SCIENCE CHINA
Information Sciences

• POSITION PAPER •

August 2026, Vol. 69, Iss. 8, 183301:1–183301:21
https://doi.org/10.1007/s11432-025-4865-x

CIS-RAN: evolving C-RAN towards cooperative, intelligent, and service-based RAN for 6G

Xiaoyun WANG[1], Nan LI[1,2*], Qi SUN[1], Yang WANG[1], Yiwei YAN[1], Sen WANG[1], Xiaofei XU[1], Na LI[1] & Jing JIN[1]

[1]*China Mobile Research Institute, Beijing 100053, China*
[2]*State Key Laboratory of Networking and Switching Technology, School of Information and Communication Engineering, Beijing University of Posts and Telecommunications, Beijing 100876, China*

Received 30 September 2025/Revised 9 February 2026/Accepted 11 March 2026/Published online 8 July 2026

Abstract The evolution toward sixth-generation (6G) mobile networks imposes unprecedented demands on radio access networks (RANs), such as extreme connectivity, ultra-low latency, high reliability, heterogeneous service support, and AI integration. To meet these challenges, this paper proposes CIS-RAN, a novel 6G RAN architecture built on three pillars of cooperative, intelligent, and service-based design. CIS-RAN enables flexible collaboration across distributed RAN nodes and multi-dimensional domains, integrates intelligence throughout the network, and supports advanced RAN capability exposure as services. To systematically analyze the architectural design, a unified modeling and optimization framework is introduced to abstract the intrinsic coupling among cooperation, intelligence, and service provisioning, providing a theoretical foundation for architectural trade-offs. Five key enabling technologies support this design, encompassing efficient heterogeneous data handling, task-driven coordinated processing, hierarchical intelligence, ultra-low-latency AI service provisioning, and unified RAN capability exposure. A case study on collaborative interference prediction demonstrates practical benefits, achieving a 36%–50% reduction in fine-tuning data requirements and 57% faster convergence compared with standalone training. In addition, a prototype-based evaluation further validates the superiority of low latency AI service provisioning, which reduces the AI service latency by 80%. Finally, practical deployment challenges are discussed, outlining directions for future research toward intelligent and adaptive 6G networks.

Keywords 6G RAN architecture, C-RAN evolution, cooperative, intelligent, service-based design

Citation Wang X Y, Li N, Sun Q, et al. CIS-RAN: evolving C-RAN towards cooperative, intelligent, and service-based RAN for 6G. Sci China Inf Sci, 2026, 69(8): 183301, https://doi.org/10.1007/s11432-025-4865-x

1 Introduction

As a cornerstone of wireless communication systems, the radio access network (RAN) plays a critical role in technological evolution, with its architecture directly shaping network performance, service quality, and cost. Rising mobile data traffic, increasingly diverse service demands, cost pressures, and the pursuit of energy efficiency have accelerated RAN innovation in recent years.

To cope with rapid traffic growth and escalating infrastructure costs, the centralized, collaborative, cloud, and clean RAN (C-RAN), proposed in 2009, offers a pioneering architecture to handle traffic surges, enhance energy efficiency, and reduce deployment costs. C-RAN centralizes baseband processing resources into a single pool, enabling dynamic management and on-demand allocation, thereby allowing mobile network operators (MNOs) to enhance scalability, cost-efficiency, and overall performance. Specifically, C-RAN consolidates baseband processing resources into centralized baseband unit (BBU) pools, which are connected to remote radio units (RRUs) via high-capacity transport networks. This centralized processing supports statistical multiplexing, simplifies site infrastructure, and provides a foundation for network-wide coordination. Building on this, cloudification and virtualization introduce dynamic resource allocation and "soft" BBUs, where baseband functions run on general-purpose hardware through network functions virtualization (NFV) or software-defined networking (SDN) technologies, improving flexibility, reducing capital and operational expenditures, and enabling rapid service deployment. Furthermore, the centralized architecture facilitates advanced cooperative techniques, such as coordinated multi-point (CoMP) transmission and interference management, thereby enhancing spectral efficiency and cell-edge performance.

* Corresponding author (email: linan@chinamobile.com)

C-RAN has been highly successful and has become a foundational architecture for modern mobile networks. Its efficiency, scalability, and cost-effectiveness have enabled widespread deployment in long term evolution (LTE) and fifth generation (5G) by leading operators and vendors such as Nokia, SK Telecom, Korea Telecom, and NTT DOCOMO. However, the forthcoming sixth generation (6G) era will demand far more from the RAN. In particular, 6G is expected to support ubiquitous intelligence, immersive experiences, and pervasive sensing, which introduce a broad spectrum of emerging applications such as autonomous driving, industrial automation, extended reality, and large-scale digital twins. These applications impose stringent and heterogeneous requirements on the network, including not only extreme throughput, ultra-low latency, and unprecedented reliability, but also seamless connectivity, massive access, and consistent service delivery across highly dynamic environments. Meeting such requirements and supporting diverse new services call for highly dynamic, context-aware resource management and real-time adaptability. Conventional C-RAN, designed primarily to guarantee quality of service (QoS) through fixed rules and passive optimization, cannot meet these challenges, and this limitation motivates the evolution toward more advanced RAN architectures for 6G.

To address these challenges, both researchers and industry have proposed several innovative paradigms, which can be broadly categorized into service-oriented architectural redesign and capability-specific integration approaches. In detail, the former is exemplified by the service-based architecture RAN (SBA-RAN), which decomposes monolithic functions into microservices that can be dynamically orchestrated and elastically scaled, thereby enabling flexible network slicing, containerized deployment, and vendor-agnostic interoperability [1–3]. Integration-focused approaches enhance RAN with emerging technologies to support specialized services. A prominent trend is artificial intelligence-enabled RAN, which not only leverages AI/ML techniques to enhance RAN intelligence, but also exploits RAN resources to support AI services and embraces AI-native design principles for joint evolution with RAN functions [4, 5]. Similarly, computing-and-communication integrated RAN incorporates computing resources and related functions into the RAN, supporting edge computing and enabling joint optimization of communication and computing resources. This design benefits latency-sensitive applications such as augmented reality and autonomous systems, while also providing a foundation for advanced AI-driven optimization beyond the capabilities of standalone AI-RAN [6, 7]. Moreover, sensing-and-communication integrated RAN paradigms have attracted considerable attention for 6G and beyond. Hydra-RAN, a dual-functional architecture proposed in [8], integrates radar, lidar, and visual sensing with conventional communication functions to create perceptive networks with environmental awareness and proactive optimization. This integration demonstrates strong synergy as sensing data enhance beam management, handover prediction, and blockage avoidance, while communication infrastructure enables large-scale distributed sensing across urban environments.

While these architectures enhance modularity and functional integration, they generally target only a subset of capabilities or partial function fusion, without offering a holistic design for 6G RAN. Moreover, they lack a unified framework that enables on-demand deployment of intelligent functions capable of jointly optimizing communication, sensing, computing, and service requirements.

In view of these considerations, we propose CIS-RAN, a cooperative, intelligent, and service-based RAN architecture, designed from a holistic perspective. Building on the C-RAN foundation, CIS-RAN establishes a hierarchical and cooperative framework that enables dynamic multi-node collaboration across communication, sensing, and AI domains, surpassing the limitations of traditional C-RAN. The architecture integrates native intelligence across enhanced RRUs, BBUs, and a newly introduced centralized RAN AI unit (CRAU) to support AI-driven network optimization and AI-native applications. Its service-based design allows flexible capability exposure and programmable service orchestration through standardized interfaces, enabling multi-dimensional resources to be composed dynamically according to service demands. This design is enabled by technologies for efficient data collection and transmission, task-driven coordinated data processing, hierarchical intelligence, ultra-low latency AI services provision and assurance, and unified capability exposure, supporting scalable, flexible, and AI-native RAN operations.

The key contributions of this work are threefold.

- A comprehensive analysis of 6G RAN requirements is presented, along with a holistic identification of the fundamental limitations inherent in existing C-RAN architectures when supporting emerging 6G scenarios.
- A novel RAN architecture, termed CIS-RAN, is proposed, with a detailed description of its core components and functions, along with interfaces impact and the theoretical modeling of RAN design trade-offs that inform RAN decisions. Five key enabling technologies of CIS-RAN are also provided, encompassing heterogeneous data collection, task-driven coordinated processing, hierarchical intelligent distribution, ultra-low-latency AI service provisioning, and RAN capability exposure.
- Extensive simulations and prototype experiments are conducted to validate CIS-RAN from both AI4RAN and RAN4AI perspectives, including an uplink interference prediction simulation and indoor-outdoor prototype trials of

AI visual detection services. The former demonstrates the advantages of CIS-RAN-enabled cooperative intelligence in improving learning efficiency and system-level throughput under diverse interference conditions, while the latter verifies the superiority of CIS-RAN in ultra-low-latency AI service provisioning and RAN capability exposure.

The remainder of this paper is organized as follows. Section 2 reviews the limitations of current RANs and introduces the key features and provides a detailed description of the CIS-RAN architecture. Section 3 describes the enabling technologies supporting its operation. Section 4 presents two case studies demonstrating the effectiveness of CIS-RAN from both AI4RAN and RAN4AI perspectives. Section 5 discusses potential future research directions, and Section 6 concludes the paper.

2 From C-RAN to CIS-RAN: RAN evolution for 6G

2.1 Challenges of current RAN

As wireless communications evolve toward 6G, RAN architectures face unprecedented challenges in meeting future demands across performance, functionality, service delivery, and resource efficiency, collectively necessitating a fundamental redesign of RAN.

As outlined in Section 1, diverse capabilities beyond communication are expected to be integrated into the RAN in future 6G systems, which drives an architectural evolution toward multi-domain convergence. In particular, native intelligence is critical to enable predictive optimization, autonomous configuration, and proactive fault management, thereby supporting the stringent performance requirements of 6G. At the same time, 6G is expected to support diverse AI-driven applications, such as embodied intelligence and highly autonomous driving, which are computationally intensive, latency-sensitive and often require real-time interaction with the environment through continuous analytics. Meeting these service demands necessitates seamless integration of communication and computing, deterministic orchestration of heterogeneous resources, and strict end-to-end service assurance. Further, integrated sensing and communication (ISAC) extends this paradigm by jointly supporting radio sensing, positioning, and environmental perception alongside data transmission, increasing the complexity of cross-domain optimization. Collectively, these developments require a fundamental redefinition of the RAN as an integrated platform with multi-dimensional capabilities, far surpassing architectures designed primarily for connectivity.

The above multi-domain evolution also amplifies service delivery complexity, rendering current static, peak-provisioned resource allocation approaches fundamentally inadequate. Existing RAN architectures allocate resources based on fixed service profiles, whereas 6G scenarios demand agile, differentiated, and fully customized services to accommodate diverse applications and rapidly changing user contexts. Specifically, applications such as massive Internet of Things deployments, healthcare monitoring, and smart city infrastructure each impose distinct QoS requirements, resource profiles, and security constraints. Addressing these demands requires real-time service identification and scheduling, advanced network slicing with customized protocol stacks and optimization algorithms, and precise cross-slice coordination. However, these requirements exceed the capabilities of current RAN management frameworks.

All of these advancements operate under stringent resource constraints, including spectrum scarcity, energy limitations, sustainability pressures, and capital expenditure bounds, which together pose significant challenges for RAN design. High spectral efficiency remains a fundamental goal, while energy efficiency becomes increasingly critical in 6G, as processing complexity and network densification escalate amid mounting regulatory and environmental pressures. In addition, deploying advanced base stations with massive multiple input multiple output (MIMO), edge computing, and fiber backhaul further amplifies cost and operational complexity. Therefore, multi-dimensional optimization across spectrum, power, computing, and infrastructure resources is required, which far exceeds the capabilities of current RAN architectures.

