



Future Networks Supporting AI Development: From AI in Networks to Networks for Agentic AI

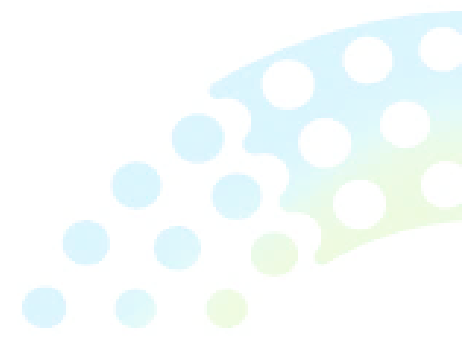
Technical Research and Standardization Considerations

China Mobile



中国移动
China Mobile

中国移动(香港)创新研究院
China Mobile (Hong Kong) Innovation Research Institute



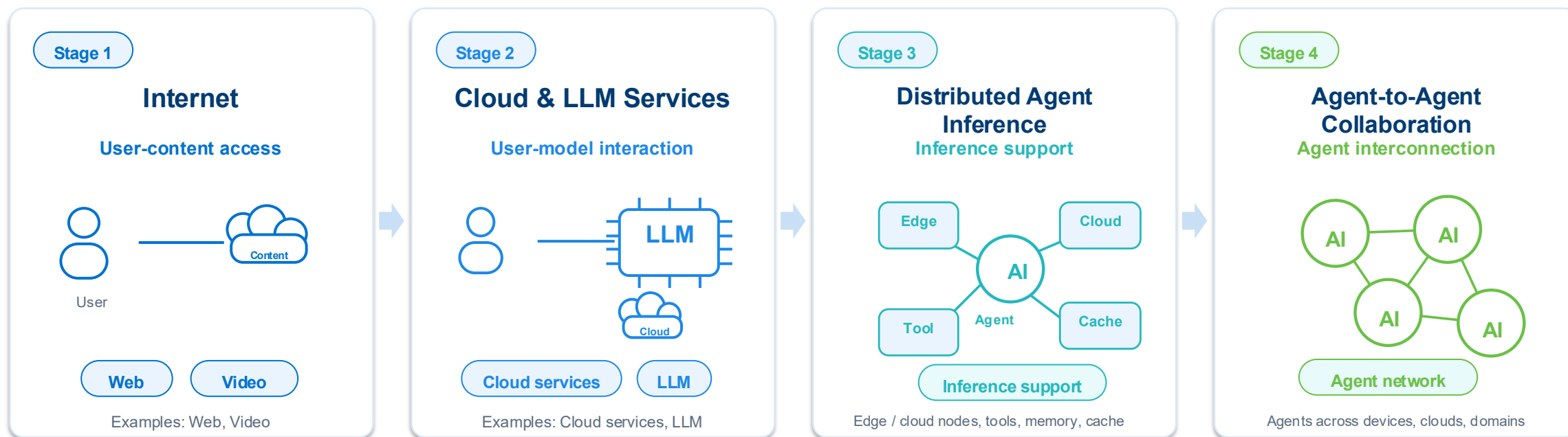
A Concrete Evolution Path: From Content Delivery to Agentic AI Services

AI for Networks

AI-assisted network operation, optimization, automation, assurance, and energy efficiency.

Networks for AI

Distributed inference, computing-network-storage coordination, low latency, high concurrency, and reliable AI delivery.



Two questions for future networks

Q1

How can networks support agent inference?

Q2

How can networks interconnect agents?

Dimension I: Distributed agent inference creates network requirements

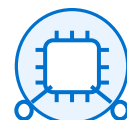
Network support for AI agents starts with the distributed inference services behind them.



Predictable Inference Performance

Stable TTFT and TPOT

High-concurrency requests need stable first-token and per-token latency across nodes.



Efficient Resource Coordination

Computing-network-storage optimization

Coordinate computing, networking, and storage to serve more requests with fewer resources.



Simplified Service Invocation

Simple APIs, hidden complexity

Users and vertical applications call simple services while the network hides scheduling.



