Xiaodong DUAN, Zhenglei HUANG, Shiyu LIANG, Shaowen ZHENG, Lu LU, Tao SUN, 2025. AI-agent communication network for 6G: vision, architecture, and key technologies. *Frontiers of Information Technology & Electronic Engineering*, 26(11):2065-2080. https://doi.org/10.1631/FITEE.2500582

AI-agent communication network for 6G: vision, architecture, and key technologies

Key words: Artificial intelligence agent; Sixth-generation mobile networks; Network architecture; Multimodality interaction; Multi-agent coordination

Corresponding author: Tao SUN
E-mail: suntao@chinamobile.com
ORCID: https://orcid.org/0009-0003-3491-8813

Motivation

Demands and challenges for 6G networks in empowering AI agents:

1. The service objects are shifting from people to heterogeneous AI agents.

2. The interaction content and mode of AI agents differ from those of human beings.

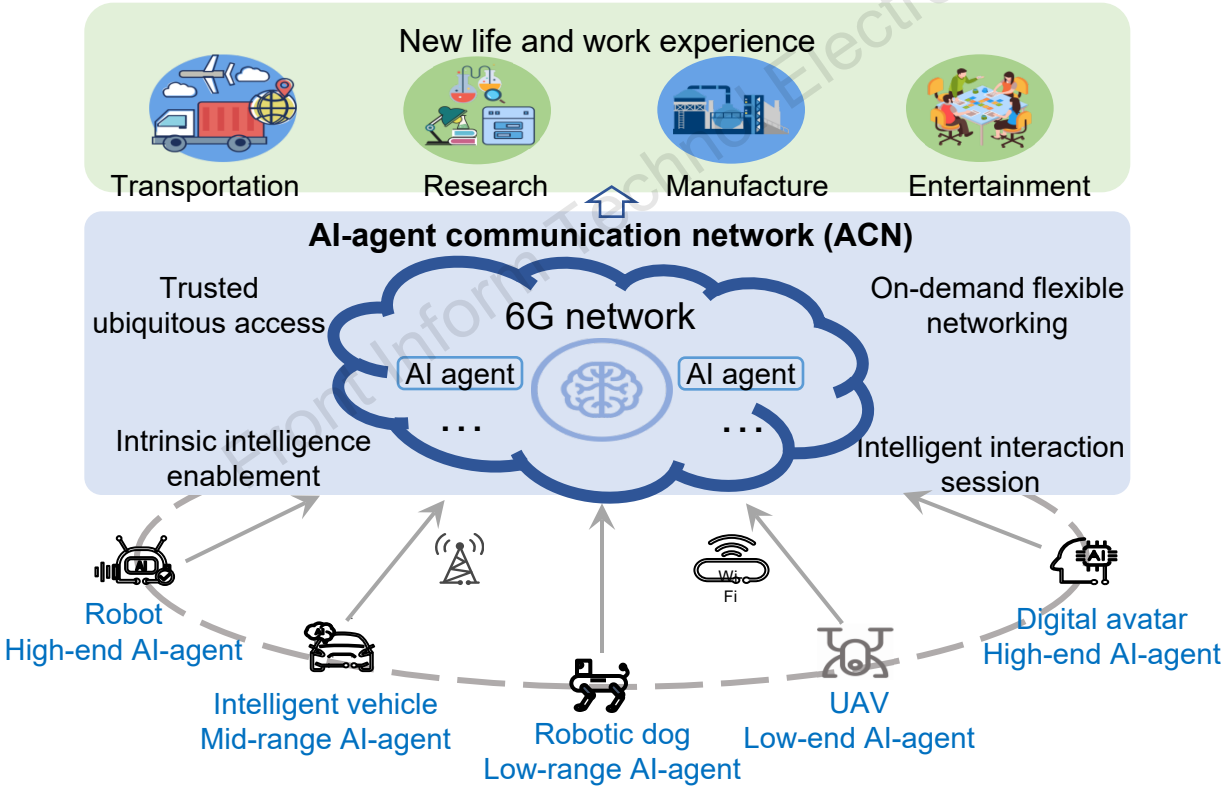
3. There are new capability requirements of AI agents beyond connectivity, such as computing and sensing capabilities.