To address these challenges, several RAN designs have been proposed in recent years. As summarized in Table 1, while these architectures address specific 6G requirements (e.g., communication QoS assurance, service flexibility, intelligence, or sensing), they largely remain domain-specific and fragmented solutions. Their limited capability to natively support functions beyond communication, combined with insufficient cross-domain synergy and resource management, prevents them from fully realizing the potential of 6G, restricting support for extreme performance, multi-service integration, and agile service differentiation under constrained resources. This motivates the development of innovative architectural paradigms that seamlessly integrate multi-dimensional capabilities and enable holistic, cross-domain optimization to tackle these tightly coupled challenges.

Table 1 Current RAN design limitations for 6G.

Architecture	Key features	Limitations for 6G
C-RAN	Centralized BBUs, CoMP support, NFV/SDN-enabled soft-BBUs	Connectivity-centric design with limited support for AI, sensing, and heterogeneous service orchestration; suboptimal for multi-service 6G scenarios.
SBA-RAN	Service-based microservices, dynamic orchestration, network slicing	Enhance service flexibility but with limited cooperative designs, such as a restricted cooperation scope and lack of support for cooperative sensing and intelligence.
AI-RAN	AI/ML-driven optimization, RAN intelligence	Focus on GPU-based infrastructure optimization but lack systematic, logical RAN-native AI function design.
Computing-and-communication integrated RAN	Embedded computing resources, edge processing	Serve as an integrated communication and computing platform for AI services and enables joint communication and computing optimization, but still lacks sufficient sensing integration and cooperative communication.
Sensing-integrated RAN	Environmental awareness, ISAC, proactive optimization	Improve sensing-communication synergy but lack holistic AI-service integration and efficient multi-node cooperation.

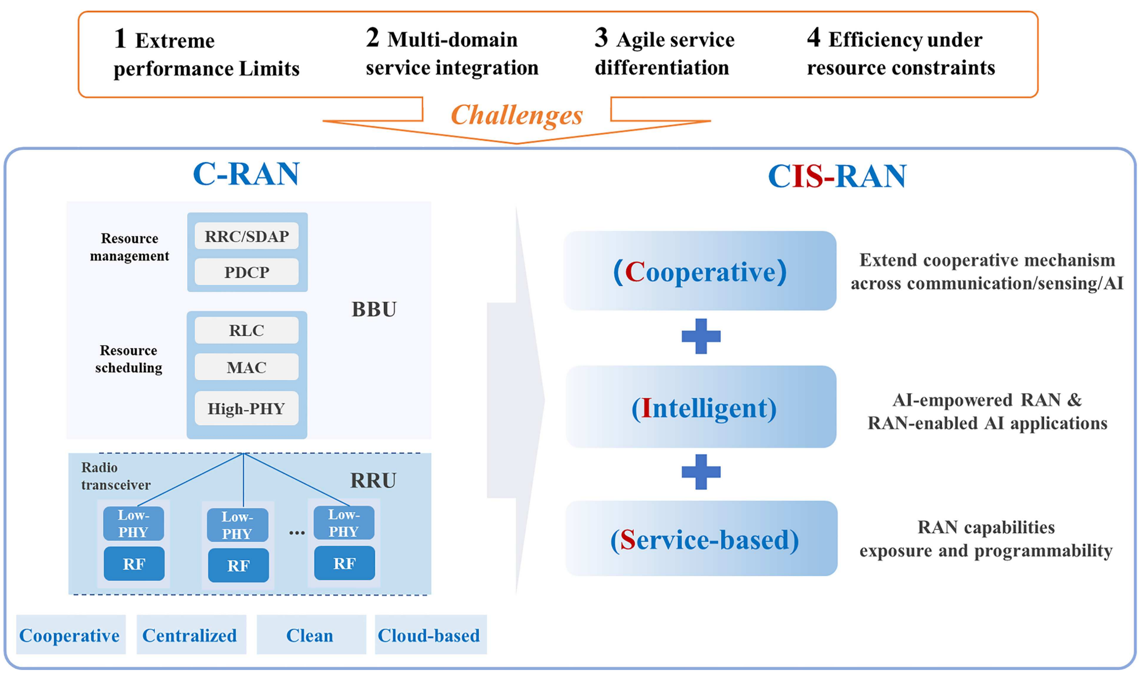

Figure 1 (Color online) Evolution from C-RAN to CIS-RAN.

2.2 Essential features for CIS-RAN

To address the limitations of C-RAN in meeting 6G requirements, CIS-RAN is proposed as its evolutionary successor, designed to achieve enhanced adaptability, autonomous optimization, and end-to-end diverse service support through three key advanced features. Specifically, cooperative mechanism is enhanced to break the static boundaries of traditional C-RAN, enabling dynamic multi-node collaboration across multiple domains, such as cooperative communication, cooperative sensing and cooperative intelligence; intelligent processing allows adaptive RAN optimization and native support for diverse AI applications, while service-based design exposes RAN functionalities through standardized open interfaces, enabling programmable and flexible service delivery across a variety of 6G scenarios. Figure 1 illustrates this evolution from C-RAN to CIS-RAN, highlighting both the key design features and the challenges addressed in the process.

2.2.1 *Cooperative*

As 6G evolves to support ubiquitous connectivity and diverse services with stringent latency, reliability, mobility, and heterogeneity requirements under limited resource conditions, cooperation emerges as a foundational enabler that improves resource efficiency, enhances robustness, and fosters seamless integration of communication, sensing, and intelligence. In CIS-RAN, cooperation extends across these domains, forming a unified paradigm for multi-dimensional optimization over distributed nodes. This cooperative design is reflected in two critical dimensions: user-centric cooperative communication and extended cooperative sensing and AI.

At the communication level, user-centric cooperative transmission represents a significant evolution from conventional cell-centric designs, enabling always-on and seamless connectivity. Instead of static single-point attachment, multiple distributed radio units adapt their collaboration based on real-time channel conditions and service requirements. Through fine-grained synchronization, cooperation further evolves from non-coherent to coherent joint transmission, unlocking cooperative gains in spectral and energy efficiency. Moreover, the cooperative scope extends beyond the single BBU pools to boundary-free coordination across multiple pools, eliminating cell-level constraints and supporting seamless mobility [9].

Beyond communication, cooperative sensing enables distributed radio units and auxiliary sensing nodes to share and jointly process echo, channel, and signal feature data, thereby achieving network-wide environment perception. In parallel, cooperative AI supports distributed model training and inference, where computation and data are adaptively coordinated among BBUs and edge nodes for low-latency decision-making [10, 11].

Realizing these cooperative capabilities requires substantial architectural enhancements, including improved multi-node synchronization, real-time inter-BBU coordination, expanded sensing and data acquisition, and multi-point data fusion and processing. Together, these changes fundamentally redefine how RANs perceive, adapt to, and respond to dynamic environments.

2.2.2 *Intelligent*

The increasing complexity of 6G networks, characterized by heterogeneous service demands and dynamic operating conditions, necessitates intelligence as a fundamental architectural capability. In CIS-RAN, intelligence manifests in two dimensions: AI for RAN (AI4RAN), which leverages AI/ML techniques to optimize network operations, and RAN for AI (RAN4AI), which provides proximal AI services.

On the one hand, AI enhances the entire RAN processing procedure through its advanced capabilities in feature extraction, high-speed computing, and non-deterministic polynomial (NP)-hard optimization problem solving. In particular, AI-enabled feature extraction supports accurate high-dimensional channel acquisition with low overhead in scenarios involving high mobility and massive MIMO, thereby reducing CSI errors and improving spectrum efficiency. In addition, AI-powered real-time inference enables efficient execution of large-scale matrix operations, especially for massive MIMO systems, while its advanced problem-solving capabilities address NP-hard optimization tasks for dynamic network management and resource optimization. Collectively, these advancements overcome the performance and efficiency bottlenecks of current RAN architectures.

On the other hand, CIS-RAN serves as a key enabler for pervasive AI applications, providing the foundational infrastructure for 6G's AI-native services. It supports diverse applications, such as intelligent agents and immersive experiences, by offering enhanced transmission, localized data processing, and computing task offloading at the RAN edge. These capabilities enable proximity-based computation and communication, reducing latency and improving service quality, all powered by the intelligent functions embedded in CIS-RAN.

The bidirectional relationship between AI and RAN creates a tightly coupled ecosystem where network intelligence and AI applications co-evolve. AI4RAN capabilities improve network efficiency and capacity, creating favorable conditions for hosting diverse services, while RAN4AI infrastructure generates rich operational data and feedback that further refines AI4RAN algorithms. This synergy enables CIS-RAN to emerge as both an intelligent communication infrastructure and a platform for AI-driven service delivery, delivering unprecedented gains in spectral efficiency, energy efficiency, system capacity, user experience, and service capabilities extending beyond traditional communications.

2.2.3 *Service-based design*

The heterogeneous and dynamic requirements of 6G scenarios expose fundamental limitations in traditional C-RAN architectures that rely on fixed functional deployments and monolithic processing units [12, 13]. These limitations motivate the adoption of service-based design, which transforms RAN functionalities through capability exposure and programmability of the capabilities to enable flexible resource allocation and differentiated service support [14, 15].

RAN capability exposure allows authorized consumers to discover and access a variety of RAN functionalities, including communication capabilities, sensing capabilities, AI-driven processing, and computational resources. A unified capability exposure framework supports service registration, discovery, and access control, ensuring that capabilities can be shared in a controlled manner while preserving the integrity of internal RAN operations. Exposure forms the foundation for higher-level service orchestration, enabling the network to provide functionality such as resource monitoring and performance reporting, which is critical for supporting dynamic, context-aware, programmable service requirements.

RAN programmability builds on the exposure to capability by allowing network operators and verticals to customize and optimize functionalities according to service requirements. Through programmable mechanism, the network can reconfigure functions for specific scenarios, for example, prioritizing ultra-low latency paths or adjusting computation scheduling. Programmability provides the internal flexibility needed for dynamic service adaptation and resource management, and supports the integration of third-party orchestrators, cross-domain optimizers, or real-time edge applications.

The realization of service-based design requires architectural evolution across multiple layers. This includes adopting modular frameworks that decompose monolithic functions into fine-grained, reusable services with standardized interfaces, deploying orchestration engines to enable dynamic composition and coordination across distributed RAN nodes, establishing intelligent resource mapping mechanisms that translate abstract service requests into concrete physical allocations, and ensuring that exposed and programmable capabilities remain compatible with internal optimization algorithms while safeguarding critical network functions from unintended interference [16, 17].

2.3 CIS-RAN: a novel architecture for 6G

2.3.1 *Architectural overview*

The proposed CIS-RAN architecture is illustrated in Figure 2. Compared to conventional C-RAN, CIS-RAN introduces several key architectural innovations: RRUs and BBUs are enhanced with sensing and AI-driven processing capabilities, a new centralized RAN AI unit (CRAU) is introduced to enable network-wide intelligence coordination, and standardized frameworks are established for capability exposure and programmability. More specifically,

- The enhanced RRUs extend beyond traditional radio frequency and physical layer signal processing to incorporate sensing and AI-driven functionalities. By integrating these capabilities directly at the radio interface, RRUs can acquire real-time radio environment data and extract preliminary measurements such as channel state information, user location, and spatial signatures. AI-assisted functionalities, such as digital predistortion, further improve signal fidelity. This distributed intelligence enables data preprocessing and feature extraction in the early-stages, and processed information is forwarded to higher layers for coordinated decision-making on scheduling, interference mitigation, and beamforming optimization.

- The BBUs in CIS-RAN undergo substantial functional evolution to serve as intelligent edge processing nodes. Beyond conventional baseband processing and protocol stack management, BBUs integrate computing and storage resources to host AI functionalities for adaptive link management, interference mitigation, and localized resource scheduling. A critical enhancement is the multi-modal data fusion capability, where BBUs combine communication measurements with sensing observations to construct enriched contextual understanding for real-time AI-driven inference. This design enables rapid local responses while maintaining coordination with centralized optimization, thereby supporting low-latency decision-making and flexible programmability.

