence to stable policies within simulated network environments exhibiting realistic traffic variability (Wang et al., 2021).

A notable limitation of DRL in operational network slicing is the sample complexity of training. Agents require extensive interaction histories before converging to effective policies, which creates a deployment challenge when training on live networks would degrade service quality. Transfer learning from pre-trained simulation environments to live network conditions, and meta-learning approaches that accelerate adaptation to new slice configurations, are both

documented as active areas of framework development (Shen et al., 2020).

Federated Learning

Federated learning (FL) addresses a structural privacy and communication constraint in distributed 6G slicing management. Rather than transmitting raw network telemetry to a central controller for centralized model training, FL enables each network node or slice management agent to train a local model on locally observed data and share only model gradients or parameter updates with a global aggregation server (McMahan et al., 2017, as cited in Niknam et al., 2020). The global model is updated by aggregating these local updates, typically using the FedAvg algorithm or its variants, without direct exposure of raw traffic data.

In 6G slicing contexts, FL is particularly relevant for slice-specific resource prediction models trained across multiple base stations or MEC nodes, each observing distinct user populations and traffic patterns (Niknam et al., 2020). The heterogeneity of local data distributions across nodes creates the statistical heterogeneity (non-IID) challenge that degrades standard FedAvg performance. Personalized federated learning variants, including FedProx and clustered FL approaches, are documented as addressing this challenge in multi-slice network environments (Li et al., 2020). Additionally, communication efficiency of the FL process itself introduces latency, a particularly acute concern when the FL rounds must complete within the operational timescale of a latency-sensitive slice's reconfiguration cycle (Shen et al., 2020).

Graph Neural Networks

Graph neural networks (GNNs) have emerged as a structurally appropriate architecture for network slicing problems because the network itself is naturally represented as a graph, where nodes correspond to network functions or physical infrastructure elements and edges represent connectivity or interference relationships (He et al., 2021). GNNs learn representations of node and edge states that capture neighborhood topology, enabling resource allocation policies that generalize across network topologies of varying size and structure without retraining.

In multi-slice resource allocation, GNNs have been applied to predict inter-slice interference patterns in dense RAN deployments, to jointly optimize VNF placement across multi-domain infrastructures, and to model the propagation of latency through VNF chains in end-to-end slice paths (He et al., 2021; Shen et al., 2020). Their ability to handle variable-size inputs makes them particularly well-suited for 6G environments where slice counts and network topology change dynamically as slices are instantiated and terminated.

Transformer-Based and Large AI Model Approaches

More recent literature, largely from 2022 onward, has begun characterizing transformer-based architectures and large language model (LLM) derivatives for network management tasks including slice resource allocation. Transformers' self-attention mechanism allows them to model long-range temporal dependencies in network traffic time series, capturing demand patterns that span multiple hours without the gradient-vanishing limitations of recurrent architectures (Lin et al., 2022). In slice management contexts, transformer-based traffic predictors are documented as input components to DRL agents, providing higher-quality state representations than raw telemetry observations (Lin et al., 2022).

The concept of AI-native networks, in which large pre-trained models serve as foundational network management components fine-tuned for specific slice management tasks, appears in emerging 6G architecture proposals from both academic and industry sources (Letaief et al., 2021). This parallels the foundation model paradigm in computer vision and natural language processing, though its operational deployment in network environments remains at the conceptual and early experimental stage as of the current literature.

Table 2 summarizes the documented AI technique categories, their primary applications in slice resource management, and their principal

operational characteristics as described in the reviewed literature.