Considering the above challenges of 6G networks in empowering AI agents, the concept of AI-agent communication network (ACN) is proposed which is a new paradigm to enable secure global information interaction and on-demand capability provisioning for single or multiple AI agents.

Vision

ACN includes the following capabilities:

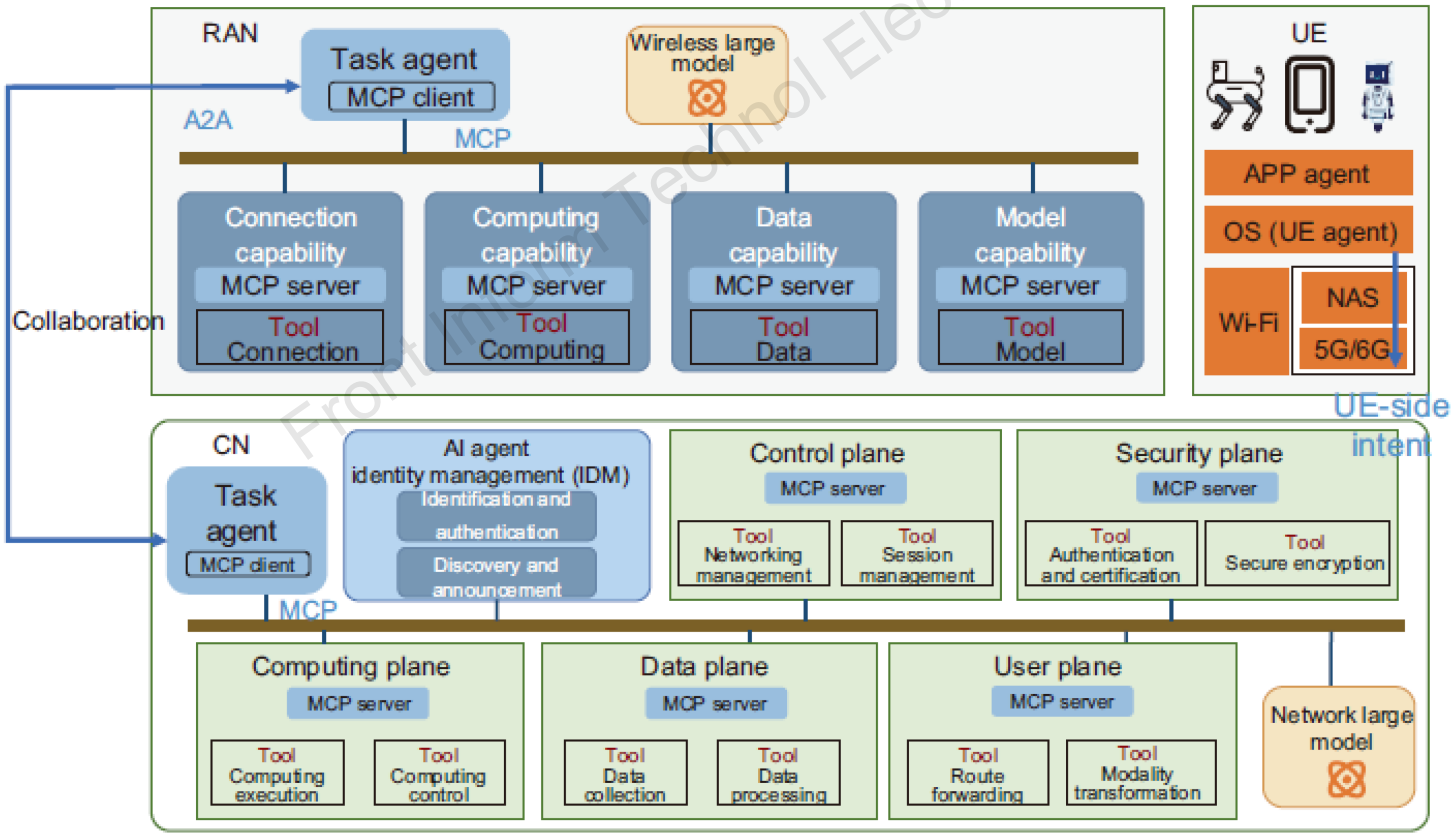
1. Trusted ubiquitous access.
2. On-demand flexible networking.
3. Intelligent interaction session.
4. Intrinsic intelligence enablement.



