SCIENCE CHINA
Information Sciences

• PERSPECTIVE •

October 2026, Vol. 69, Iss. 10, 206301:1–206301:3
https://doi.org/10.1007/s11432-025-4990-9

6G RAN cloud platform: a foundation for user-centric intelligent 6G

Nan LI[1,2], Xiaohua ZHANG[2], Qi SUN[2], Zhi WANG[2], Baichun ZHAO[2], Ting LI[2], Weichen NI[2], Jinri HUANG[2], Yuhong HUANG[2], Chih-Lin I[2], Mugen PENG[1] & Xiaoyun WANG[2]*

[1]*State Key Laboratory of Networking and Switching Technology, Beijing University of Posts and Telecommunications, Beijing 1000876, China*
[2]*China Mobile Research Institute, Beijing 100053, China*

Received 4 November 2025/Revised 7 March 2026/Accepted 3 June 2026/Published online 18 August 2026

Citation Li N, Zhang X H, Sun Q, et al. 6G RAN cloud platform: a foundation for user-centric intelligent 6G. Sci China Inf Sci, 2026, 69(10): 206301, https://doi.org/10.1007/s11432-025-4990-9

The next generation of mobile services is poised to encompass a diverse array of AI-native applications such as extended reality (XR), holographic communication, autonomous driving, and embodied intelligent agents. To meet their demanding performance requirements and therefore provide extreme user experience for the end users, 6G should undergo a fundamental paradigm shift from the traditional cell-centric architecture to a truly user-centric network (UCN) design. In literature, the prevailing user-centric paradigm has predominantly focused on the optimization of physical transmission layer, which is insufficient to maximize the potential of UCN. Moreover, the realization of intelligent UCN is constrained not only by algorithmic challenges but, more fundamentally, by the limitations of current cloud infrastructures. In this perspective it is proposed that a genuine UCN should be built upon three hierarchically dependent awareness capabilities: spatial, context, and intent awareness. Furthermore, enabling these capabilities necessitates a dedicated, AI-native 6G RAN cloud platform as its indispensable practical foundation.

User-centric transmission and sensing, resource management, and service provision. The user-centric nature of a UCN is fundamentally realized across three hierarchical levels: transmission and sensing level, resource management level, and service provisioning level. Correspondingly, we proposed that to realize the user-centric feature at different level requires three key awareness capability: spatial awareness, context awareness, and intent awareness.

The first and foundational awareness capability is spatial awareness, which is essential to realize the user-centric transmission and sensing. This capability is to enable the network to precisely perceive physical environment, such as physical locations, velocity vectors, and real-time channel state information (CSI) in order to provide optimal transmission and sensing. Spatial awareness requires two critical enabling technologies: cell-free coordinated transmission and cooperative integrated sensing and communication (ISAC). Technologies like cell-free massive MIMO leverage spatial awareness to dynamically form a serving cluster of access points (APs) for each user, based on which coordinated multipoint (CoMP) transmission could be implemented to mitigate the inter-cell interference to ensure high-quality signal reception for users [1]. On the other hand, with cooperative ISAC, signals reflected from the environment and users can be fused to construct high-resolution environmental maps and track user kinematics with high precision, which is essential for resource management [2].

Building upon the real-time spatial data feed, the second awareness capability is context awareness, which is essential to realize user-centric resource management. It is the capability of fusion and analysis of multi-dimensional data streams, including user requirements and network conditions, to predict future states and behaviors. It thereby empowers the network to transition from reactive adaptation to proactive and predictive resource optimization. Context awareness involves the fusion and analysis of multi-dimensional data streams, including user and service context such as dynamic quality of experience (QoE) requirements, service types, and network and environment Context (e.g., real-time traffic load, interference conditions). Machine learning (ML) algorithms are pivotal in correlating these disparate data sources to predict future network states and user behavior. Based on ML-driven decisions, the UCN could executes dynamic adjustments at both resource level (e.g., spectrum, power, time slots) and function level (e.g., function instantiation, scale-in, scale-out).

The highest and most transformative awareness capability is intent awareness, which is essential to realize the user-centric service provisioning. It parses high-level business abstract goals or user requests directly into quantifiable, executable and verifiable network policies. Then it further manages the complexity of user-specific service customization through autonomous closed-loop optimization.

A more detailed description could be found in Appendix A.