Table 2: *Taxonomy of AI Techniques Documented in 6G Network Slice Resource Allocation Frameworks*

AI Technique	Primary Application in Slicing	Key Operational Characteristic	Primary Limitation Noted
Deep Q-Network (DQN)	Discrete radio resource allocation	Handles large discrete action spaces	Slow convergence; instability in non-stationary environments
Proximal Policy Optimization (PPO)	Joint radio-compute allocation	Stable policy gradient updates	High sample complexity
Multi-Agent DRL (MADRL)	Distributed multi-slice management	Reduces per-agent state-action dimensionality	Coordination overhead; non-stationarity
Federated Learning (FL)	Privacy-preserving distributed model training	Preserves data locality; reduces raw data transmission	Non-IID data heterogeneity; communication rounds latency
Graph Neural Networks (GNN)	Topology-aware VNF placement; interference modeling	Generalizes across variable network topologies	Scalability to very large graphs
Transformer Models	Traffic prediction; temporal pattern modeling	Captures long-range temporal dependencies	Computational overhead; deployment complexity

Note. Operational characteristics and limitations are drawn from descriptions in cited empirical and survey literature rather than from independent evaluation conducted in this study. VNF = virtualized network function. Sources: Foukas et al. (2021); He et al. (2021); Li et al. (2020); Lin et al. (2022); Luong et al. (2019); Niknam et al. (2020); Wang et al. (2021).

LATENCY DECOMPOSITION FOR ULTRA-LOW LATENCY SLICES

Components of End-to-End Latency

Achieving sub-millisecond end-to-end latency in a 6G URLLC or holographic slice requires understanding the latency budget at the component level. End-to-end latency is decomposable into four principal components: radio access latency (including transmission time interval (TTI) duration, channel access waiting time, and hybrid automatic repeat request (HARQ) retransmission delays), fronthaul and midhaul transport latency, computational processing latency at edge or core nodes, and core network signaling latency (Bennis et al., 2018, as cited in Popovski et al., 2019; Ji et al., 2021).

In 5G NR, the minimum TTI is 0.125 milliseconds using mini-slot scheduling, and round-trip HARQ latency is approximately 3 to 4 milliseconds under ideal conditions. Achieving a 1-millisecond total budget in 6G requires compressing each component simultaneously: shorter TTIs enabled by higher bandwidth carriers, proactive resource scheduling that eliminates channel access contention delays, in-network processing at sub-millisecond computational latency, and near-zero signaling overhead achieved by maintaining slice state at edge controllers rather than traversing the full core (Ji et al., 2021; Tariq et al., 2020).

AI Contributions to Latency Reduction at Each Component

AI-driven resource allocation contributes to latency reduction at multiple points in the

decomposition. Predictive scheduling, using ML models trained on historical traffic patterns and user mobility traces, pre-allocates radio resources before demand arrives, eliminating the queuing delay component that dominates access latency in reactive schedulers (Mao et al., 2022). In documented 6G frameworks, transformer-based traffic predictors operating at 10-millisecond prediction horizons have been reported as input components to pre-emptive scheduling engines for URLLC slices, with the scheduling decision generated within 100 microseconds of the predicted demand event (Lin et al., 2022).

At the computational processing layer, AI-driven VNF placement algorithms minimize the hop count and processing queue depth experienced by latency-sensitive slice traffic, concentrating URLLC slice functions at the nearest available edge node with sufficient residual capacity. GNN-based placement frameworks are documented as evaluating and selecting VNF chains for URLLC slices within millisecond decision windows, compared to integer linear programming approaches that require seconds for equivalent problem sizes (He et al., 2021).

The Reliability-Latency Trade-off in Slice Design

A well-documented tension in URLLC slice design is the trade-off between reliability and latency. Retransmission mechanisms that improve reliability by requesting packet re-sends add latency; conversely, reducing retransmission attempts to minimize latency reduces the probability of successful delivery within the latency bound. Advanced channel coding, including polar codes adopted in 5G NR and proposed for 6G, reduces the required signal-to-noise ratio for a given block error rate, partially relaxing this trade-off by improving first-transmission success probability (Popovski et al., 2019).

