Intelligence Distribution Network

(draft-li-cats-idn-00)

Author: Qing Li[1], Hanling Wang[1], Yong Jiang[2,1], Mingwei Xu[3]

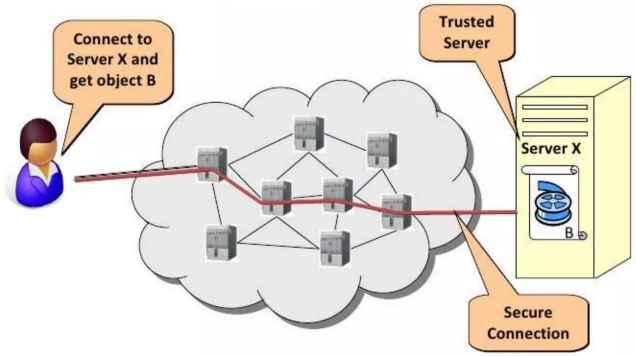
[1]Pengcheng Laboratory, [2]Tsinghua SIGS, [3]Tsinghua DCST

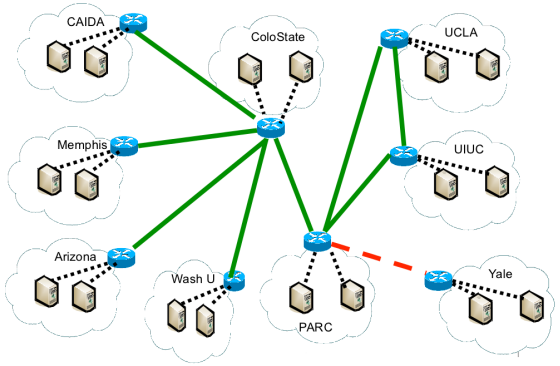
IETF 125 @ Shenzhen, March 2026

Motivation

The Internet is Undergoing a New Paradigm Shift

The core goal of the Internet is shifting from "**establishing connections**", "**distributing content**" towards providing "**intelligent services**" for users.





Connection-centric (1970s-2005)

- **Core Goal：** Establish connections
- **Typical Applications:** remote communication, file transfer, etc
- **Network Capabilities:** End-to-end connectivity
- **Network Characteristics:** Best-effort forwarding

Content-centric (2005-2023)

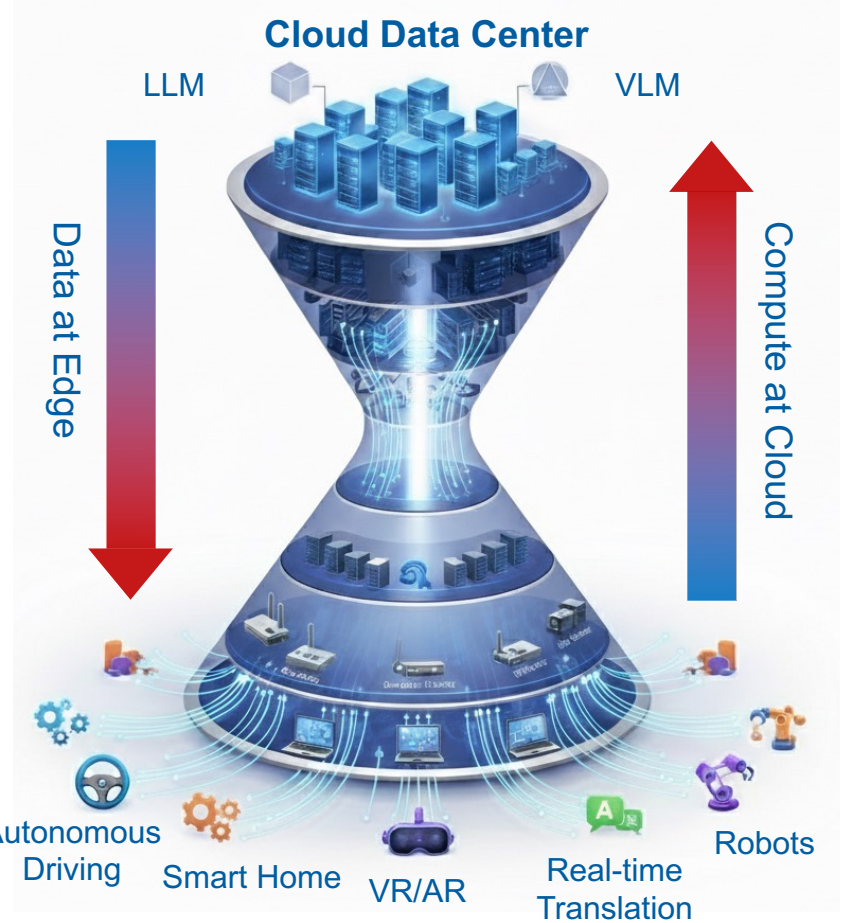
- **Core Goal:** Content distribution
- **Typical Applications:** Web services, streaming media, etc
- **Network Capabilities:** Hierarchical caching and distribution
- **Network Characteristics:** QoS/QoE, and network performance optimization