Architecture

1. Model context protocol (MCP) servers are deployed on five planes (i.e., control plane, user plane, data plane, computing plane, and security plane) and the original atomic services are converted into invocable tools for MCP servers.

2. Task agents are introduced to support user intent recognition and task decomposition, and they act as MCP clients to request various services from the MCP servers deployed on the five planes.



Technologies

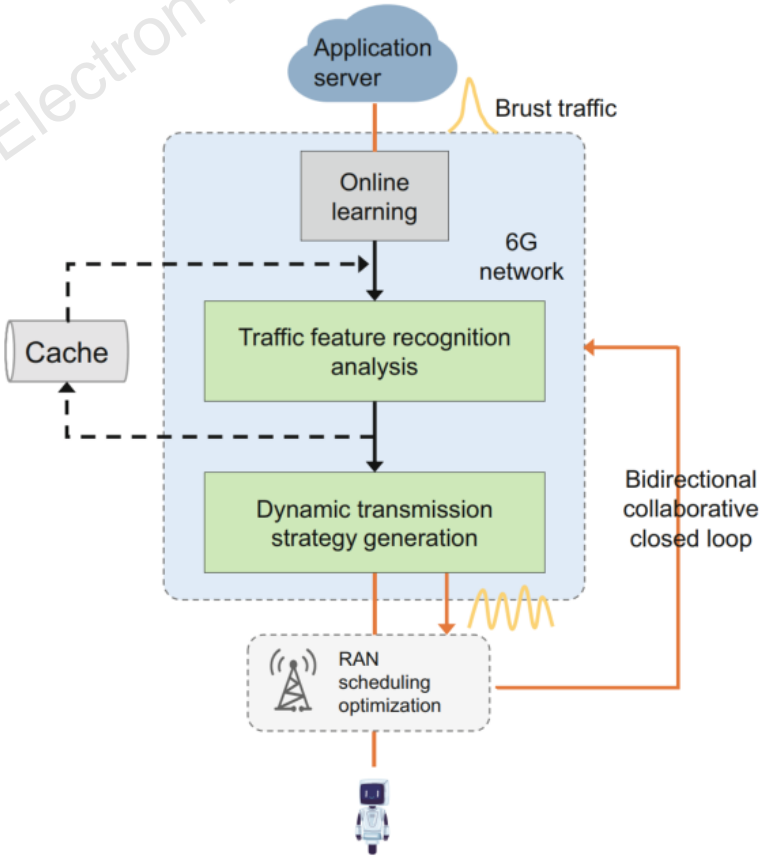
1. Multimodal traffic transmission

A transmission assurance framework based on online learning and dynamic policy generation is proposed.

a) Information on the traffic characteristics is exchanged through network-service collaboration.

b) Service feature recognition is achieved through AI algorithms. Building on these features, dynamic transmission policies are generated to guide resource scheduling on the RAN side.

c) Through real-time resource allocation based on fine-grained service perception, the exact scheduling requirements of multimodal traffic for AI agents are satisfied.



Technologies (Cont.)