- To enable network-wide intelligence and service programmability, the CRAU is introduced as a logically centralized node, providing both AI-assisted processing and orchestration of tasks across the RAN. It is responsible for training AI models, executing inference tasks, and managing their lifecycles, thus supporting intelligent decision-making across the RAN. By exposing programmable interfaces and RAN capabilities, the CRAU facilitates flexible use of computing and communication resources and enables RAN-based AI service provision.

The architecture design reflects the considerations on the trade-offs between latency and resource efficiency in terms of AI and sensing functionalities. On one hand, centralized processing and model coordination can improve statistical multiplexing and resource utilization, which is beneficial for AI training consistency and system-wide optimization. On the other hand, excessive centralization may incur additional transmission and processing delays, particularly for latency-sensitive inference and control tasks. CIS-RAN addresses this trade-off through a hierarchical intelligence design, where time-critical and lightweight AI functions are executed closer to the radio edge (e.g., BBU), while delay-tolerant and resource-intensive tasks are handled in CRAU. Consequently, latency requirements and resource efficiency are jointly integrated as architectural design dimensions, rather than being treated as independent

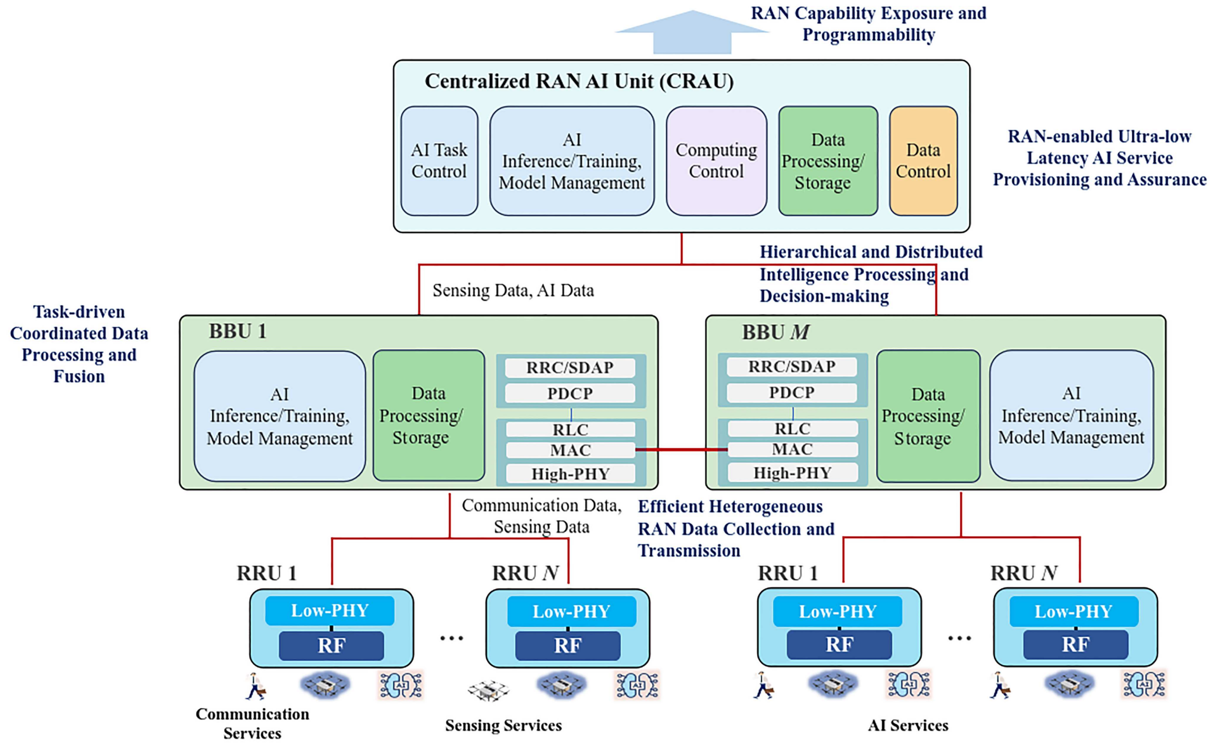

Figure 2 (Color online) CIS-RAN architecture.

Table 2 Functional components of CIS-RAN architecture. The entries in bold indicate functions inherited from traditional C-RAN systems; others represent CIS-RAN innovations. FFT: fast Fourier transform; iFFT: inverse fast Fourier transform; CP: cyclic prefix; PRACH: physical random access channel; MAC: medium access control; RLC: radio link control; RRC: radio resource control; PDCP: packet data convergence protocol; SDAP: service data adaptation protocol; API: application programming interface.

Component	Functionalities
Remote radio unit (RRU)	**– Radio frequency (RF) signal transmission/reception** **– Low-PHY processing: FFT/iFFT, CP insertion/removal, PRACH extraction, digital precoding/beam-forming [18]** – Sensing data collection and low-level observations (e.g., signal strength variation, spatial signatures)
Baseband unit (BBU)	**– High-PHY layer processing: modulation/detection, channel coding/decoding** **– MAC, RLC, RRC, PDCP, SDAP protocol stack management** – Sensing observation fusion – Local data storage and structured data preprocessing – Execution of lightweight AI models for near-real-time decisions
Centralized RAN AI unit (CRAU)	– Centralized AI training and orchestration – AI model lifecycle management – Joint optimization of compute, communication, and sensing resources – Exposure of capabilities via programmable APIs

objectives. Compared with other design options such as BBUs without AI capabilities or with over-provisioned computing resources, this approach strikes a practical balance between latency requirements, hardware investment, and overall network performance.

The functional roles of these architectural components are further summarized in Table 2, where the entries in bold denote functions inherited from C-RAN and standard entries reflect newly introduced capabilities essential for 6G.

Several interfaces are essential for the CIS-RAN. The BBU-RRU interface is capable of real-time and high-throughput transmission of communication and sensing data between a BBU and multiple RRUs. The BBU-BBU interface is designed to support UE mobility, as well as user-centric cooperative communication and sensing across multiple BBUs. The CRAU-BBU interface enables the exchange of information types, including (i) RAN data for AI/ML model training, and the trained AI/ML models, typically delay-tolerant and exchanged at time scales of hundreds of milliseconds or longer; (ii) RAN data and AI/ML inference results for cooperative optimization, with latency requirements in the order of tens of milliseconds; and (iii) fine-grained signaling and feedback for near-real-time task control, requiring sub-10-ms latency.

The CIS-RAN architecture highlights a smooth RAN evolution toward 6G by considering both legacy RAN deployment and the ongoing standardization process. First, CIS-RAN has a minimum impact on the interplay with UE, core network and transport. No requirement is imposed on the UE protocol stack as the intelligence and collaboration in CIS-RAN can be confined within the RAN side. Additional transport load from sensing, cooperative communication and AI/ML can be mitigated by advanced transport technologies, task-driven scheduling and compression. The interaction with the core network is compatible with the traditional functional split, and the associated enhancements can be streamlined by extending existing network exposure frameworks. Second, CIS-RAN is aligned with ongoing standardization efforts. Progress on RAN intelligence has been made in 3GPP since Rel-18, while cooperative sensing and communication, new RAN services and associated service-based designs are under active discussion in the Rel-20 studies for 6G. Meanwhile, O-RAN provides a practical pathway for early trials and incremental CIS-RAN deployment, where the RAN intelligent controllers can be enhanced to host the CRAU functionalities, D2 interface for advanced cooperative communication, and open APIs for service exposure, etc.

2.3.2 *Architectural trade-offs and unified optimization abstraction*

To explicitly capture these architectural trade-offs, a unified modeling and optimization framework is introduced to abstract the joint design of CIS-RAN. The objective is to balance the utilities of communication, computing, and sensing services against the associated capital and operational expenditures, leading to a holistic co-design problem across multiple architectural dimensions.

Let $\mathbf{X} = \{\mathcal{A}, \mathcal{Z}, \Theta, \mathcal{F}, \mathcal{B}, \mathcal{C}\}$ denote the set of decision variables.

- $\mathcal{A} = \left\{\mathbf{a}^n_{\text{comm}}, \mathbf{a}^n_{\text{comp}}, \mathbf{a}^n_{\text{sense}}\right\}$ denotes the function placement decisions at RAN node n (indexing RRUs, BBUs, or the CRAU). These variables determine how functions are placed at node n, such as specific function implementation selection and whether a function is locally executed or split across multiple nodes (e.g., between n_1 and n_2). This abstraction applies equally to AI model placement and partitioning.
- $\mathcal{Z} = \{\mathbf{z}^m_k\}$ denotes data processing and transmission policies, where k indexes the service and m denotes the data modality. The variable $\mathbf{z}^m_k$ governs how heterogeneous service data are prioritized, processed (e.g., via feature extraction or compression), and transmitted to satisfy QoS requirements.
- $\Theta = \{\boldsymbol{\theta}_n\}$ characterizes the multi-node cooperation configuration, including node participation in cooperative processing and the establishment of coordination relationships among distributed CIS-RAN entities.
- $\mathcal{F} = \{\mathbf{f}^m_k\}$ denotes the service function chaining decisions for service k, where $\mathbf{f}^m_k = (j, i, n)$ specifies that the j-th function in the service chain is executed as function i at node n to processes data $\mathbf{z}^m_k$. The placement of function i is determined by $\mathcal{A}$ and the data policy is determined by $\mathbf{z}^m_k$, thus linking the service function chain with function placement and data flow.
- $\mathcal{B} = \{\mathbf{b}^m_k\}$ denotes radio resource allocation for services, such as bandwidth, power, time-frequency slots, subcarriers, and antenna resources. Similarly, $\mathcal{C} = \left\{\mathbf{c}^n_{i,k}\right\}$ denotes computing resource allocation for function execution, where $\mathbf{c}^n_{i,k}$ specifies the resources allocated to function i at node n to support service k.

Based on these decision variables, the unified optimization problem is formulated as

$$\max_{\mathbf{X}} \quad \frac{\omega_1 \mathcal{U}_{\text{comm}} + \omega_2 \mathcal{U}_{\text{comp}} + \omega_3 \mathcal{U}_{\text{sense}}}{\text{OPEX} + \text{CAPEX}}, \tag{1a}$$

$$\text{s.t.} \quad \sum_{k,m} \text{Res}(\mathbf{b}^m_k) \leqslant B_{\text{tot}}, \tag{1b}$$

$$\sum_{i,k} \mathbf{c}^n_{i,k} \leqslant C_n, \tag{1c}$$

$$\text{Req}(\mathbf{a}^n_{\text{comm}}, \mathbf{a}^n_{\text{comp}}, \mathbf{a}^n_{\text{sense}}) \leqslant \sum_k \mathbf{c}^n_{i,k}, \tag{1d}$$

$$g_1(\mathbf{z}^m_k, \mathbf{f}^m_k, \mathbf{b}^m_k, \mathbf{c}^n_{i,k}) \leqslant U_{\min}, \tag{1e}$$

$$g_2(\mathbf{z}^m_k, \mathbf{f}^m_k, \mathbf{b}^m_k, \mathbf{c}^n_{i,k}) \geqslant U_{\max}, \tag{1f}$$

where $\mathcal{U}_{\text{comm}}$, $\mathcal{U}_{\text{comp}}$, and $\mathcal{U}_{\text{sense}}$ denote the utilities of communication, computing, and sensing functions in CIS-RAN, respectively, with ω_1–ω_3 as weighting coefficients. The specific utility metrics are scenario- or service-dependent, such as throughput, spectrum efficiency, end-to-end latency accounting for both communication and computing, AI inference accuracy, and sensing accuracy.

The optimization is subject to resource limitations and performance requirements. Constraints (1b) and (1c) represent the total radio and computing resource capacities constraints, where $\mathrm{Res}(\mathbf{b}_k^m)$ denotes the radio resource usage. Constraint (1d) stipulates that the allocated computing resources must satisfy the minimum computing requirement mandated by the function placement decisions, particularly for resource-intensive capabilities empowered by AI. Finally, Constraints (1e) and (1f) enforce QoS constraints, distinguishing between metrics to be minimized (e.g., latency, energy, error rates) and those to be maximized (e.g., throughput).