Mapping UCN visions to 6G RAN cloud platform requirements. While cloud-native principles offer undeniable benefits in flexibility and scalability, general-purpose cloud platforms are architecturally misaligned with the core demands of the three-layer UCN framework. Traditional cloud platforms are optimized for statistical multiplexing and best-effort workloads, not for the deterministic,

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

hard-real-time, and tightly synchronized operations required by the RAN, especially in a UCN context. To draw the requirements over cloud platform, the idea is to map these abstract "awareness" visions of UCN into concrete network functions (NFs) and furthermore, the underlying cloud platform requirements for infrastructure as a service (IaaS) and platform as a service (PaaS) layers. As illustrated on the left side of Figure 1, this hierarchical mapping acts as a logical translator, where each awareness pillar imposes stringent deterministic demands.

6G RAN cloud platform requirements to support spatial awareness: Spatial awareness coordinated communication and sensing encompass the essential functions of joint transmission and real-time environmental perception. To realize this capability, the 6G cloud-native platform should be able to support the coherent and real-time coordination of distributed radio and compute resources. This overarching requirement includes two aspects. On the one hand, it demands a deterministic real-time foundation for IaaS, encompassing (1) sub-microsecond synchronization to enable coherent joint transmission and sensing, and (2) deterministic ultra-low latency processing for real-time functions. On the other hand, it requires (3) high-throughput networking on IaaS layer for massive data exchange between distributed units, and (4) latency-bounded computation on PaaS layer to execute AI-native sensing and channel prediction models within the RAN slot cycle.

6G RAN cloud platform requirements to support context awareness: By leveraging multi-dimensional context information provided by spatial awareness or intent-derived network policies, the UCN control plane could predict user demands and network fluctuations, dynamically adapting resources to optimize personalized service performance. To enable this predictive optimization, the 6G cloud-native platform should meet a core, overarching requirement: enabling dynamic, stateful, and highly elastic resource management. This overarching requirement logically includes two key aspects. First, it demands agile resource control and provisioning, which encompasses (1) granular resource guarantees and isolation for IaaS and (2) dynamic computing resource management as well as observability for PaaS. Second, it should ensure seamless service continuity, which requires (3) support for stateful mobility, facilitated by (4) high-throughput intra-BBU networking for efficient state synchronization on the IaaS layer.

6G RAN cloud platform requirements to support intent awareness: The cloud platform should provide an autonomous, AI-driven orchestration framework capable of converting abstract goals into real-time, executable network policies or configurations, which could further guide context-aware optimization. This capability is built upon two essential pillars for the PaaS layer. The first is the execution engine, demanding the ability to implement translated policies with extreme speed and reliability, enabling rapid and responsive service instantiation. This requires 6G RAN cloud platform enhancing (1) real-time resource availability monitoring and (2) dynamic resource scheduling. The second is the coordination fabric, requiring a robust and scalable system for AI agents to manage these actions, which is achieved through an (3) agent workflow and (4) agent gateway.

A more detailed description could be found in Appendix B.

6G RAN cloud platform architecture. These requirements, derived from the three foundational pillars of spatial, context, and intent awareness and detailed in the previous section, cannot be fulfilled by generic cloud architectures. To bridge this critical implementation gap and provide a practical foundation for deploying UCN, a 6G RAN cloud platform architecture is proposed and shown on the right side of Figure 1. In this architecture, components that are enhanced beyond the O-RAN O-Cloud baseline are highlighted in light blue within both the IaaS and PaaS layers, while entirely new capabilities introduced for the 6G RAN cloud platform are marked in orange.

IaaS. The IaaS layer of the 6G RAN cloud platform provides the essential hardware and software infrastructure resources crucial for deploying and operating 6G network functions and applications. The hardware encompasses compute, network, and storage components along with heterogeneous hardware accelerators like FPGAs, GPUs, and DPUs. The software includes a real-time operating system (OS), container runtime, network components, and container orchestration control plane components. A key new component is the real-time operating system equipped with a preemptive kernel. It incorporates technologies like CPU isolation, IRQ affinity, and CPU pinning to guarantee the deterministic scheduling latency required for 6G RAN, ensuring tasks like data processing to meet the sub-10 microsecond target.

The container runtime provides the environment for containerized workloads, offering a lightweight alternative for deployment. To meet stringent networking demands, specialized network components are integrated. This includes the OVS+DPDK solution, which bypasses the kernel network stack to achieve high-speed, low-latency packet processing vital for RAN, and SR-IOV, which enables direct hardware access to NICs for near-native performance and isolation. Logically above these, Kubernetes CNI plugins orchestrate network connectivity for pods, providing control-plane agility for dynamic UCN clusters.