2. Design of AI agent identity

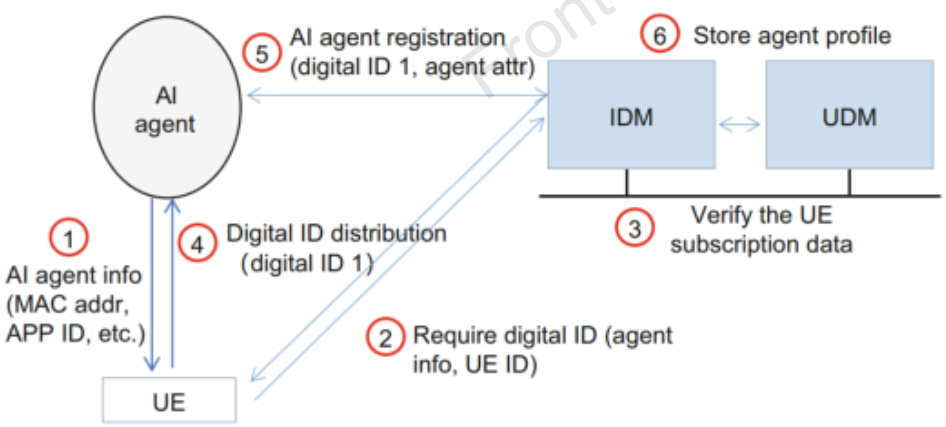
Inspired by the design principle of the uniform resource identifier (URI) and DIDs, the proposed AI agent identity (ID) syntax is as follows:

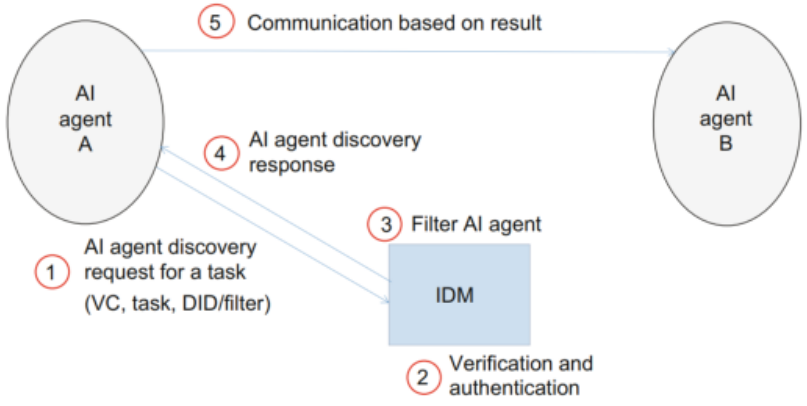
Scheme: Method: Method-specific string.

3. AI agent registration and discovery

a) The UE can obtain a digital ID for the AI agent from the IDM by providing its own UE ID.

b) An AI agent sends a discovery request to the IDM for a specific task. The IDM queries its local database to identify agents that fulfill the specified criteria.



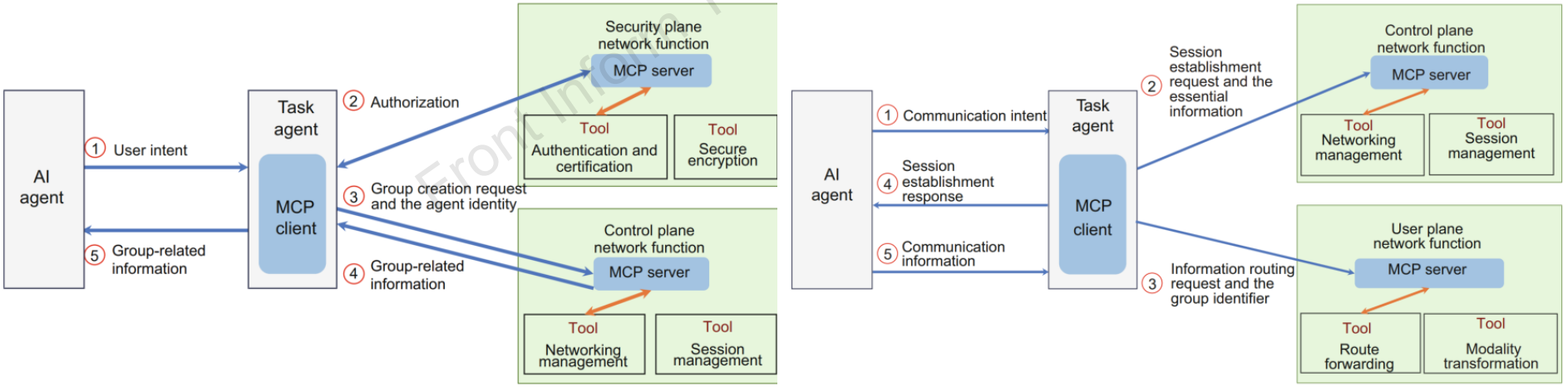


Technologies (Cont.)