The explicit inclusion of these metrics within a unified objective function ensures that the architecture can navigate the trade-offs between processing latency and resource utilization efficiency. The OPEX term mainly reflects energy consumption and runtime resource usage, whereas CAPEX is primarily associated with infrastructure deployment, upgrade, and maintenance.

By formalizing function placement, data transmission and processing, cooperation, and resource allocation within a unified optimization framework, this model quantifies trade-offs between service performance and costs, offering principled guidance for CIS-RAN design, resource orchestration, and system-level optimization.

2.3.3 *Architectural advantages*

With these capabilities integrated across RRUs, BBUs, and the CRAU under the guidance of the trade-off considerations established by the unified modeling framework, CIS-RAN provides a foundation for scalable, high-performance, and service-flexible operations, as detailed below.

(1) **Sustainable and agile infrastructure deployment.** In CIS-RAN, communication, sensing, and AI capabilities are co-integrated within the RAN architecture, which reduces duplication of hardware functions and lowers both capital and operational costs. Integrated AI functions allow the system to adjust resource utilization in real time—for example, redistributing baseband workloads across units or dynamically adapting carrier and cell configurations to match traffic variations—so that efficiency is improved without compromising service continuity. In addition, the design supports software-based function upgrades and gradual capacity expansion, enabling infrastructure evolution without frequent hardware replacement. Such a deployment paradigm not only extends the lifecycle of RAN components but also allows the infrastructure to flexibly adapt to heterogeneous service requirements.

(2) **Ultra high performance across communication, sensing and AI.** CIS-RAN achieves enhanced performance by facilitating tight interconnection among communication, sensing, and AI functionalities. For example, communication reliability is improved through the incorporation of sensing-derived information into radio link adaptation, beam management, and mobility estimation. Concurrently, sensing accuracy benefits from joint processing of radio signals across nodes, while AI can refine these measurements into predictive knowledge that further supports network control. Continuous access to traffic and channel statistics further strengthens AI inference capabilities at the network edge. This mutual enhancement across communication, sensing, and AI allows CIS-RAN to sustain high data rates, fine-grained situational awareness, and adaptive decision-making, surpassing the capabilities of conventional C-RAN architectures.

(3) **Adaptive and on-demand multi-dimensional service enablement.** CIS-RAN adopts a service-based design, allowing communication, sensing, and AI capabilities to be orchestrated according to current application requirements. For example, a smart factory may request combined positioning and uplink reliability support, while a vehicular network may require low-latency AI inference combined with real-time environment sensing. These capabilities are supported by AI-assisted functions within the RAN, enabling dynamic configuration of radio resources and processing to satisfy service requirements. By exposing these capabilities through standardized interfaces, CIS-RAN enables flexible support for diverse applications.

3 Key enabling technologies and mechanisms of CIS-RAN

The CIS-RAN architecture presented in the above section establishes a structural foundation for cooperative, intelligent, and service-based design. However, realizing these architectural capabilities in practice requires a set of enabling technologies that address fundamental challenges in data handling, processing coordination, intelligence distribution, service provisioning, and capability exposure. This section identifies and discusses five key enabling technologies that underpin the CIS-RAN framework. Figure 3 provides an overview of the associations between features and technologies, setting the stage for the detailed discussion of each enabling technology in the following subsections.

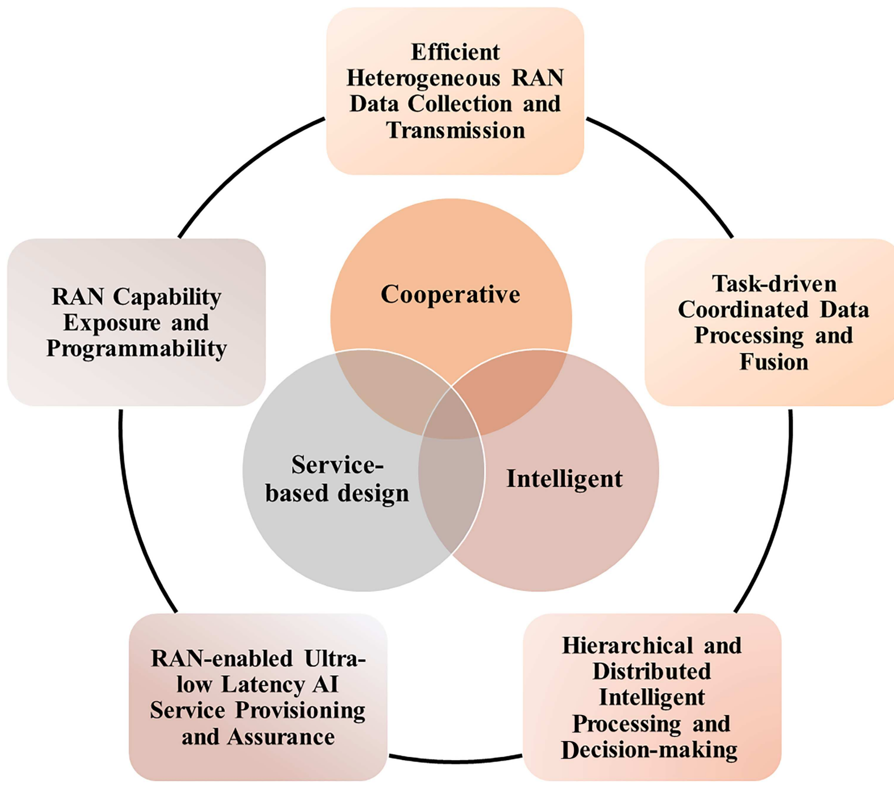

Figure 3 (Color online) Key technologies enable the essential features of CIS-RAN.

3.1 Efficient heterogeneous RAN data collection and transmission

The integration of communication, sensing, and AI-native capabilities in future 6G RAN, together with the diversity of emerging AI applications inevitably leads to the generation of highly heterogeneous data. These include raw I/Q samples, channel state information, sensing echoes, mobility traces, and semantic features extracted by RAN edge intelligence, each with distinct temporal resolutions, reliability requirements, and transmission priorities.

Conventional RAN architectures are not well suited for handling these heterogeneous data streams. First, they typically adopt a uniform transport strategy, treating raw signals, control parameters, and semantic features in the same way, which leads to excessive bandwidth consumption and redundant transmission. Second, the lack of task-driven prioritization results in suboptimal utilization of network resources, since not all data contribute equally to the performance of downstream tasks such as beam management or object detection. Third, conventional designs often rely on single-node measurements, making it difficult to achieve reliable sensing under non-line-of-sight (NLOS) or limited field-of-view conditions. Finally, heterogeneous data captured by different nodes often arrive in an unsynchronized manner, and the absence of precise time-frequency alignment significantly degrades the effectiveness of multi-node fusion. Together, these issues prevent existing RAN designs from achieving efficient and timely heterogeneous data handling.

To address these challenges, the proposed CIS-RAN introduces three key enabling technologies for efficient heterogeneous data collection and transmission. First, advanced synchronization mechanisms are employed to achieve precise time-frequency alignment across distributed nodes, which is essential for accurate MIMO beamforming and multi-point sensing fusion [19]. By establishing coherent phase relationships and compensating for propagation delays, CIS-RAN enables cooperative transmission gains and statistical fusion of sensing observations from geographically separated units. Second, intelligent edge data processing and compression are leveraged to reduce transmission overhead while preserving core information within tasks. AI-equipped RRUs and BBUs perform task-oriented preprocessing and compression, extracting key features such as angle of arrival (AoA), angle of departure (AoD), propagation delay, and received signal strength indicator (RSSI), and selectively filtering and encoding data based on its relevance to downstream tasks. For ISAC applications, pilot or reference signals can be locally processed to detect obstacles and optimize beamforming, with only processed channel state information and environmental reconstructions transmitted over the backhaul [20]. Third, a novel AI-optimized bearer architecture is established for RAN-side closed-loop data flows, complementing conventional control and user plane bearers. Unlike standard control plane bearers, which are limited by payload size, or user plane bearers, which are constrained by strict latency requirements, the dedicated AI bearer enables flexible scheduling and resource allocation tailored to diverse AI application data, including computation-intensive, latency-sensitive sensing data, model parameters, and inference results. Through the coordinated operation of these three technologies, CIS-RAN transforms heterogeneous data handling from a bandwidth-constrained transport problem into an intelligent information processing pipeline, thereby providing efficient and high-quality support for emerging intelligent 6G services.

3.2 Task-driven coordinated data processing and fusion

6G networks are expected to support multi-dimensional tasks such as cooperative communication, cooperative sensing, and cooperative intelligence, which necessitates a transition from individual node processing to synchronized coordination across distributed RAN elements. However, a fundamental conflict exists between the requirement for high-precision data fusion and the constraints of fronthaul/backhaul bandwidth. Conventional C-RAN architectures struggle with this trade-off due to their rigid functional splits and a traffic-centric design philosophy that is unknown to task semantics. In such frameworks, the selection of coordination clusters is often static, and data processing lacks the flexibility to adapt to real-time factors like user location or channel conditions, leading to either unsustainable signaling overhead or significant information loss.

CIS-RAN introduces a task-driven coordinated data processing and fusion paradigm in which coordination scope, processing depth, and participating nodes are dynamically determined according to service objectives and network context. For user-centric cooperative communication, the architecture optimizes flexible Low-PHY and High-PHY split processing between RRUs and BBUs to handle the massive data volumes required for joint precoding. By continuously monitoring network conditions, the system identifies the optimal set of BBUs for coordination, replacing semi-static cell-clustering with real-time, learning-based selection. This approach avoids the mismatched boundaries and handover interference common in traditional coordination, thereby maximizing spatial multiplexing gains with minimal resource wastage. Similarly, for cooperative sensing tasks like UAV detection, the system employs a flexible fusion hierarchy guided by geometry-aware policies. By prioritizing BBUs that offer diverse observation angles and favorable visibility, the architecture ensures that only the most informative nodes participate in the task. Instead of transmitting raw waveforms, these nodes extract features—such as delay/angle/Doppler spectrums—which the CRAU then fuses to achieve higher accuracy and robustness than any single node. This feature-level compression significantly reduces the fronthaul bandwidth requirement while preserving target signatures, allowing high-precision detection without congesting the transport network. The proposed task-driven integration further extends to cooperative intelligence where the CRAU coordinates AI /ML tasks with BBUs, as shown in Figure 4; the orchestration plane manages dataset heterogeneity and model aggregation to facilitate efficient distributed learning. To enhance system responsiveness, techniques such as model compression can be utilized to reduce update sizes, while asynchronous training mitigates the straggler effects inherent in distributed environments. Coupled with resource-aware scheduling that balances inference accuracy against execution efficiency, this approach enables AI-native functions to scale without triggering unsustainable data transfer overhead.

Crucially, the scalability of this coordination is facilitated by task-relevant clusters—such as dynamic BBU sets for user-centric communication or localized BBU-CRAU groups for model fine-tuning—thereby avoiding the unsustainable signaling overhead and complexity of global coordination. By allowing these clusters to form and evolve dynamically based on the task context, the architecture ensures that the RAN can sustain scalable multi-node cooperation while maintaining high efficiency.

3.3 Hierarchical and distributed intelligent processing and decision-making

The evolution of 6G RAN toward an intelligent communication fabric and an AI-enabling infrastructure introduces heterogeneous timing constraints and decision scales, ranging from microsecond-level radio control to long-term network optimization. Purely centralized designs fail to meet stringent latency requirements, while fully localized schemes lack the global awareness and coordination necessary for large-scale optimization.

The diverse service requirements, computing resources, and data availability necessitate a hierarchical and distributed intelligence framework that balances rapid responsiveness with large-scale optimization.