Finally, the container orchestration control plane components, with Kubernetes as the reference design, manage the deployment, scaling, and operation of containers. Its key elements (API server, controller manager, scheduler) ensure reliable management of containerized 6G RAN workloads. Together, this IaaS layer forms the deterministic, high-performance foundation supporting the hierarchical awareness framework of UCN.

PaaS. The PaaS layer of the 6G RAN cloud platform acts as the intermediary between the IaaS substrate and containerized RAN functionalities, providing the essential tools and services to manage application complexity without the need for manual infrastructure tuning. A primary objective of this layer is to facilitate the development and deployment of intelligent, user-centric applications by abstracting resource management. In addition to standard services such as scalable databases, message queues for asynchronous coordination, and package management, several specialized functionalities are proposed in the architecture.

First, to realize the high-level automation required by the UCN paradigm, we propose the integration of several new PaaS components. For example, an agent gateway is introduced as a critical broker for agent handshakes, supporting the agent-to-agent (A2A) protocol and the model context protocol (MCP). This gateway translates A2A identities and MCP context into secure, policy-governed calls, allowing models to discover tools while maintaining audit trails and cost caps. Complementing this is the newly proposed MLOps Toolset, which integrates machine learning workflows into network operations to manage the lifecycle of models used for RAN and edge applications.

Furthermore, we propose the inclusion of a service mesh specifically tailored to accelerate communication between microservices within the RAN functions and edge AI applications. For effective system management, the hardware accelerator management (HAM) is proposed to handle the mapping of heterogeneous silicon resources to specific RAN and AI containers. Finally, an observability toolset is integrated to provide high-resolution monitoring of performance metrics and traffic loads, enabling proactive scaling and energy-saving optimizations through generative AI and advanced ML technologies. These components collectively form the foundation for the agent workflow, ensuring that the network remains pinned to the user's intent.

A more detailed description could be found in Appendix C.

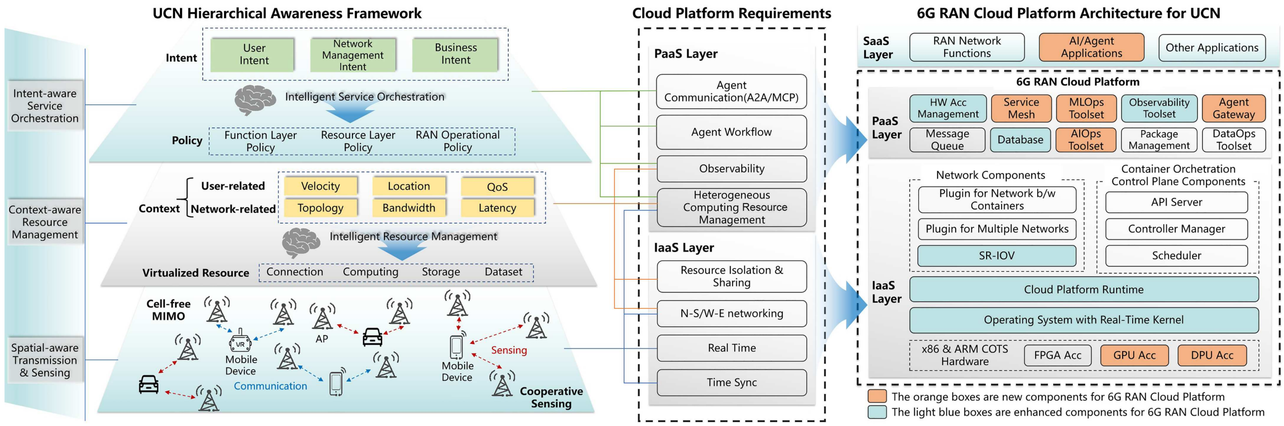

Figure 1 (Color online) UCN hierarchical awareness framework, 6G RAN cloud platform requirements and architecture.

Experimental validation. An open 6G RAN cloud platform testbed was developed and comprehensive experimental evaluation was conducted to empirically validate the capability to serve for UCN by meeting the stringent, synergistic requirements derived from its awareness framework. The implemented testbed, as summarized in Appendix D, integrates heterogeneous processing units (CPUs, GPUs, FPGAs, DPUs) and leverages mature open-source cloud components, including a real-time Linux kernel and Kubernetes, to create a reproducible cloud-native RAN environment.