Intelligence-centric (2023-now)

- **Core Goal:** Provide Intelligence
- **Typical Applications:** Generative AI, LLM agents, etc
- **Network Capabilities:** underexplored
- **Network Characteristics.:** underexplored

Key Challenges

Demand for “Intelligence” originates from the network edge, while computing resources are concentrated in cloud data centers. The current paradigm of “**data transmission via network + centralized cloud inference**” leads to: 1) poor user experience, 2) high network pressure, 3) underutilized computing resources, and 4) high security risks.



Poor User Experience

Cloud-based inference depends on multi-hop network transmission, increasing **end-to-end latency and jitter**, difficult to meet the real-time and stability requirements.

High Network Pressure

Massive volumes of data must be sent back to the cloud, consuming large amounts of network resources, increasing congestion and **limits service scalability**.

Underutilized Resources

Ubiquitous computing resources lack unified abstraction and scheduling mechanisms, leaving **large amounts of computing power idle.**

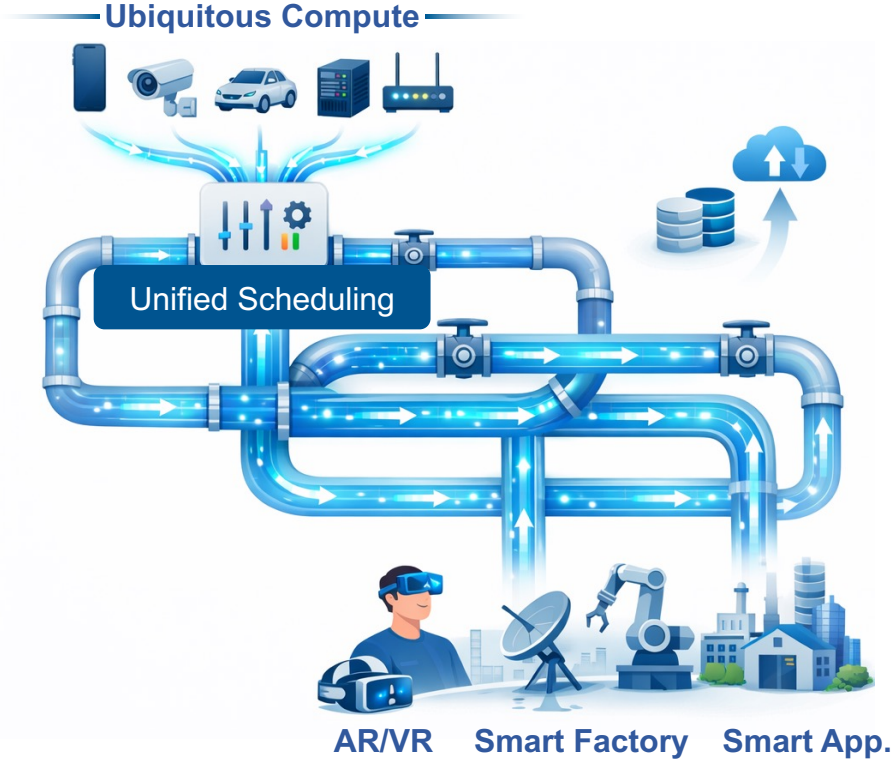
High Security Risks

Users’ private data must be transmitted to the cloud for processing, introducing **privacy risks and data compliance challenges**.

How to design the next-generation Internet architecture for intelligent services?

Goals

Goal: Enable computing power to flow like water and electricity. The use of computing power become plug-and-play and elastically supplied through unified scheduling.