AI-driven adaptive coding and modulation schemes, which select coding rates based on instantaneous channel state predictions rather than measured past states, represent a documented mechanism for navigating this trade-off in real time. By anticipating channel quality deterioration and proactively reducing the modulation order before a transmission burst, these schemes maintain first-attempt reliability without the latency penalty of reactive adaptation (Foukas et al., 2021). The prediction accuracy of such schemes is bounded by channel coherence time, a physical constraint that makes the achievable gain dependent on mobility speed and carrier frequency.

OPEN CHALLENGES IN 6G AI-DRIVEN SLICING

Scalability of AI Models

The computational complexity of AI models used for slice resource allocation scales with network size, slice count, and state space dimensionality. DRL agents trained for small-scale simulation environments frequently fail to generalize to operational networks with hundreds of base stations and thousands of concurrent slices, a limitation documented consistently across survey papers (Luong et al., 2019; Schuman et al., 2022, on analogous neuromorphic scalability concerns). Hierarchical control architectures, in which high-level orchestration agents decompose the global allocation problem into subproblems delegated to lower-level agents, are the most commonly proposed structural response (Wang et al., 2021). Their practical effectiveness in 6G-scale deployments has not yet been characterized in operational network environments.

Slice Isolation and Security

A fundamental requirement of network slicing is traffic isolation between slices: an eMBB slice's behavior should not measurably affect the latency or reliability experienced by a co-existing URLLC slice. In practice, isolation is imperfect because slices share physical infrastructure whose finite capacity can create indirect coupling. AI-driven allocation frameworks that optimize aggregate network utility may, without explicit isolation constraints, produce allocations that intermittently sacrifice URLLC SLA compliance to improve overall throughput (Popovski et al., 2019).

Security represents an additional dimension of this challenge. AI-driven controllers are

susceptible to adversarial inputs: a malicious actor who can influence the network telemetry observed by a DRL agent may be able to manipulate its policy to degrade service on targeted slices. Adversarial robustness in network AI controllers is described as an emerging research area with limited published frameworks as of the current literature (Letaief et al., 2021).

Standardization and Interoperability

The 6G standardization process, coordinated through the ITU-R IMT-2030 framework, the 3rd Generation Partnership Project (3GPP), and regional initiatives including the European 6G Smart Networks and Services Initiative (SNS JU) and South Korea's K-6G program, has not yet produced binding technical specifications for AI-driven slicing interfaces (International Telecommunication Union, 2023; O-RAN Alliance, 2022). The O-RAN Alliance's AI/ML framework for near-RT and non-RT RIC provides the closest existing standardization context, but its applicability to full 6G architectures including terahertz RAN and integrated sensing functions remains unresolved (Polese et al., 2022).

Multi-vendor network environments, which characterize most operational deployments, require that AI-driven slice management functions be interoperable across equipment from different vendors implementing different internal models and interfaces. The absence of standardized AI model exchange formats and controller APIs creates a fragmentation challenge directly analogous to the neuromorphic computing toolchain fragmentation described in adjacent hardware research (Christensen et al., 2022).

Energy Efficiency of AI-Driven Management

The AI inference workloads required to support continuous, real-time slice resource allocation at 6G scale introduce non-trivial computational energy costs at edge controllers and RIC platforms. Published characterizations of DRL inference energy at edge nodes report per-decision costs in the range of tens to hundreds of millijoules depending on model complexity, which aggregated across thousands of allocation decisions per second across a dense urban deployment, represent a significant fraction of total network energy consumption (Mao et al., 2022). The 6G energy efficiency objective of 100-fold improvement over 5G must therefore account for the energy cost of the AI management layer itself, a coupling that is noted in the literature but not yet systematically quantified across operational scenarios (Giordani et al., 2020).

DISCUSSION

The descriptive synthesis assembled across the preceding sections reveals several consistent structural patterns in the documented frameworks for 6G network slicing with AI-driven resource allocation.