RAN functions also exhibit heterogeneous timing constraints, with physical-layer control requiring millisecond-level reactions, MAC scheduling and RRC operations operating on tens to hundreds of milliseconds, and cross-BBU cooperation spanning longer timescales. While the RAN already incorporates layered units (such as RRU and BBU), embedding intelligence introduces challenges such as environmental dynamics, resource heterogeneity, and synchronization errors. Therefore, hierarchical intelligence must not merely reflect architectural layering but must also enable adaptive decision-making to effectively manage these complexities [21].

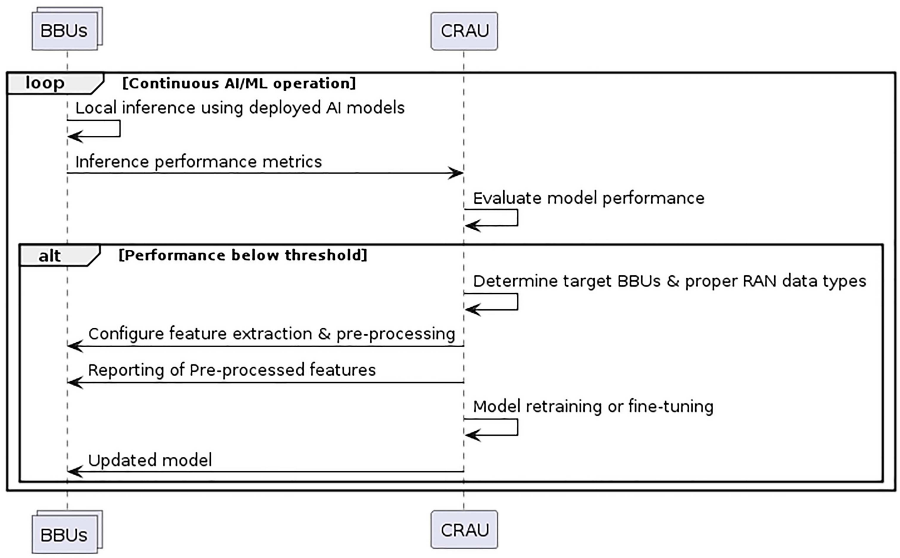

Figure 4 (Color online) CRAU coordinates AI/ML tasks with BBUs.

To address these challenges, intelligence is distributed across multiple tiers with complementary roles. The ultra-low-latency AI/ML models operating at microsecond timescales, such as digital pre-distortion, require inference to be tightly coupled with the radio front-end in the RRUs. For the link-level real-time control with millisecond-level latency (e.g., CSI prediction), the inference complexity mostly ranges from 0.1 MFLOPs to 100 MFLOPs, and therefore is best placed at the BBU, which provides sufficient computing capability while maintaining tight interaction with PHY and MAC processing. For the near-real-time and non-real-time inference tasks (100 ms and above) at the network level (e.g., energy saving) and model training tasks, the latency requirements are much more relaxed, and so the CRAU is an ideal location with the benefit from statistical multiplexing of multiple tasks. By dynamically allocating decision rights and workloads according to task urgency, latency budgets, and node capabilities, the system ensures that each task is executed at the most appropriate tier, thereby reconciling responsiveness with overall efficiency.

Ensuring robustness and adaptability is equally critical. Local controllers must maintain service continuity by reverting to simplified strategies under resource or connectivity shortages, while model management must support rapid iteration and seamless switching. Centrally pretrained models can be locally fine-tuned, and systems should flexibly transition between AI versions or even conventional algorithms. Such adaptability enables fast upgrades, smooth environment adaptation, and safe rollback, ensuring stable performance under dynamic conditions [22].

Through hierarchical deployment, coordinated decision-making, and adaptive lifecycle management, CIS-RAN offers a scalable and resilient framework for 6G RAN intelligence. It delivers rapid local reactions, regionally coordinated collaboration, and globally optimized control, while maintaining robustness and evolving with diverse and dynamic service demands.

3.4 RAN-enabled ultra-low latency AI service provisioning and assurance

For emerging AI-driven applications like collaborative robots, a deterministic ultra-low service latency is crucial. However, the end-to-end (E2E) latency of such services is a composite of both air-interface transmission and backend processing delays. Traditional RAN is oriented toward connectivity only, lacking the structural visibility and control mechanisms required to coordinate computing resources. Consequently, even with high-speed radio links, unpredictable processing jitter or long user-plane paths through the core network can easily violate the stringent latency budgets of AI tasks.

To address these challenges, connectivity QoS/QoE assurance has an important role. In the proposed CIS-RAN, cooperative transmission schemes across multiple BBUs help mitigate inter-cell interference and maintain comparable channel quality at the cell edge to that at the cell center. Sensing and positioning information can further improve these schemes by enabling optimized beam steering and seamless handovers. In addition, service-level assistance information collected from the application layer provides valuable input for AI/ML-based algorithms in the RAN. This information may include the dynamics of service data flows (e.g., traffic patterns) and AI service-specific QoE metrics, such as time-to-first-token. Leveraging these inputs, the RAN can perform intelligent scheduling to

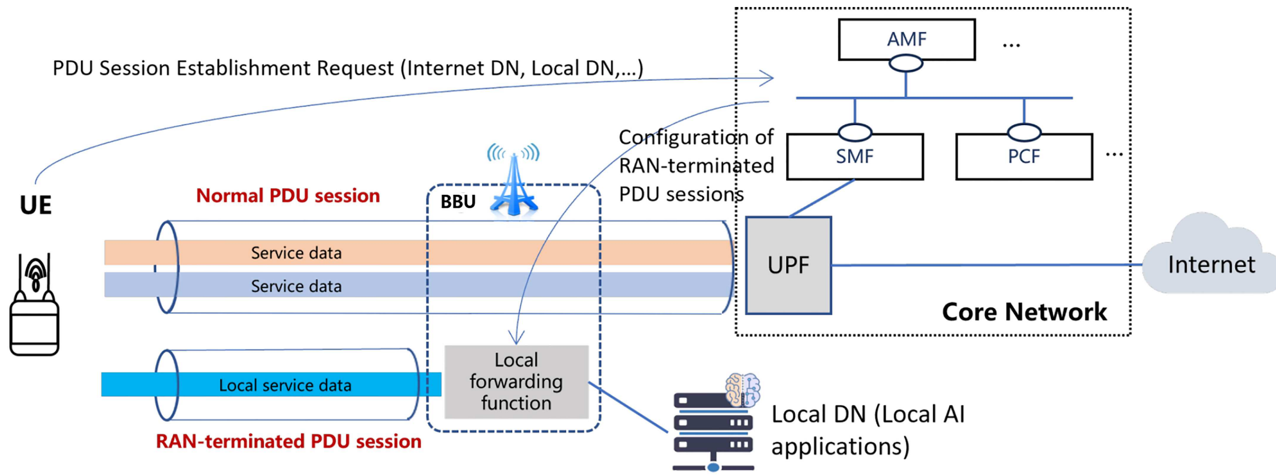

Figure 5 (Color online) RAN-terminated PDU session.

align service requirements with radio channel variations.

Beyond connectivity, computing service provisioning enables CIS-RAN to host local AI applications with reduced latency, which involves orchestration and local breakout. For orchestration, diverse AI accelerators such as graphics processing units (GPUs), neural processing units (NPUs), and data processing units (DPUs) need to be efficiently managed for concurrent AI/ML tasks. Techniques such as multi-instance GPU (MIG) for isolated multi-instance execution and multi-process service (MPS) for shared multi-instance utilization can be adopted for dynamic and scalable allocation of computing resources. Unified orchestration across heterogeneous computing platforms can be realized through unified computing models, accelerator abstraction layers (AAL) [23], virtualization, and containerization. These approaches help reduce development complexity, improve resource utilization, and enhance cross-platform interoperability. Local breakout further reduces latency by providing direct access to applications hosted within CIS-RAN, thereby avoiding the longer 5G user plane path through the UPF. In this case, the user-plane forwarding function at the base station is enhanced to support on-demand traffic offloading. One possible solution is the introduction of new PDU sessions that are terminated within the RAN (Figure 5). These sessions are configured by the core network in response to user equipment (UE) requests for local services and enforced at the BBU (not in UE), therefore minimizing the impact on the legacy UEs.

Furthermore, the joint optimization of communication and computing resources is crucial for the optimal allocation of both types of resources, when CIS-RAN offers both communication and computing services for AI service applications. In this scenario, the end-to-end user experience would be affected by both communication and computational QoS. To that end, the RAN must be capable of real-time monitoring of radio and computing performance, and dynamic scheduling of multi-dimensional resources, to guarantee end-to-end performance of AI services. Such synergistic orchestration not only ensures service reliability but also aligns with the sustainability objectives of 6G by maximizing the utilization of distributed infrastructures [24]. For instance, when the air-interface latency worsens due to degraded radio link quality, the system can allocate additional computing resources at RAN in real time to reduce the processing delay, so as to maintain the end-to-end delay. In the second example illustrated in Figure 6, handover (HO) of UEs served by local services can be optimized by CRAU according to the availability of distributed RAN computing resource. The CRAU determines the HO type (with or without computing HO) and instructs the target BBU either to establish the required RAN computing environment or to configure the data path for delivering the local service flow back to the source BBU. These mechanisms collectively transform the RAN into a deterministic AI service platform rather than a best-effort connectivity pipeline.

3.5 RAN capability exposure and programmability

The 6G RAN is expected to provide diversified services beyond communication (e.g., data services and computing services) for diverse consumers. To fully leverage these capabilities, vertical consumers require agile development environments and customized RAN solutions tailored to their specific operational needs. However, conventional RAN architectures, are characterized by closed, proprietary designs with rigid interfaces, where capability exposure is primarily restricted to the core network. This centralized exposure model lacks the granularity to reflect rapid fluctuations in radio environments or the localized availability of computing resources, making it difficult for third-party applications to interact with the RAN in a timely and responsive manner.

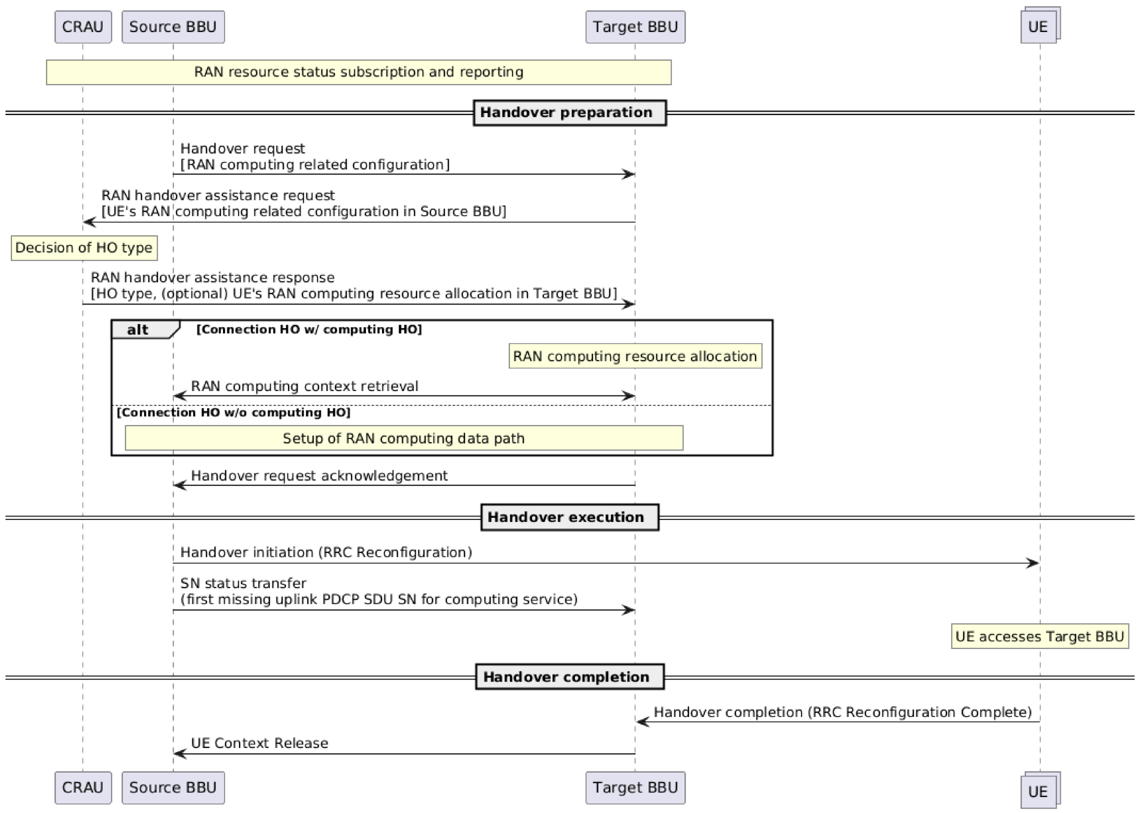

Figure 6 (Color online) Joint optimization of communication and computing.