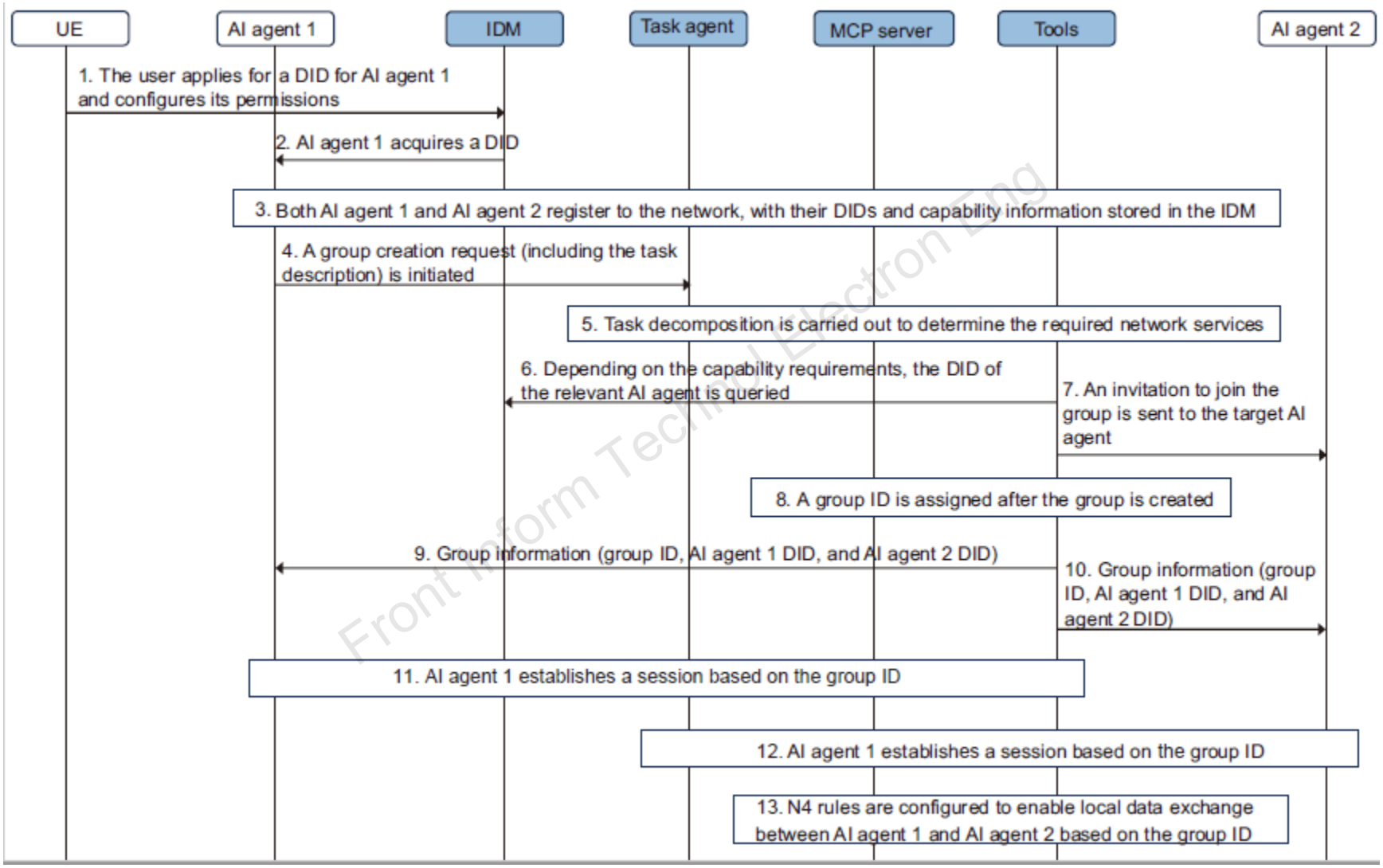
4. Dynamic networking on demand

a) Users can create, delete, and update AI agent communication groups via non-access stratum (NAS) signaling, and their information can be routed based on the corresponding group identifiers.