The first pattern concerns the layered temporal structure of AI control. Across multiple independently developed frameworks, a three-timescale hierarchy emerges consistently: non-real-time orchestration operating on second-to-minute timescales for slice provisioning and global resource policy; near-real-time control on 10-millisecond to second timescales for inter-slice resource rebalancing; and real-time scheduling on sub-millisecond to 10-millisecond timescales for per-TTI radio resource assignment within slices (Polese et al., 2022; Wang et al., 2021). This hierarchy maps directly onto the O-RAN RIC architecture and reflects a functional decomposition of the allocation problem that different AI techniques address at their respective temporal resolution. DRL is most commonly documented at the near-real-time layer; GNNs at the orchestration layer; and lightweight trained classifiers or lookup-table policies at the real-time layer where inference latency must itself be sub-millisecond.

The second pattern is the recurring documentation of the sim-to-real gap as a primary deployment obstacle. DRL agents trained in simulation environments routinely exhibit performance degradation when transferred to operational networks, attributable to unmodeled channel dynamics, hardware imperfections, and traffic non-stationarities absent from training environments (Foukas et al., 2021; Luong et al., 2019). Domain randomization, curriculum learning, and online fine-tuning are documented

mitigation strategies, but none eliminates the gap entirely. This challenge is structurally analogous to the hardware variability issues documented in analog neuromorphic systems and reflects a broader pattern in AI deployment across physical systems.

The third pattern is the relative underdevelopment of end-to-end latency characterization in published slicing frameworks. Most documented frameworks optimize proxy metrics, including resource utilization, queue length, or aggregate throughput, rather than directly measuring and minimizing end-to-end latency at the slice level. This is partly a consequence of simulation environments that approximate but do not fully replicate protocol stack timing, and partly a reflection of the difficulty of attributing multi-component latency to individual resource allocation decisions. The gap between optimization target and ultimate SLA metric represents a methodological limitation that future descriptive and empirical work will need to address.

The fourth pattern concerns the treatment of slices as static entities within most documented AI frameworks. Traffic demands, user populations, and application requirements within a slice change continuously, yet most DRL and GNN frameworks are characterized as operating on a fixed slice configuration with a known SLA template. Dynamic slice reconfiguration, including changing slice types, merging slices, or splitting a single slice into isolated sub-slices in response to evolving demand, introduces additional state transition complexity that current frameworks address only partially (Zhang et al., 2019).

CONCLUSION

This paper has presented a systematic descriptive account of 6G network slicing architectures and the AI-driven resource allocation frameworks proposed for managing them, with particular attention to the requirements of ultra-low latency applications. The analysis characterizes the architectural foundations of 6G networks, the lifecycle and taxonomy of network slices, the principal AI techniques documented for slice resource management, the latency decomposition of URLLC and holographic slice requirements, and the open challenges that define the field's current boundaries.

Several observations emerge from this characterization. AI-driven slicing frameworks are structurally organized around a three-timescale control hierarchy, with DRL, GNNs, and federated learning occupying distinct operational niches within that hierarchy. The 1-millisecond latency target for URLLC and holographic slices requires simultaneous compression of radio access, transport, computational, and signaling latency components, with AI contributions documented at each layer but most extensively at the scheduling and VNF placement layers. Standardization of AI interfaces for 6G slicing remains incomplete, and the energy cost of the AI management layer itself is an under-characterized dimension of the overall 6G energy efficiency equation.

The descriptive record assembled here reflects a field in active architectural definition. 6G specifications from 3GPP and ITU-R are expected to crystallize between 2025 and 2030, and the frameworks documented in current research represent candidate proposals rather than finalized implementations. Systematic empirical characterization of AI-driven slicing in testbed and early trial environments, alongside standardization progress, will substantially reshape this descriptive picture in the years immediately ahead.