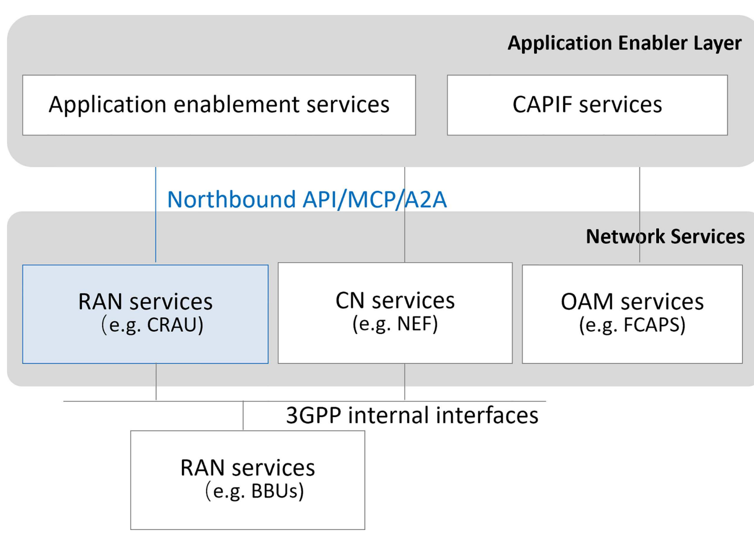

Figure 7 (Color online) RAN capability exposure diagram.

For that purpose, a unified capability exposure framework would be helpful. To overcome these barriers, the proposed CIS-RAN introduces a unified capability exposure framework. Such a framework should support service registration, service discovery and access control, etc., and take into account the specificity of RAN services (e.g., service availability may change due to fluctuations of radio conditions and UE mobility). One solution could be built on top of the common API framework (CAPIF) [25], e.g., by integrating its CAPIF core function with the CRAU. It is important that the RAN capability exposure functionalities need to collaborate with the similar functionalities in the core network including the NEF [26], so that the different services are integrated and exposed coherently, and this integrated, coherent exposure architecture is illustrated in Figure 7. Based on this framework, a brief RAN service exposure flow is illustrated in Figure 8.

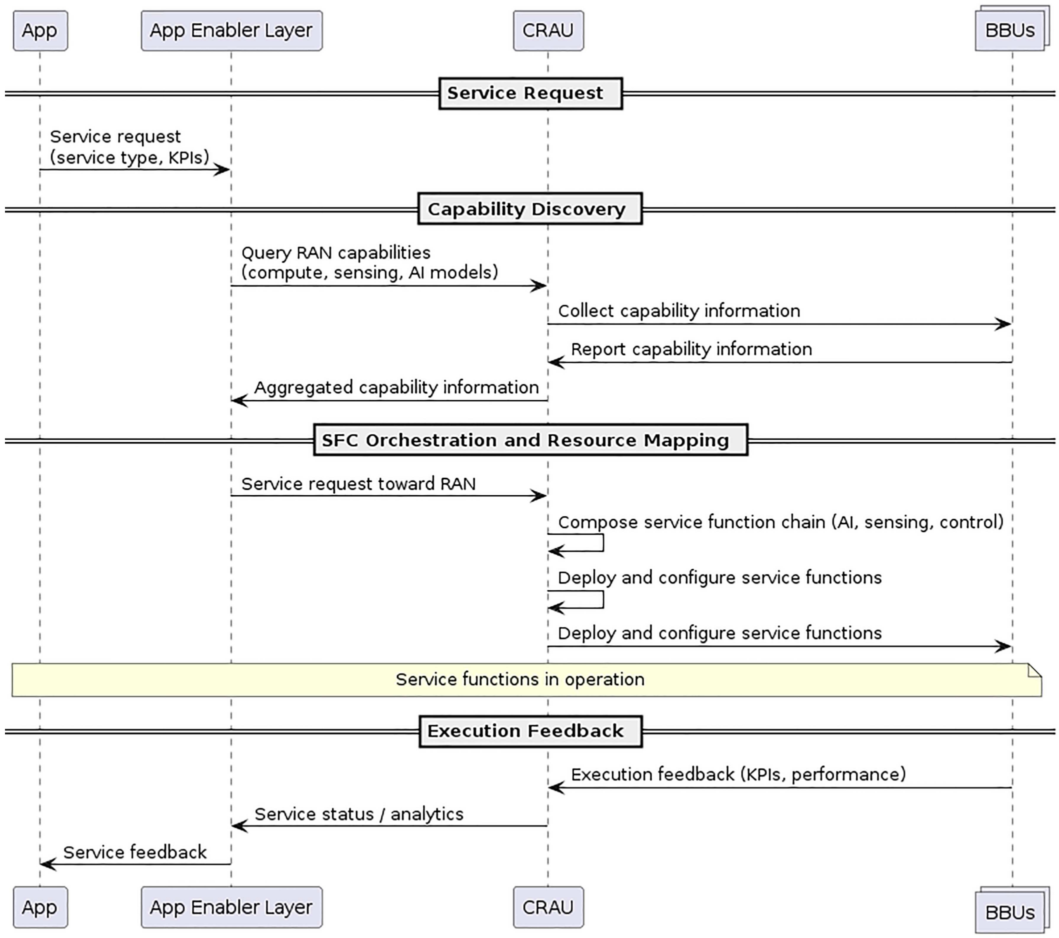

Figure 8 (Color online) RAN service exposure flow diagram.

As a special type of RAN capability exposure, RAN programmability allows the MNOs and/or the verticals to customize and optimize RAN internal processes, according to the diversified service requirements. The concept may be dated back to the SDN and is being extended to RAN in the past few years. The programmability in CIS-RAN can be further exploited in terms of task types and realtimeness. On task types, the CIS-RAN may expose open APIs for the control of site-/scenario-specific AI/ML model training task, or the clustering of BBU/RRUs for a multi-static sensing task. On realtimeness, new techniques and research in recent years have proven the feasibility of manipulating the real-time processes in the lower layers of RAN [27]. Last but not least, the reliability of CIS-RAN needs to be maintained with the programmability. Techniques such as digital twins [28] may play a crucial role in testing/validating the modified internal process before it is activated.

To effectively support such programmability, the monolithic, integrated BBU architecture can be decomposed into finer-grained atomic functions interconnected via uniform APIs. This decomposition allows on-the-fly assembly or reassembly of processing chains according to task-specific requirements. AI techniques are expected to play a central role in this pipeline, including intelligent function decomposition, context-aware service perception, automated orchestration of functional chains, optimal selection of execution nodes, and joint allocation of multi-dimensional resources, as summarized in Table 3. By integrating these capabilities, CIS-RAN can realize both openness and adaptability, meeting the heterogeneous demands of emerging 6G services.

4 Case study

4.1 Case 1: collaborative and distributed intelligence for interference prediction based on CIS-RAN

In this subsection, a case study is presented to demonstrate the advantages of cooperative intelligence enabled by the CIS-RAN architecture. Specifically, an uplink interference prediction-assisted UE throughput optimization scenario is investigated. The study first evaluates the performance of uplink interference prediction by comparing

Table 3 AI-driven RAN programmability enablers.

Case	AI-driven functions	Outputs	Benefits
Intelligent reconstruction of atomic functions	Intelligent decomposition of RAN functions	RAN services	Optimization the function decomposition policy
Intelligent service perception	Real-time analysis of application requirement	Requirement profile (e.g., latency, bandwidth, computing power, QoS)	Improving accuracy of requirement recognition
Intelligent function orchestration	Dynamic generation and deployment of service function chains (SFC)	Executable SFC according to application requirement and network topology	Reducing the delay of SFC generation and deployment
Intelligent allocation of multi-dimensional resources	Joint optimization of communication/computing /storage resources	Resource allocation policy	Improving energy and resource efficiency

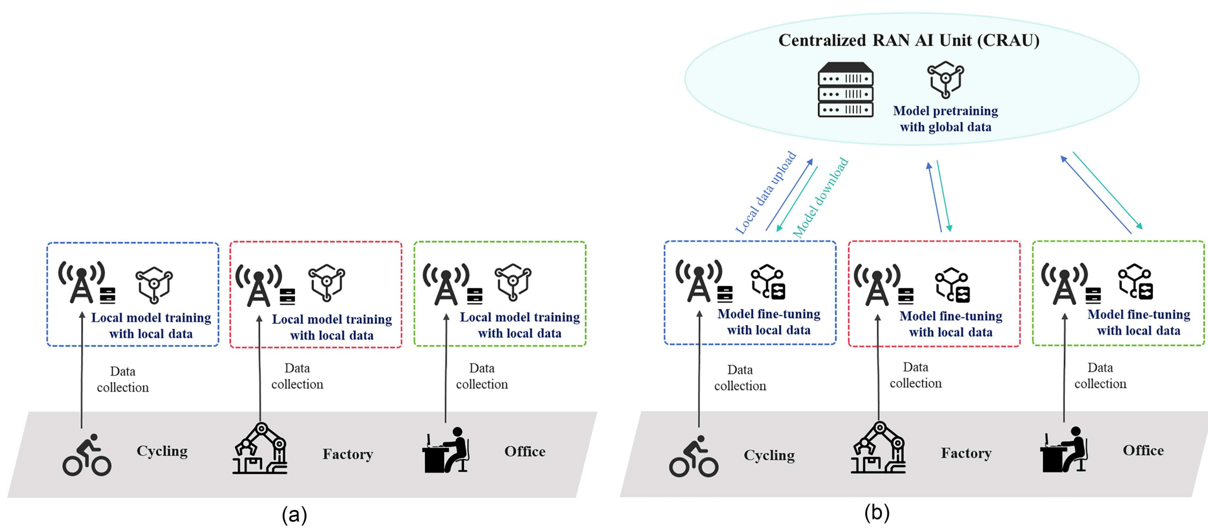

Figure 9 (Color online) Workflow comparison between (a) standalone local training (SLT) and (b) centralized pretraining with local fine-tuning (CPLF) schemes.

two training strategies: the CIS-RAN-enabled centralized pre-training with local fine-tuning (CPLF) approach and the conventional single-BBU local training (SLT) scheme, as illustrated in Figure 9. The comparison focuses on key metrics including prediction accuracy, training overhead, and convergence speed. Subsequently, the predicted interference information is leveraged to derive optimized modulation and coding schemes (MCS) and radio resource allocation, with the objective of enhancing UE throughput. Finally, system-level simulations are conducted to assess overall UE throughput performance, thereby illustrating the effectiveness and performance gains brought by the proposed CIS-RAN-enabled CPLF framework.

Simulations are conducted in `ns-3` with 4 BBUs and 16 UEs. The air interface adopts OFDM and incorporates both large-scale and small-scale fading, as well as inter-cell interference, to emulate realistic wireless conditions. Three representative 6G scenarios are considered to reflect diverse interference dynamics.

- Cycling: UEs at 15–25 km/h, characterized by rapidly varying interference caused by frequent handovers and intermittent blockages.
- Factory: Static UEs with periodic interference bursts induced by industrial machinery and dense metallic environments.
- Office: Pedestrian mobility with relatively persistent interference from neighboring cells and indoor reflections.