The validation has been carried out systematically from four critical dimensions. First, real-time performance tests assessed the platform's deterministic behavior under varying CPU loads. With 11 test cases and 1.38 billion sampled interrupts per test case, the results confirmed that the maximum interrupt scheduling latency remained at or below 10 ms in all test cases, with an average latency of approximately 2 ms. This meets the sub-10 microsecond budget essential for maintaining RAN stack state within a slot duration, thereby fulfilling the ultra-low latency processing requirement for spatial-aware cooperative transmission and beamforming in dynamic UCN clusters.

Second, clock synchronization tests utilizing the precision time protocol (PTP) demonstrated the platform's ability to achieve the sub-microsecond synchronization precision mandated for coherent multi-point operation. Test results have shown that after a brief convergence period, the time offset between nodes stabilized consistently within a narrow range of 0–7 ns, far exceeding the $\leqslant$65 ns target [3, 4]. This high-precision synchronization is fundamental for enabling APs to function as a coordinated antenna array, preventing signal interference and underpinning joint sensing and communication.

Third, container network performance tests evaluated two key data plane technologies to support massive data flows. SR-IOV technology achieved the full 100 Gbps line rate, providing the deterministic, near-native performance required for radio-rate fronthaul traffic (spatial awareness) and high-throughput east-west data exchange between collaborating BBUs (context awareness). In contrast, the OVS+DPDK provided the software-based programmability required for complex east-west traffic and multi-tenant isolation, albeit with a lower throughput of 30 Gbps. The platform's architecture incorporating both provides a balanced dual-plane networking solution for diverse UCN workload needs.

Finally, AI capability and GPU power consumption tests focused on dynamic resource management for context-aware elasticity. By employing multi-instance GPU (MIG) technology, the platform enabled fine-grained partitioning and on-demand binding of GPU resources. A key finding was that an optimal partitioning strategy (using 4 MIG instances and binding 2) reduced GPU power consumption by approximately 15% compared to a less granular approach, despite serving the same number of active containers. This demonstrates the platform's inherent support for intelligent, power-efficient, and elastic computing resource management, which is crucial for aligning AI acceleration resources dynamically with user and service context.

A more detailed description could be found in Appendix D.

Conclusion. This work presented the 6G RAN cloud platform as a foundational enabler for user-centric intelligent 6G networks. Three key awareness capabilities, including spatial, context and intent awareness, have been proposed to realize UCN, which are further mapped and derived the requirements on the underlined 6G RAN cloud platform. To realize these capabilities to achieve the UCN vision, a dedicated cloud-native platform architecture is proposed with enhanced or newly introduced components, structured into an IaaS layer providing deterministic real-time performance, high-throughput networking, and heterogeneous acceleration, and a PaaS layer offering intelligent orchestration and MLOps toolkits.

Comprehensive experimental validation on an open testbed confirms the platform's practical viability. It demonstrates the ability to meet the synergistic demands of the UCN: sub-10 microsecond deterministic scheduling, nanosecond-precision clock synchronization, 100 Gbps networking throughput, and efficient, elastic AI resource management via technologies like MIG. This work provides not only a conceptual roadmap but also a concrete, empirically validated architectural blueprint, which serves as a critical stepping stone towards user-centric intelligent 6G.

Acknowledgements This work was supported by National Science and Technology Major Project of China (Grant No. GXB-2-2025-5).

Supporting information Appendixes A–E. The supporting information is available online at info.scichina.com and link.springer.com. The supporting materials are published as submitted, without typesetting or editing. The responsibility for scientific accuracy and content remains entirely with the authors.

References

1 Elhoushy S, Ibrahim M, Hamouda W. Cell-free massive MIMO: a survey. IEEE Commun Surv Tutorials, 2021, 24: 492–523
2 Ade Krisna Respati M, Lee B M. A survey on machine learning enhanced integrated sensing and communication systems: architectures, algorithms, and applications. IEEE Access, 2024, 12: 170946
3 Cao Y, Wang P, Zheng K, et al. Experimental performance evaluation of cell-free massive MIMO systems using COTS RRU with OTA reciprocity calibration and phase synchronization. IEEE J Selected Areas Commun, 2023, 41: 1620–1634
4 O-RAN Alliance. O-RAN fronthaul control, user and synchronization plane specification. Wg4.cus.0-v07.00, 2023. https://www.o-ran.org/specifications