b) When an AI agent within a group intends to communicate with another AI agent in the same group, it establishes a group session with the target AI agent.



Overall process



Evaluation results

These evaluations are performed under two distinct scenarios: with and without the integration of large model inference executed by the Qwen3 model.

1. The latency of CN transmission and tool invocation meets the stringent low-latency requirements of 6G networks.
2. The proposed architecture and workflow meet stringent low-latency requirements, with no significant latency increase as the number of group members grows.
3. The issue of high inference latency can be mitigated via hardware upgrades and joint optimization methods.

Table 3 Latency of the robot registration process (without large-model inference)

Procedure	Latency (ms)		
	First test	Second test	Third test
Applying for a DID	157	151	159
Registering with a DID	81	78	82

DID: decentralized identifier

Table 4 Latency of the robot registration process (with large-model inference)

Procedure	Latency (s)		
	First test	Second test	Third test
Applying for a DID	9.865	6.207	9.250
Registering with a DID	7.737	8.866	7.492

DID: decentralized identifier

Table 5 Latency of creating AI agent communication groups with varying numbers of robots (without large-model inference)

Number of robots	Latency (ms)		
	First test	Second test	Third test
Two robots	445	453	441
Ten robots	479	485	486

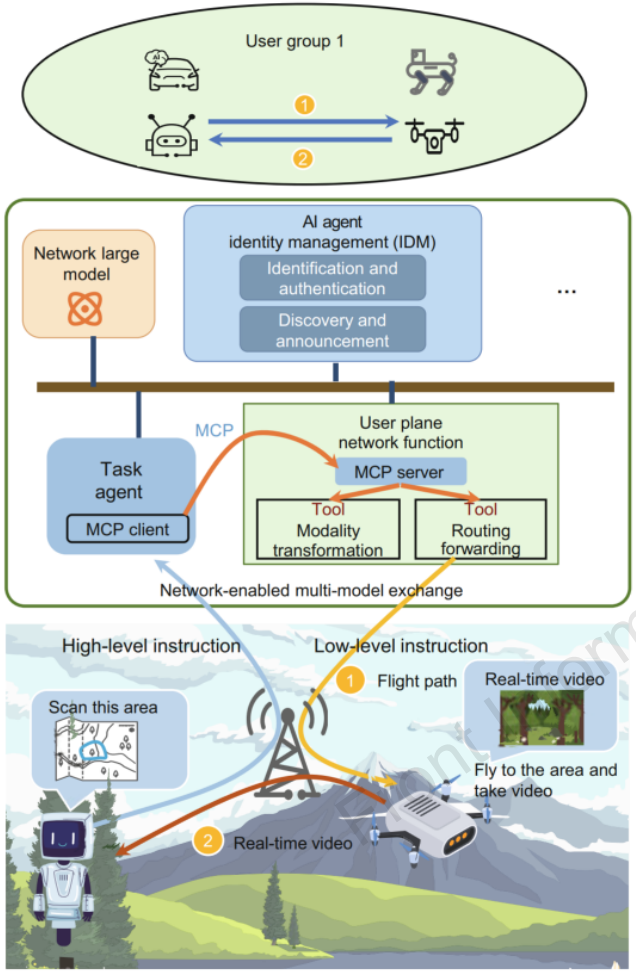
AI: artificial intelligence

Table 6 Latency of creating AI agent communication groups with varying numbers of robots (with large-model inference)