Real-Time Knowledge Access

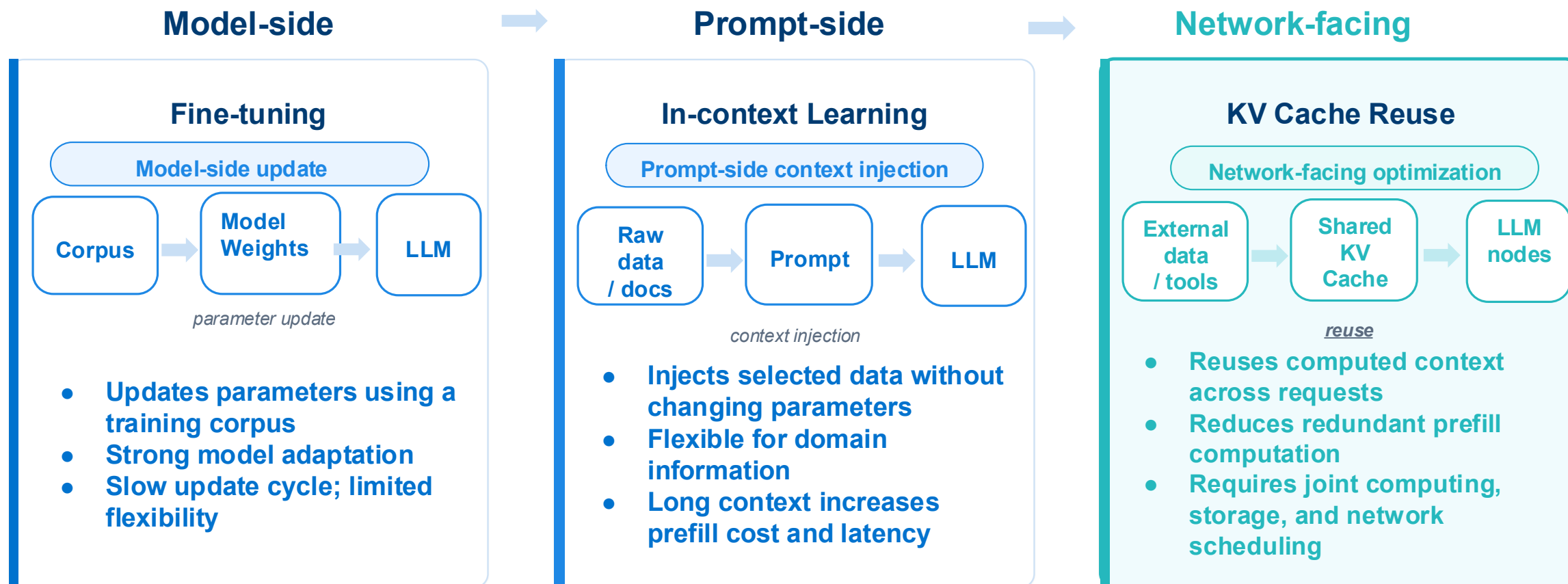
Fresh information and domain expertise

Networks connect agents to real-time data, tools, and knowledge without frequent retraining.

Agent inference is not only a model problem. It is also a network coordination problem.

KV Cache is the network-facing entry point for inference support

Fine-tuning and in-context learning mainly optimize the model or prompt

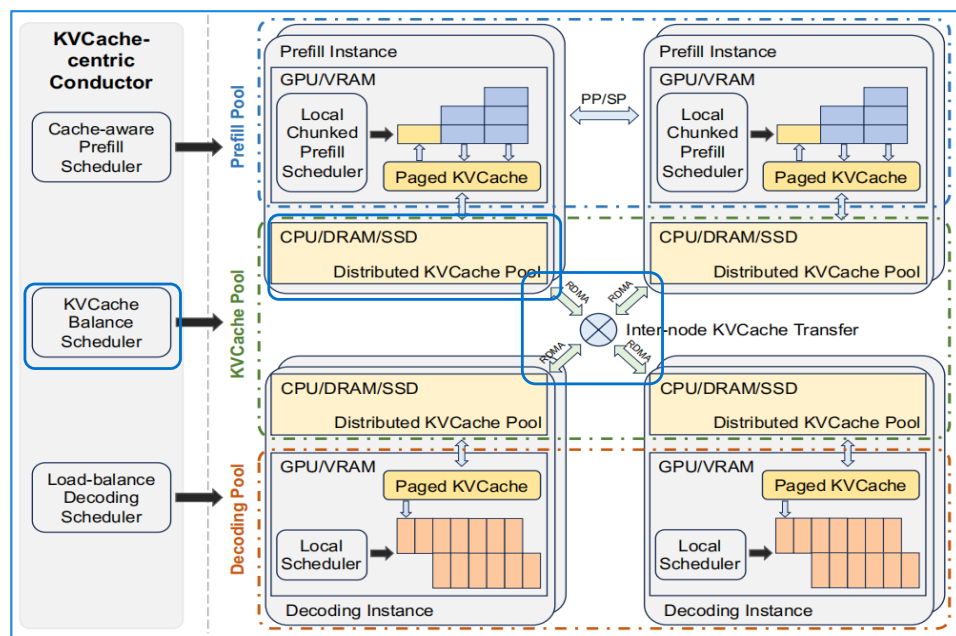


KV Cache turns inference support into a computing-network-storage coordination problem.
Network role: schedule, transmit, store, and reuse KV Cache.