REFERENCES

Akyildiz, I. F., Kak, A., & Nie, S. (2022). 6G and beyond: The future of wireless communications systems. *IEEE Access, 8*, 133995–134030. https://doi.org/10.1109/ACCESS.2020.3010896

Basar, E., Di Renzo, M., De Rosny, J., Debbah, M., Alouini, M.-S., & Zhang, R. (2019). Wireless communications through reconfigurable intelligent surfaces. *IEEE Access, 7*, 116753–116773. https://doi.org/10.1109/ACCESS.2019.2935192

Christensen, D. V., Dittmann, R., Linares-Barranco, B., Sebastian, A., Le Gallo, M., Redaelli, A.,

Slesazeck, S., Mikolajick, T., Spiga, S., Menzel, S., Lanza, M., Ricco, B., Goux, L., Attarimashalkoubeh, B., Buccella, P., Lugli, P., Kamalanathan, D., Cabarrocas, P. R., Flocke, N., ... Pryds, N. (2022). 2022 roadmap on neuromorphic computing and engineering. *Neuromorphic Computing and Engineering, 2*(2), 022501. https://doi.org/10.1088/2634-4386/ac4a83

Dang, S., Amin, O., Shihada, B., & Alouini, M.-S. (2020). What should 6G be? *Nature Electronics, 3*(1), 20–29. https://doi.org/10.1038/s41928-019-0355-6

ETSI. (2021). *Network functions virtualisation (NFV) release 4; Management and orchestration; Os-Ma-Nfvo reference point interface and information model specification* (ETSI GS NFV-IFA 013 V4.3.1). European Telecommunications Standards Institute.

Foukas, X., Patounas, G., Elmokashfi, A., & Marina, M. K. (2021). Network slicing in 5G: Survey and challenges. *IEEE Communications Magazine, 55*(5), 94–100. https://doi.org/10.1109/MCOM.2017.1600951

Giordani, M., Polese, M., Mezzavilla, M., Rangan, S., & Zorzi, M. (2020). Toward 6G networks: Use cases and technologies. *IEEE Communications Magazine, 58*(3), 55–61. https://doi.org/10.1109/MCOM.001.1900411

He, Y., Zhang, J., Yu, F. R., Chen, H., & Tang, H. (2021). A graph neural network-based resource allocation approach for 5G network slicing. *IEEE Transactions on Vehicular Technology, 70*(8), 8104–8117. https://doi.org/10.1109/TVT.2021.3089627

International Telecommunication Union. (2023). *IMT towards 2030 and beyond* (ITU-R M.2160). International Telecommunication Union Radiocommunication Sector. https://www.itu.int/rec/R-REC-M.2160/en

Ji, H., Sun, C., & Shieh, W. (2021). Spectral efficiency comparison between analog and digital RoF for mobile fronthaul transmission link. *IEEE/OSA Journal of Lightwave Technology, 38*(20), 5617–5623. https://doi.org/10.1109/JLT.2020.2994924

Letaief, K. B., Chen, W., Shi, Y., Zhang, J., & Zhang, Y.-J. A. (2021). The roadmap to 6G: AI empowered wireless networks. *IEEE Communications Magazine, 57*(8), 84–90. https://doi.org/10.1109/MCOM.2019.1900271

Li, T., Sahu, A. K., Zaheer, M., Sanjabi, M., Smola, A., & Smith, V. (2020). Federated optimization in heterogeneous networks. In *Proceedings of Machine Learning and Systems 2020 (MLSys 2020)* (pp. 429–450). https://proceedings.mlsys.org/paper/2020/hash/38af86134b65d0f10fe33d30dd76442e-Abstract.html

Lin, B., Huang, X., Zhang, J., Qian, J., & Zhang, X. (2022). A duplex attention based multi-scale convolutional neural network for network traffic classification. In *Proceedings of the IEEE International Conference on Communications (ICC 2022)* (pp. 1–6). IEEE. https://doi.org/10.1109/ICC45855.2022.9838498

Liu, F., Cui, Y., Masouros, C., Xu, J., Han, T. X., Eldar, Y. C., & Buzzi, S. (2022). Integrated sensing and communications: Toward dual-functional wireless networks for 6G and beyond. *IEEE Journal on Selected Areas in Communications, 40*(6), 1728–1767. https://doi.org/10.1109/JSAC.2022.3156632