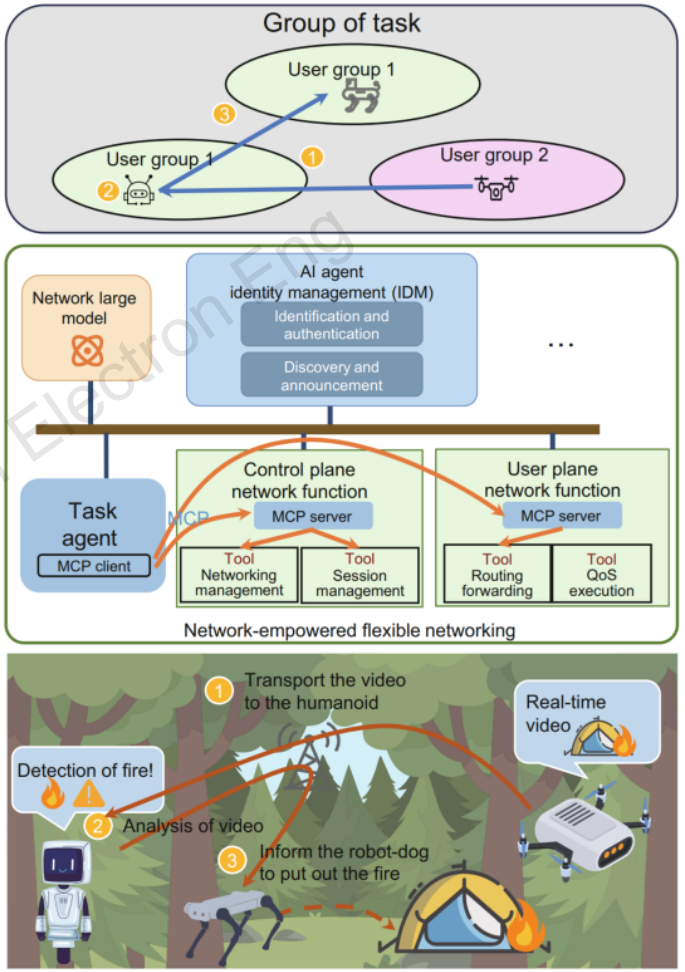
Number of robots	Latency (s)		
	First test	Second test	Third test
Two robots	6.245	8.803	7.852
Ten robots	9.248	8.267	7.467

AI: artificial intelligence

Potential use cases



Taking the selection of camping site when a family goes camping as an example, it demonstrates the communication and collaboration among AI agents within a single group.



Taking the case of multiple families temporarily forming a night security team as an example, it demonstrates the dynamic networking and communication of AI agents among multiple groups.

Conclusions

- ACN represents a new paradigm designed to enable secure information interaction and on-demand capability provisioning for AI agents. The vision of ACN includes trusted ubiquitous access, on-demand flexible networking, intelligent interaction sessions, and intrinsic intelligence enablement.
- An agent-overlay architecture framework and some potential solutions are proposed for ACN. Ultimately, ACN is positioned to become a foundational network service in future mobile networks, providing computing, AI, and sensing capabilities to various AI agents.

Xiaodong DUAN is a Vice President of the China Mobile Research Institute (CMRI), and leader of the network technology group of IMT-2030 (6G). His research interests include 5G/6G architecture, computing force network, and new IP technology.

Zhenglei HUANG is a principal researcher in CMRI, and leader of the ACN work group of CCSA. His research interests include 5G-A/6G network architecture, immersive XR communication, and agent communications.

Shiyu LIANG is a researcher in CMRI. Her research interests include 5G-A/6G network architecture and agent communications.

Lu LU is a Deputy Director of the Department of Basic Network Technology of CMRI, leader of the core network group of CCSA TC5, and a Vice Chairman of ITU-T SG13. Her research interests include 5G-A/6G network architecture and computing force network.

Tao SUN is a Chief Expert of CMRI, and a vice chairman of the SA group in 3GPP. His research interests include mobile network architecture design, IP technology, and computing force network.