Strategic Direction

- ✓ **Standardization**: specifications for computing networks
- ✓ **Internationalization**: Global deployment of intelligent services
- ✓ **Connection:** cross-domain connection of computing resources

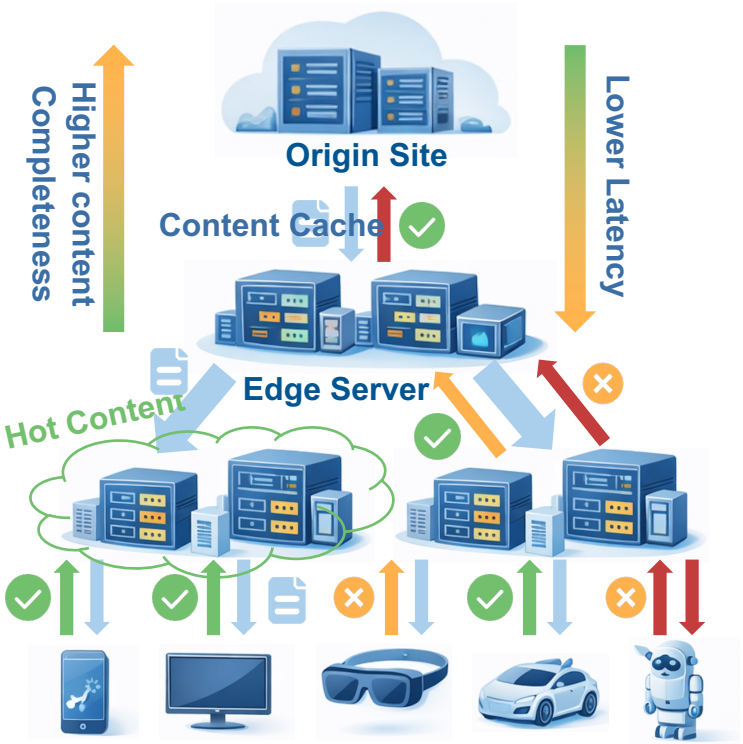
Service Goal

- ✓ **Low Latency:** Reduce latency to improve service quality
- ✓ **High Throughput:** Increase throughput support high concurrency
- ✓ **Low Cost:** Reduce costs while improving overall efficiency

Content Delivery Network (e.g., CDN) enables content to flow like water and electricity. How can the next-generation network architecture enable computing power to flow in the same way?

Key Insight

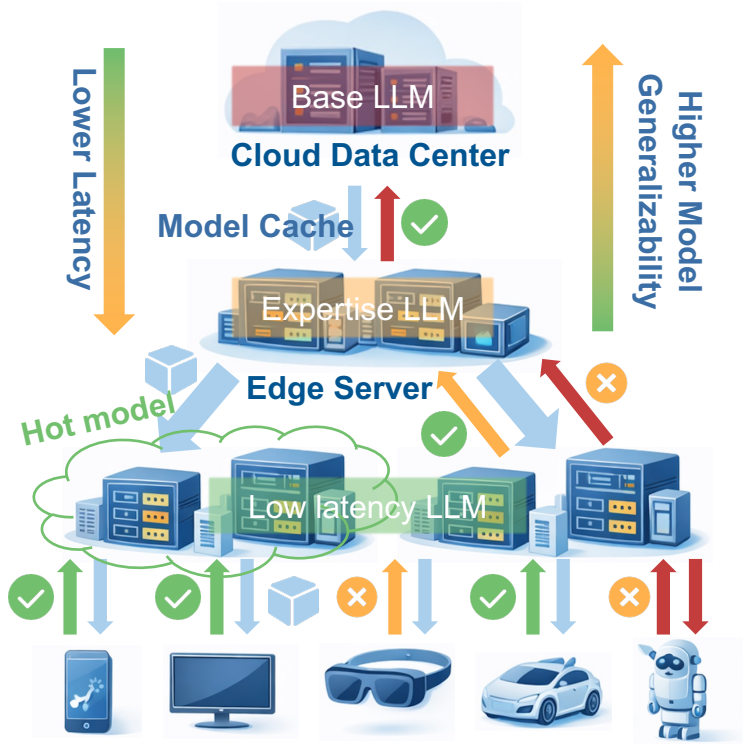
Computing resources cannot flow, but intelligence can. By building a hierarchical intelligent computing interconnection architecture, we can cache popular intelligent services. Similar to CDNs, intelligent service capabilities can be pushed into distributed computing networks.