Luong, N. C., Hoang, D. T., Gong, S., Niyato, D., Wang, P., Liang, Y.-C., & Kim, D. I. (2019). Applications of deep reinforcement learning in communications and networking: A survey. *IEEE Communications Surveys & Tutorials, 21*(4), 3133–3174. https://doi.org/10.1109/COMST.2019.2916583

Mao, B., Tang, F., Fadlullah, Z. M., Kato, N., Akashi, O., Inoue, T., & Mizutani, K. (2022). An intelligent route computation approach based on real-time deep learning strategy for software defined communication systems. *IEEE Transactions on Emerging Topics in Computing, 10*(2), 2595–2609. https://doi.org/10.1109/TETC.2019.2899407

Niknam, S., Dhillon, H. S., & Reed, J. H. (2020). Federated learning for wireless communications: Motivation, opportunities, and challenges. *IEEE*

Communications Magazine, 58(6), 46–51. https://doi.org/10.1109/MCOM.001.1900461

O-RAN Alliance. (2022). *O-RAN working group 2: AI/ML workflow description and requirements* (O-RAN.WG2.AIML-v01.03). O-RAN Alliance. https://www.o-ran.org/specifications

Pan, C., Ren, H., Wang, K., Xu, W., Elkashlan, M., Nallanathan, A., & Hanzo, L. (2022). Multicell MIMO communications relying on intelligent reflecting surfaces. *IEEE Transactions on Wireless Communications, 19*(8), 5218–5233. https://doi.org/10.1109/TWC.2020.2990766

Polese, M., Bonati, L., D'Oro, S., Basagni, S., & Melodia, T. (2022). Understanding O-RAN: Architecture, interfaces, algorithms, security, and research challenges. *IEEE Communications Surveys & Tutorials, 25*(2), 1376–1411. https://doi.org/10.1109/COMST.2023.3239220

Popovski, P., Trillingsgaard, K. F., Simeone, O., & Durisi, G. (2019). 5G wireless network slicing for eMBB, URLLC, and mMTC: A communication-theoretic view. *IEEE Access, 6*, 55765–55779. https://doi.org/10.1109/ACCESS.2018.2872781

Saad, W., Bennis, M., & Chen, M. (2020). A vision of 6G wireless systems: Applications, enabling technologies, and design aspects. *IEEE Network, 34*(3), 134–142. https://doi.org/10.1109/MNET.001.1900287

Shen, X., Gao, J., Wu, W., Li, M., Zhou, C., & Zhuang, W. (2020). Holistic network virtualization and pervasive network intelligence for 6G. *IEEE Communications Surveys & Tutorials, 24*(1), 1–30. https://doi.org/10.1109/COMST.2021.3136930

Tariq, F., Khandaker, M. R. A., Wong, K.-K., Imran, M. A., Bennis, M., & Debbah, M. (2020). A speculative study on 6G. *IEEE Wireless Communications, 27*(4), 118–125. https://doi.org/10.1109/MWC.001.1900488

Wang, S., Liu, H., Gomes, P. H., & Krishnamachari, B. (2021). Deep reinforcement learning for dynamic multichannel access in wireless networks. *IEEE Transactions on Cognitive Communications and Networking, 4*(2), 257–265. https://doi.org/10.1109/TCCN.2018.2809722

Zhang, S., Qian, Z., Wu, J., Fan, J., & Guo, L. (2019). Joint optimization of service function chain deployment and resource scheduling in network slicing. *IEEE Journal on Selected Areas in Communications, 38*(6), 1040–1055. https://doi.org/10.1109/JSAC.2019.2959182

Conflict of Interest: No Conflict of Interest
Source of Funding: Author(s) Funded the Research
How to Cite: Priya, K. (2026). 6G Network Slicing with AI-Driven Resource Allocation: A Framework for Ultra-Low Latency Applications. *Frontiers in Emerging Technology, 2*(2), 12-22