Networks can help schedule, transmit, store, and reuse KV Cache

KV Cache turns agent inference support into a computing-network-storage coordination problem

Technical reference: distributed KV Cache architecture



Schedule

Place cache where it is needed

Coordinate inference requests, KV Cache placement, and distributed inference nodes to reduce redundant prefill computation.



Store

Move warm and cold cache near inference

Offload reusable KV Cache from GPU memory to nearby storage to relieve GPU memory pressure.



Transmit

Move cache across domains

Support high-throughput transfer of large KV Cache data across edge, cloud, and computing nodes.



Reuse

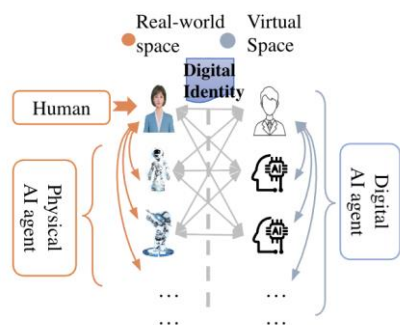
Avoid repeated computation

Reuse computed context across requests, agents, and inference nodes to improve resource efficiency.

Network role: schedule, store, transmit, and reuse KV Cache across distributed inference nodes.

Dimension II: Agent-to-agent collaboration creates network requirements

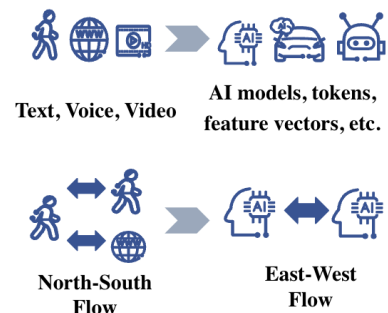
AI agents need identity, dynamic task networking, agent traffic forwarding, and multimodal interaction support.



Identity Management

Unified agent identity

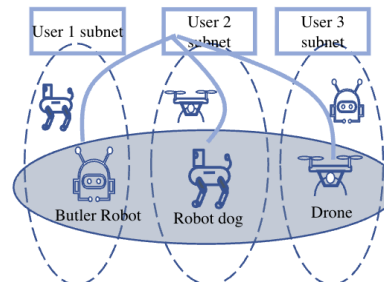
Associate human users, physical agents, and digital agents under unified identities.



Agent Traffic Forwarding

From human media to machine media

Support not only text, voice, and video, but also models, tokens, memory, and feature vectors.



Dynamic Task Networking

Task-oriented agent networks

Agents dynamically form collaboration networks across devices, clouds, and domains.

Instant Search	Remote Guidance	Virtual Tour Guide/Companion
1 interaction per second	Short-term interaction @ minute-level	Long-term interaction @ hour-level
Experience		RTT
Good (1.2s)		50 ms
Excellent (Near-human level 650ms)		20 ms
Ultimate (Human-level 400ms)		10 ms

Multimodal Interaction

Predictable low latency

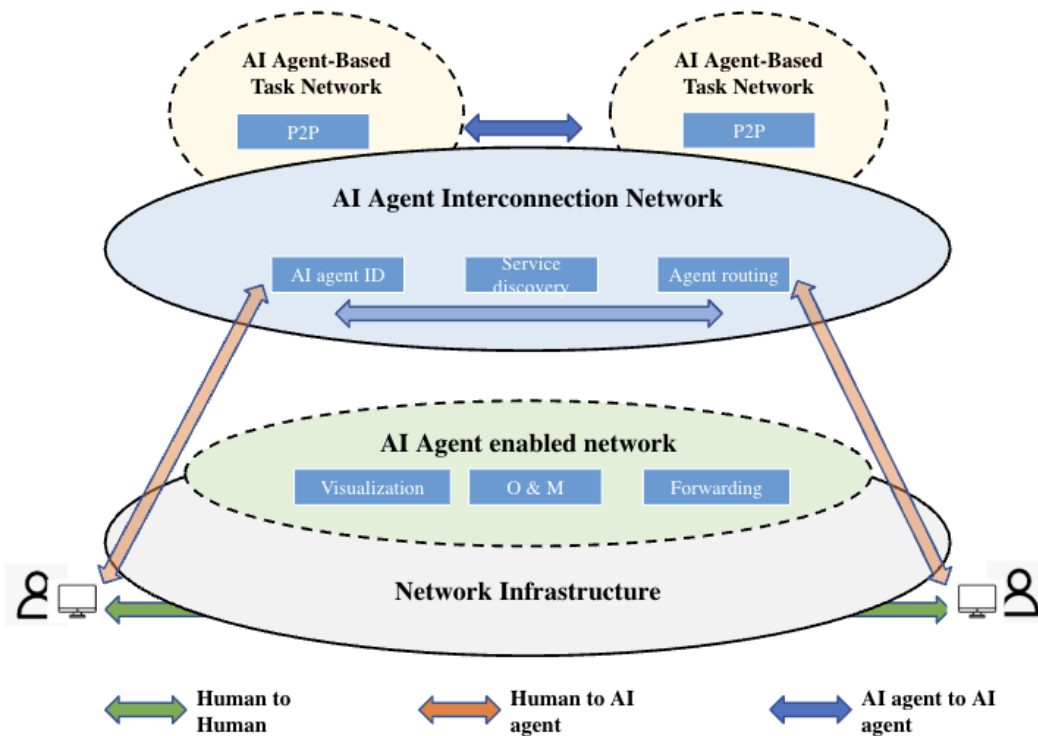
Enable near-human real-time interaction for voice, video, perception, and embodied agent collaboration.