Content Delivery Network (CDN)

Key Idea

- Local respond first
- Hierarchical Architecture
- Hotspot awareness
- Load balancing



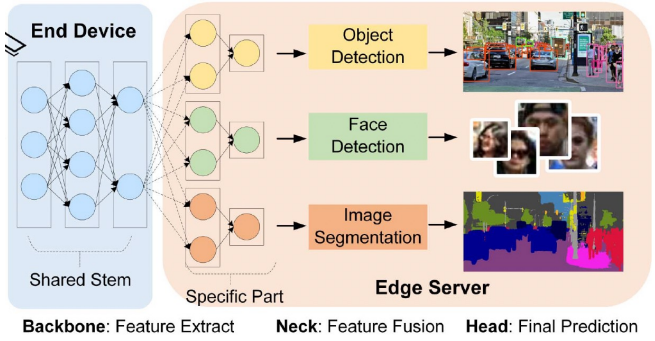
Intelligence Delivery Network (IDN)

Challenges

AI models have **complex structures** and cannot be arbitrarily decomposed; they **require substantial computing resources**, making deployment difficult; intelligent service **requests are massive, dynamic, and heterogeneous**, making it challenging to guarantee high-quality service.

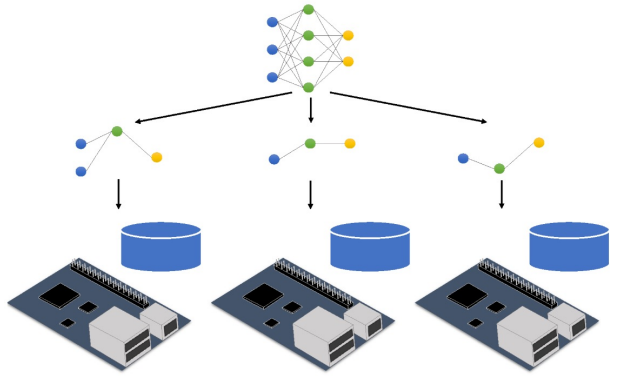
Challenge 1: How to split model?

AI models are structurally complex with highly coupled parameters and strong dependencies between components.



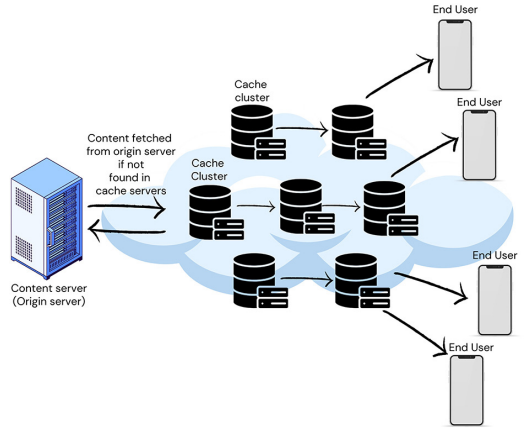
Challenge 2: How to deploy models?

AI models have strict requirements on compute types, memory, and storage capacity and are hard to be deployed on heterogeneous devices.



Challenge 3: How to route requests?

Intelligent requests are massive, dynamic, and heterogeneous, with varying service quality requirements.