Table 4 compares the data volume and training time required by different interference prediction mechanisms across these scenarios. Across all three scenarios, the CPLF mechanism demonstrates superior efficiency. Compared with scenario-specific retraining, the collaborative pre-training and fine-tuning mechanism inherent to CIS-RAN

Table 4 Comparison of training data volume and training time at the BBU side under different interference prediction mechanisms.

Scenario	Data volume (samples)				Training time (s)		
	Re-train	Pre-train	Fine-tune	Reduction (%)	SLT	CPLF	Reduction (%)
Cycling	90175		45010	50.1	2812	639.651	77.25
Factory	70340	69842	45010	36.0	3039	881.835	70.98
Office	90175		45010	50.1	2531	550.485	78.25

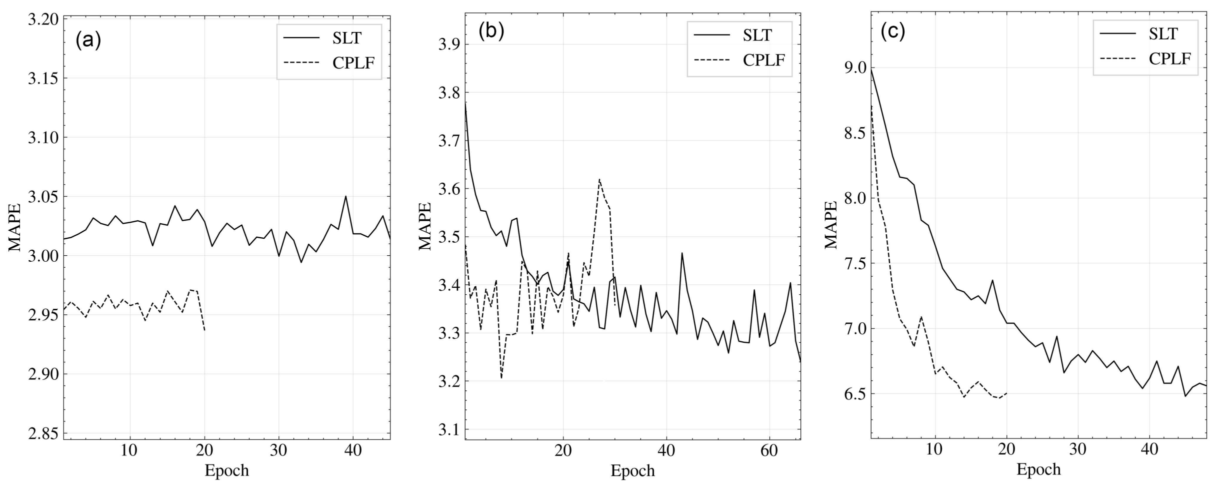

Figure 10 Convergence of interference prediction error under different learning mechanisms, measured by mean absolute percentage error (MAPE), defined as the average relative deviation between predicted and actual interference levels. (a) Cycling scenario; (b) factory scenario; (c) office scenario.

Table 5 System-level performance comparison in terms of average UE uplink throughput.

Scenario	Baseline (Mbps)	SLT scheme		CPLF scheme	
	(Non-AI)	Throughput	Gain (%)	Throughput	Gain (%)
Cycling	6.791	7.975	17.43	8.046	18.48
Factory	12.843	14.008	9.07	14.006	9.06
Office	12.028	13.295	10.53	13.427	11.63

significantly reduces the amount of training data at the BBU side, with data volume reductions ranging from 36% to 50% across scenarios. Meanwhile, the corresponding training time is consistently shortened, achieving a reduction of approximately 71%–78%, indicating that collaborative learning can effectively alleviate both data preparation and model update overhead.

Figure 10 illustrates the convergence behavior of the interference prediction error under different mechanisms. Despite the interference dynamics of the three scenarios, the CPLF approach consistently exhibits faster convergence and reaches a lower steady-state error compared with training solely on local data. This trend remains consistent across different interference dynamics, suggesting that shared representations learned from heterogeneous scenarios improve generalization while reducing the number of epochs required for convergence.

Table 5 shows that both SLT and CPLF mechanisms improve the average UE uplink throughput compared to the non-AI baseline. Crucially, the CPLF scheme achieves performance nearly identical to SLT (with gains of 9%–18% over the non-AI baseline), confirming that the significant overhead reductions observed earlier do not come at the cost of air-interface performance. This demonstrates that CIS-RAN can deliver high system-level gains through efficient, cooperative intelligence.

Collectively, the results indicate that the CIS-RAN architecture effectively addresses diverse and dynamic 6G interference environments through centralized and distributed collaborative intelligence. By aggregating global knowledge at centralized AI entities while enabling fast local adaptation at distributed nodes, the architecture jointly exploits global representation capability and local responsiveness, resulting in simultaneous gains in learning efficiency and system-level throughput performance.

Importantly, although the case study is conducted on a limited-scale topology for tractable evaluation, the scalability of CIS-RAN primarily stems from its architectural design rather than the network size itself. The hierarchical

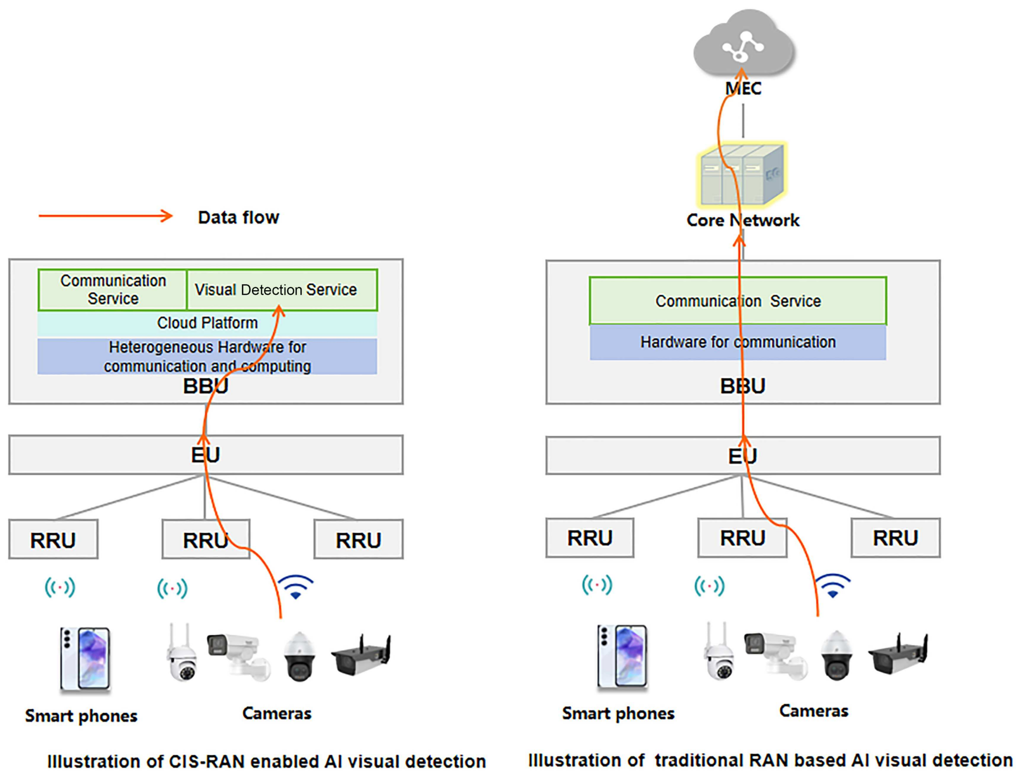

Figure 11 (Color online) Illustrations of the CIS-RAN prototype.

collaboration mechanism confines intensive coordination and model exchange within task-relevant clusters instead of enforcing network-wide synchronization, thereby preventing signaling overhead from growing linearly with the number of cooperating nodes. Meanwhile, centralized pre-training enables cross-scenario knowledge reuse, allowing newly added BBUs or cells to inherit pretrained representations and require only lightweight local fine-tuning, which mitigates the computational burden associated with large-scale expansion. From a system perspective, the hierarchical coordination between global knowledge aggregation and localized inference further enables AI computing resources to benefit from statistical multiplexing, ensuring that performance improvements do not rely on proportional increases in training cost or infrastructure investment as the network densifies.

From a deployment perspective, CIS-RAN also provides a practical evolution path for existing C-RAN infrastructures, where centralized RAN AI functions can be co-located with BBU pools or integrated into edge data centers to harvest AI computing multiplexing gains while preserving low-latency local intelligence and overall RAN flexibility. These characteristics indicate that the observed performance trends are structurally aligned with the requirements of large-scale 6G RAN deployments rather than being confined to the evaluated scale.

4.2 Case 2: low-latency AI visual detection assurance based on CIS-RAN

In this section, a prototype-based evaluation is conducted to validate the practical effectiveness of the proposed CIS-RAN architecture, with a particular focus on ultra-low-latency AI service provisioning and RAN capability exposure. Visual detection is selected as the test case due to its widespread application across domains characterized by heterogeneous tasks and stringent end-to-end latency requirements. However, it poses significant challenges for current RAN architectures due to the high latency caused by forwarding uplink video data to MEC servers via the UPF.

In the CIS-RAN architecture, this limitation is addressed by computing service provision powered by computing capability integration and RAN-terminated PDU session bearers, as introduced in Subsection 3.4. As illustrated in Figure 11, a cloud platform built on heterogeneous hardware is integrated into BBU, supporting both communication and computing capabilities. Further, to support diverse visual detection services, AI models including the Qwen2.5-VL multimodal large language model (LLM) and a YOLO-based visual detection model are deployed, which opportunistically utilize idle computing resources. The LLM is responsible for interpreting user intents and processing multimodal inputs, whereas the YOLO model performs real-time object detection and localization. As a result, AI inference can be supported directly at the BBU, and these capabilities are exposed to external entities

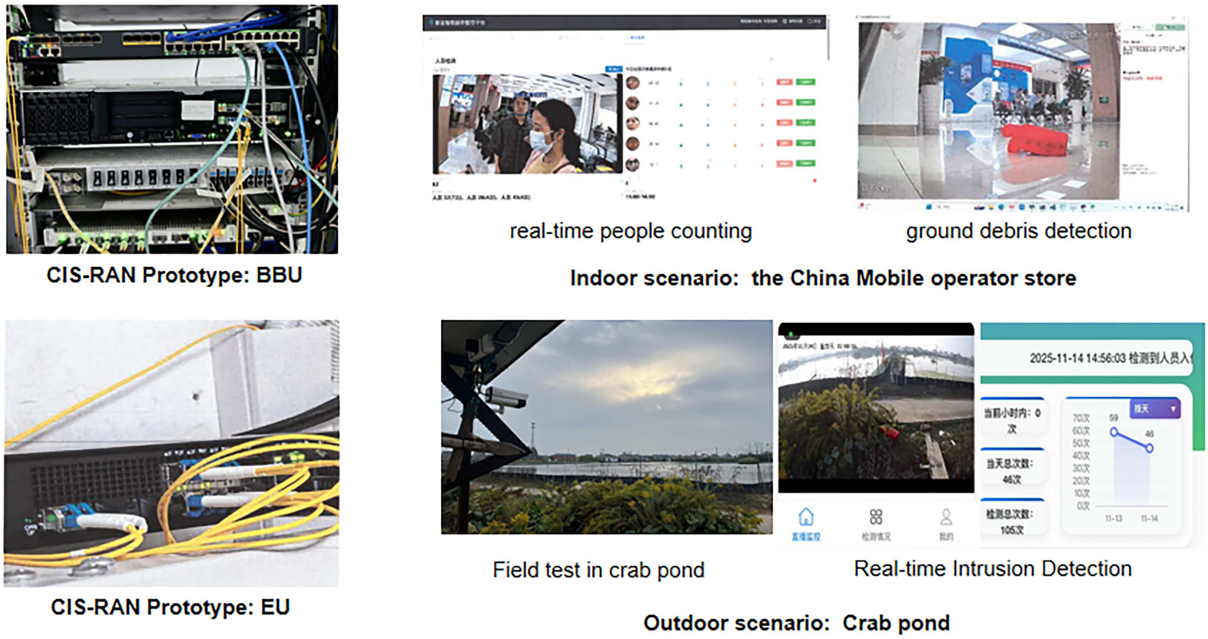

Figure 12 (Color online) Practical test scenario of the CIS-RAN prototype.

through the RAN exposure and programmability mechanisms introduced in Subsection 3.5. The BBU supports multiple RRUs serving heterogeneous terminals (e.g., smartphones and various types of cameras) via an edge unit.