Agent may become first-class network entities.

A layered architecture for AI agent interconnection

Agent interconnection requires layered network capabilities

Reference architecture: AI agent interconnection network



Task Network

Agent-based collaboration

Agents form temporary task networks for specific objectives, based on peer-to-peer interaction.

Interconnection Network

Agent ID, discovery, routing

Provides agent identity, service discovery, and agent routing across domains.

Agent-Enabled Network

Network enhancement

Integrates agent capabilities into network operation, visualization, and forwarding optimization.

Network Infrastructure

Connectivity foundation

Provides the underlying connectivity, transport, and forwarding capabilities.

Layered architecture maps agent requirements into network capabilities: identity, discovery, routing, task networking, and forwarding.

What needs to be standardized?

It should turn inference support and agent interconnection into interoperable network capabilities.

Common Terminology



Define AI agent, agent interconnection, AI agent network, inference support, and related concepts.

Use Cases & Requirements



Cover distributed inference, KV Cache support, agent collaboration, latency, reliability, and security.

Reference Architecture



Place agent IDs, service discovery, routing, task networks, and network enhancement functions.

Interfaces & Procedures



Specify registration, discovery, authentication, authorization, session establishment, task collaboration, and service invocation.

Security



Support cross-vendor and cross-network operation while addressing trust, privacy, compliance, and fallback behavior.

From technical requirements to common building blocks.

Standardization should cover both dimensions: inference support and agent-to-agent communication.

1. Near-term: SG13 - Q20

Scope:

Service-oriented, terminal-side, and application-side agents.

Outputs:

Terminology, scenarios, requirements, network enhancement capabilities, and AI agent definition.

Role:

Convert converged requirements into concrete SG13 work.

2. Exploratory: Focus Group

Scope:

Network-native agent functions.

Outputs:

Management and operations, control plane, user plane, pre-standard research, and candidate work items.

Role:

Explore new network capabilities before formal standardization.

3. Coordination: SGs and SDOs

Scope:

SG17, SG21, JCA AI/ML, 3GPP, IETF, and industry ecosystems.

Outputs:

Security, application scenarios, protocol alignment, architecture alignment, and interoperability.

Role:

Avoid fragmentation across domains.

From common building blocks to coordinated standardization actions.

1 Two dimensions

Future networks need to support both distributed inference behind agents and communication among agents.

2 Two technical entry points

KV Cache is a concrete inference-support entry point; agent interconnection needs identity, discovery, routing, security, multimodal traffic, and layered architecture.

3 Standardization directions

Common terminology, requirements, architecture, interfaces, security, and interoperability are required across vendors, networks, and agent platforms.

China Mobile looks forward to collaborating with industry partners and standardization communities to facilitate innovation in the AI era.

Thank You!

AI agent interconnection is attracting standardization attention

AI agent interconnection is gaining increasing attention across 3GPP, IETF discussions, industry protocols such as MCP/A2A, and China Mobile ACN work.

Proposing the Conceptual Vision for ACN
2024.6

First Discussion of the 3GPP 6G Requirements Project
2024.11

Second Discussion on 3GPP 6G Requirements Project
2025.2

Concept Vision Refinement and Prototype Validation ACN
2025.3

At MWC Shanghai, China Mobile released the white paper "AI Agent Communication Network ACN."

Multiple countries submitted contributions related to AI agents, promoting consensus that AI agents may become a new capability for 6G terminals.

A total of 15 companies from multiple countries submitted AI agent proposals, and the definition of AI agents is introduced for the first time.

At MWC 2025, China Mobile collaborated with 16 global partners to release the 2.0 version of the ACN white paper and showcased the ACN prototype device.

2024.11

2025.03

2025.04

2025.05

2025.06

2025.07

Anthropic released open-sourced the MCP protocol.

Cisco Announces the Open-Source Organization AGNTCY

Google Releases A2A Protocol

Cisco Submits First Paper to the IETF

Coexistence Mode Established: MCP Processing Tool Invoked, A2A Coordination Tool

IETF123 Hosts Multiple Workshops on Agent Interconnection