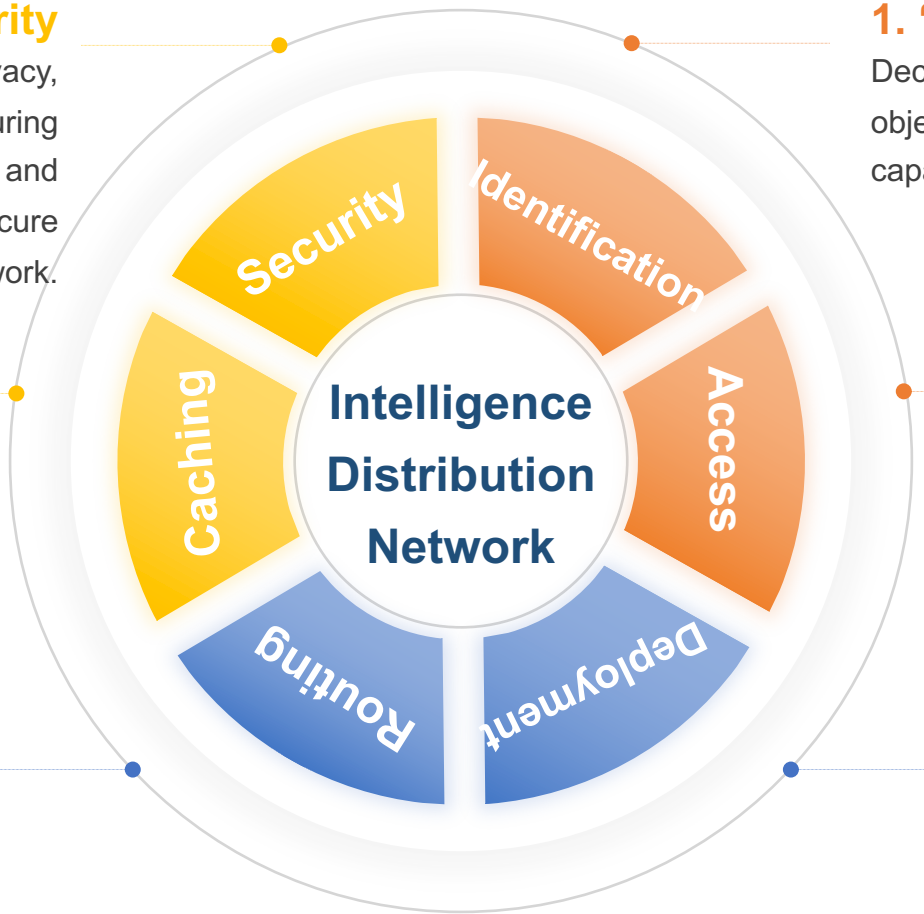
Roadmap

6. “Intelligence” Security

Ensure model security, data privacy, and inference integrity during intelligent service distribution and collaboration, forming a secure intelligent service network.

1. “Intelligence” Identification

Decompose large models into structured objects, characterized by service type, capability boundaries, and required resources.



5. “Intelligence” Caching

Cache intermediate inference results (e.g., KV Cache) within computing networks to reduce redundant computation and improve long-context inference efficiency.

2. “Intelligence” Access

Build a hierarchical computing network access mechanism allowing heterogeneous devices to join by declaring their capabilities, forming a computing resource pool.

4. “Intelligence” Routing

Schedule tasks and select routes among multiple computing nodes to handle massive heterogeneous user requests while maintaining high service quality.

3. “Intelligence” Deployment

Explore demand-driven collaborative deployment mechanisms within computing networks, supporting dynamic reallocation of intelligent services.

Follow us!

Submitted IETF draft (ongoing):

- **Title:** A Framework of Intelligence Delivery Network (IDN) for Deep Learning Inference (draft-li-cats-idn-00)
- **Link:** https://datatracker.ietf.org/doc/draft-li-cats-idn/
- **Author:** Qing Li, Hanling Wang, Yong Jiang, Mingwei Xu

Contact us:

- **Email:** hanling.thu@gmail.com
- **Wechat:** YanyeKong
- **Website:** https://smartinternet.group/

Smart Internet Group

Who we are?

Smart Internet Group (SIG) from Tsinghua Shenzhen International Graduate School, Peng Cheng Laboratory and Southern University of Science and Technology. Our research interests include: General in-network intelligence, Intelligent video transmission/analysis, DRL-based intelligent network routing, etc.

Thank you!

An Open, Decentralized, and Scalable Framework for Large Language Model Inference

(draft-wang-cats-odsi-00)

Author: Hanling Wang[1], Qing Li[1], Yong Jiang[2,1], Mingwei Xu[3]

[1]Pengcheng Laboratory, [2]Tsinghua SIGS, [3]Tsinghua DCST

IETF 125 @ Shenzhen, March 2026

Background: LLM Inference Era

- **LLM inference has become the dominant workload**
  - Conversational AI (multimodal understanding)
  - Image/video/voice generation
  - Recommendation, decision support, etc
- **Inference workloads exhibit new characteristics**
  - Massive **concurrency** and wide geographic distribution
  - Strong **latency** sensitivity for interactive applications
  - Increasing **privacy**, **security**, and **regulatory** constraints
  - Highly **heterogeneous** in terms of complexity, accuracy/latency requirements

LLM Applications

Model	#Param
DeepSeek V3.1	671B
Llama 3.1	405B
Claude Sonnet 3.5	175B

New requirements for compute and network infrastructure

Background: Why Centralized Inference is not Enough?