As shown in Figure 12, the evaluation is carried out in two representative scenarios: the China Mobile operator store (indoor) and Crab Pond (outdoor). In the operator store, AI visual detection services are configured for real-time people counting, occupancy estimation, ground debris detection, and identification of users with special service needs, such as premium customers or individuals exhibiting anomalous behavior. In Crab Pond, services are focused on intrusion detection, abnormal behavior monitoring (e.g., theft, poisoning), and danger alerts for drowning. These services can be activated either concurrently or selectively in response to venue dynamics, enabling continuous monitoring and detection. The experimental results are shown in Figure 13. In both indoor and outdoor scenarios, the average end-to-end latency of the prototype-based detection service is maintained below 20 ms. By contrast, under a conventional architecture where inference is executed at an MEC/edge server and service data are forwarded through the UPF, the end-to-end latency increases to approximately 60 ms. These results demonstrate that, by terminating service data paths and integrating AI inference directly at the RAN, CIS-RAN enables tightly coupled perception-decision-action loops with significantly reduced end-to-end latency. Such closed-loop service support provides a fundamental capability that can be naturally extended to embodied intelligence scenarios, where deterministic latency assurance can be further achieved by enabling joint communication-computing optimization within the RAN building upon the validated prototype.

5 Future directions

Although the CIS-RAN architecture and its key enabling technologies have been presented and evaluated through simulations above, the practical implementation and deployment aspects have not been adequately explored. These challenges are paramount to achieving the full potential of CIS-RAN.

A critical challenge stems from the need for heterogeneous hardware support. Multi-dimensional capabilities and computing resources integrated into CIS-RAN are highly likely to rely on diverse platforms (e.g., GPU, central processing unit, field programmable gate array, application specific integrated circuit), which exhibit significant variations in latency, power efficiency, and programmability. Consequently, building a scalable intelligent computing infrastructure capable of supporting advanced AI-native functions across this hardware diversity presents a major hurdle. Specific research is needed on unified hardware abstraction layers and lightweight cross-platform runtimes that can decouple high-layer task logic from underlying instruction sets (e.g., CUDA vs. OpenCL), ensuring deterministic execution latency across heterogeneous nodes.

Beyond infrastructure, the flexible deployment of CIS-RAN components presents a multifaceted challenge. A key consideration is the deployment of the newly introduced CRAU, which must dynamically account for traffic

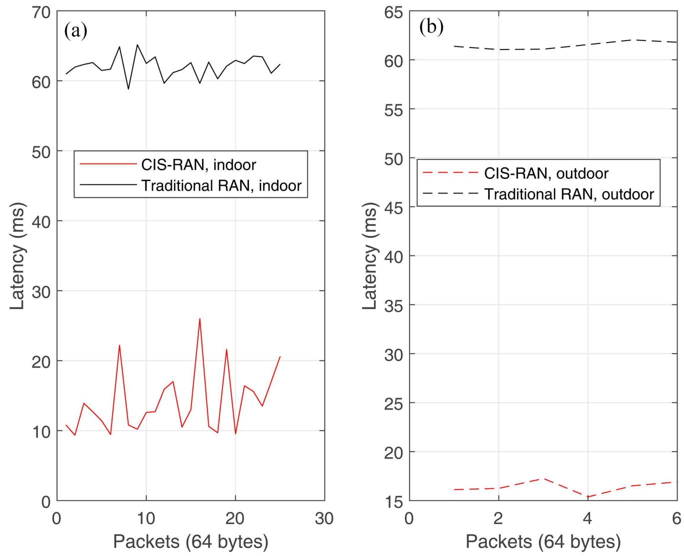

Figure 13 (Color online) End-to-end latency performance comparison. (a) Indoor scenario; (b) outdoor scenario.

prediction, energy-performance trade-offs, and efficient hardware-resource mapping. Additionally, the inherently dynamic nature of service demands and non-stationary network conditions necessitates lifecycle-aware orchestration frameworks with robust adaptive reconfiguration capabilities. Future efforts should focus on online reinforcement learning-based orchestration that can handle sub-millisecond mapping between task-level requirements and physical resources under non-stationary environments. Furthermore, determining the optimal functional split between eCU and eDU for joint sensing-communication streams remains an open problem, particularly concerning the tradeoff between raw data front-hauling and local feature extraction.

Effective interaction mechanisms are crucial for CIS-RAN to enable low-latency, reliable cooperation across communication, sensing, and intelligent processing. This requires redesigning signaling procedures to support adaptive collaboration, including optimizations like synchronization signal block (SSB) procedures, improving RRC signaling, and enabling dynamic, cooperative cluster formation for multi-TRP transmission. Potential directions include the design of low-overhead semantic signaling to replace traditional verbose RRC procedures, and the development of dynamic, cluster-based synchronization protocols that allow multiple TRPs to collaboratively sense and communicate without exhaustive reference signal overhead.

6 Conclusion

This paper has presented CIS-RAN, a holistic RAN architecture designed to overcome the limitations of current networks in the 6G era. By integrating cooperation, intelligence, and service-based design, CIS-RAN transforms traditional C-RAN into an intelligent, adaptive, and extensible platform that synergizes communication, sensing, and computing to meet the heterogeneous demands of emerging 6G applications. To provide a theoretical foundation for architectural design, a task-driven unified modeling and optimization framework is further introduced to abstract the intrinsic coupling among cooperation, intelligence, and service provisioning. Its feasibility was demonstrated through a case study on collaborative AI-based interference prediction, showing that centralized pre-training combined with distributed fine-tuning achieves faster convergence and improved data efficiency compared with isolated local training, underscoring the importance of architectural support for distributed intelligence. In addition, representative prototype-based experimentation further verifies the capability of CIS-RAN to enable end-to-end intelligent service provisioning within the CIS-RAN. While several practical challenges remain, including managing hardware heterogeneity, enabling lifecycle-aware orchestration, and ensuring low-latency cross-domain cooperation, this work provides a significant step toward intelligent, flexible, and service-oriented RAN architectures for the 6G future.

References

1 Schmidt R, Nikaein N. RAN engine: service-oriented RAN through containerized micro-services. IEEE Trans Netw Serv Manage, 2021, 18: 469–481
2 Li N, Liu G, Zhang H, et al. Service-based RAN: the next phase of cloud RAN. In: Proceedings of 2022 IEEE GlobeCom Workshops (GC Wkshps), 2022. 1206–1211

3 Ding H, Wang Y, Zheng X, et al. Design and implementation of a service-based radio access network. In: Proceedings of 2023 IEEE 97th Vehicular Technology Conference (VTC2023-Spring), 2023. 1–5
4 Khan N A, Schmid S. AI-RAN in 6G networks: state-of-the-art and challenges. IEEE Open J Commun Soc, 2024, 5: 294–311
5 Kundu L, Lin X, Gadiyar R, et al. AI-RAN: transforming RAN with AI-driven computing infrastructure. IEEE Commun Mag, 2026, 64: 168–174
6 Huang Y, Li N, Sun Q, et al. Communication and computing integrated RAN: a new paradigm shift for mobile network. IEEE Netw, 2024, 38: 97–112
7 Fang Z, Yuan C, Tian L, et al. D^5RAN: a substrate platform for service-based RANs toward the integration of computing and communication. IEEE Commun Mag, 2025, 63: 32–39
8 Abd R I, Findley D J, Kim K S. Hydra-RAN perceptual networks architecture: dual-functional communications and sensing networks for 6G and beyond. IEEE Access, 2024, 12: 2162–2185
9 I C L, Huang J, Duan R, et al. Recent progress on C-RAN centralization and cloudification. IEEE Access, 2014, 2: 1030–1039
10 Liu F, Masouros C, Petropulu A P, et al. Joint radar and communication design: applications, state-of-the-art, and the road ahead. IEEE Trans Commun, 2020, 68: 3834–3862
11 Wu W, Zhou C, Li M, et al. AI-native network slicing for 6G networks. IEEE Wireless Commun, 2022, 29: 96–103
12 Foukas X, Marina M, Kontovasilis K. Orion: RAN slicing for a flexible and cost-effective multi-service mobile network architecture. In: Proceedings of the 23rd Annual International Conference on Mobile Computing and Networking, 2017. 127–140
13 Polese M, Bonati L, D'oro S, et al. Understanding o-ran: architecture, interfaces, algorithms, security, and research challenges. IEEE Commun Surv Tutor, 2023, 25: 1376–1411
14 International Telecommunication Union (ITU). Framework and overall objectives of the future development of IMT for 2030 and beyond. ITU-D SG2 Q3/2 Workshop on 5G Cybersecurity, 2024. https://www.itu.int/dms pub/itu-d/oth/07/31/D07310000090015PDFE.pdf
15 Chowdhury M Z, Shahjalal M, Ahmed S, et al. 6G wireless communication systems: applications, requirements, technologies, challenges, and research directions. IEEE Open J Commun Soc, 2020, 1: 957–975
16 GSMA. Operator platform requirements and architecture. Technical Specification OPG.02-v8.0, GSMA, 2025. OpenGateway Architecture Framework. https://www.gsma.com/solutions-and-impact/technologies/networks/wp-content/uploads/2025/03/OPG.02-v8.0-Operator-Platform-Requirements-and-Architecture.pdf
17 Li B, Liu T, Wang W, et al. Agent-as-a-service: an AI-native edge computing framework for 6G networks. IEEE Netw, 2025, 39: 44–51
18 O-RAN Alliance. O-ran control, user and synchronization plane specification 18.0. Technical Report, 2025. https://specifications.o-ran.org/specifications
19 Wang B, Zhu F, Liu M, et al. Multi-sources information fusion learning for multi-point NLOS localization. In: Proceedings of 2024 IEEE 99th Vehicular Technology Conference (VTC2024-Spring), 2024. 1–6
20 Wei Z, Liu H, Yang X, et al. Carrier aggregation enabled integrated sensing and communication signal design and processing. IEEE Trans Veh Technol, 2024, 73: 3580–3596
21 Vandikas K, Larsson H, Ickin S, et al. On the adoption of distributed intelligence toward 6G networks. IEEE Access, 2025, 13: 149035
22 Yang Y, Wu J, Chen T, et al. Task-oriented 6G native-AI network architecture. IEEE Netw, 2024, 38: 219–227
23 O-RAN Alliance. O-RAN acceleration abstraction layer general aspects and principles v11.00. Technical Report O-RAN.WG6.AAL-GAP-v11.00, O-RAN Alliance, 2025. https://specifications.o-ran.org/specifications
24 You X H, Huang Y M, Zhang C, et al. When AI meets sustainable 6G. Sci China Inf Sci, 2025, 68: 110301
25 3GPP. Functional architecture and information flows to support common API framework for 3GPP northbound APIs. TS 23.222, 2025. https://www.3gpp.org/ftp/Specs/archive/23 series/23.222/23222-j50.zip
26 3GPP. Procedures for the 5G system (5gs). TS 23.502, 2025. https://www.3gpp.org/ftp/Specs/archive/23 series/23.502/23502-j40.zip
27 Foukas X, Radunovic B, Balkwill M, et al. Taking 5G RAN analytics and control to a new level. In: Proceedings of the 29th Annual International Conference on Mobile Computing and Networking, 2023. 1–16
28 Huang Y, Xie Y, Chen Z, et al. AI empowered modeling, closed-loop optimization, and field trials of RAN digital twin. IEEE Netw, 2026, 40: 154–164