Centralized inference works well in data center

- High-bandwidth, low-latency fabrics (NVLink, InfiniBand)
- Homogeneous GPUs and tightly controlled scheduling
- Trusted execution environment
- Optimal solution for low latency and high throughput

Accuracy Alone Is Not the Whole Problem

- **Elasticity and Peak Cost**
  - Inference demand is bursty and highly correlated.
  - Peak provisioning dominates cost, not average utilization
- **Compute Ownership and Access**
  - Only a few providers can afford large-scale inference clusters
  - Inference access, pricing, and policy are centrally dictated
- **Cross-Organization and Trust-Constrained Inference**
  - In many settings (e.g., healthcare), data cannot be centralized and models cannot be fully replicated.
- **Geography and Tail Latency**
  - Data centers are geographically sparse
  - Users far from regions experience high tail latency

Centralized inference is not a one-size-fits-all solution!

Motivation: What is Expected for LLM Inference?

- **More Flexible Cost Structure**
  - Use already-existing, geographically distributed compute
  - On-demand scaling without centralized overprovisioning
  - A cost structure closer to pay-as-you-go at system scale
- **Active Computational Participants**
  - Idle GPUs, edge servers, and institutional clusters
  - Make inference participatory and composable
  - Inference service is economically shareable
- **Decentralized Compute Access**
  - No central control of computing resource
  - Partial model execution across domains/entities
  - No single trusted coordinator
- **Higher Execution Elasticity**
  - Inference execution can be placed closer to users and data
  - Reduces worst-case latency or tail latency
  - Provide higher throughput and enable larger model

Distributed LLM Inference is the Future Paradigm!

Distributed Inference: Challenges

Distributed Inference Is Strictly Harder

- Strict layer-wise dependencies
- Millisecond-scale per-layer deadlines
- Heterogeneous, unreliable, untrusted nodes
- No global scheduler or trusted coordinator

How can inference remain correct, timely, and live under these constraints?

➢ Challenge 1 -- The Activation Delivery Problem (ADP)

- Each inference request needs to traverse a chain of *L* dependent layers
- Each inference request needs to transmit *number_of_node – 1* activations
- Each activation must be timely, correct, and delivered to the next executor

☐ Why Existing Protocols Are Insufficient?

- **TCP / QUIC:** No notion of deadlines or dependencies
- **RPC frameworks:** Retries violate inference timing
- **Generic P2P overlays:** Ignore compute and latency constraints and collapse under churn

Distributed Inference: Challenges

➢ Challenge 2 -- Security and Incentives

- Nodes are independently owned and rational
- Correct execution cannot be assumed
- Detection alone does not ensure honest behavior

☐ Learning from Bitcoin and Blockchain Systems

- **Cryptographic identities:** public key as identity, private key as authority
- **Verifiable actions:** Correctness can be validated independently
- **Costly misbehavior:** Invalid blocks waste computation and yield no reward
- **Economic incentives:** System progress does not rely on altruism

ModelFlow: Unified Design Philosophy

ODSI: A Distributed and Decentralized LLM Inference Sytem

A Layer-Aware Transport protocol

- LLM inference is sequential, latency-sensitive, and deadline-driven
- The transport protocol understands model layer boundaries, execution order, and per-layer deadlines

A Coordination Protocol for Heterogeneous Peers

- Peers differ in compute speed, memory capacity, network latency, and availability
- Coordination must be decentralized, adaptive, and predictive

An Economic Protocol for Rational Participants

- Nodes are not altruistic. Compute and bandwidth have real cost
- The economic protocol defines who is allowed to execute inference and how rewards are distributed

Follow us!

Submitted IETF draft (ongoing):

- **Title:** An Open, Decentralized, and Scalable Framework for Large Language Model Inference (draft-wang-cats-odsi-00)
- **Link:** https://datatracker.ietf.org/doc/draft-wang-cats-odsi/
- **Author:** Hanling Wang, Qing Li, Yong Jiang, Mingwei Xu

Contact us:

- **Email:** hanling.thu@gmail.com
- **Wechat:** YanyeKong
- **Website:** https://smartinternet.group/

Smart Internet Group

Who we are?

Smart Internet Group (SIG) from Tsinghua Shenzhen International Graduate School, Peng Cheng Laboratory and Southern University of Science and Technology. Our research interests include: General in-network intelligence, Intelligent video transmission/analysis, DRL-based intelligent network routing, etc.

Thank you!