Xiaoyun WANG, Xiaodong DUAN, Kehan YAO, Tao SUN, Peng LIU, Hongwei YANG, Zhiqiang LI, 2024. Computing-aware network (CAN): a systematic design of computing and network convergence.
Frontiers of Information Technology & Electronic Engineering, 25(5):633-644.
https://doi.org/10.1631/FITEE.2400098

Computing-aware network (CAN): a systematic design of computing and network convergence

Front Inform Technol Electron Eng

Key words: Coordination of networking and computing; Computility network; Network architecture

Xiaoyun WANG, Tao SUN (Corresponding author)
E-mail: wangxiaoyun@chinamobile.com; suntao@chinamobile.com
ORCID: https://orcid.org/0000-0002-3574-9746;
https://orcid.org/0009-0003-3491-8813

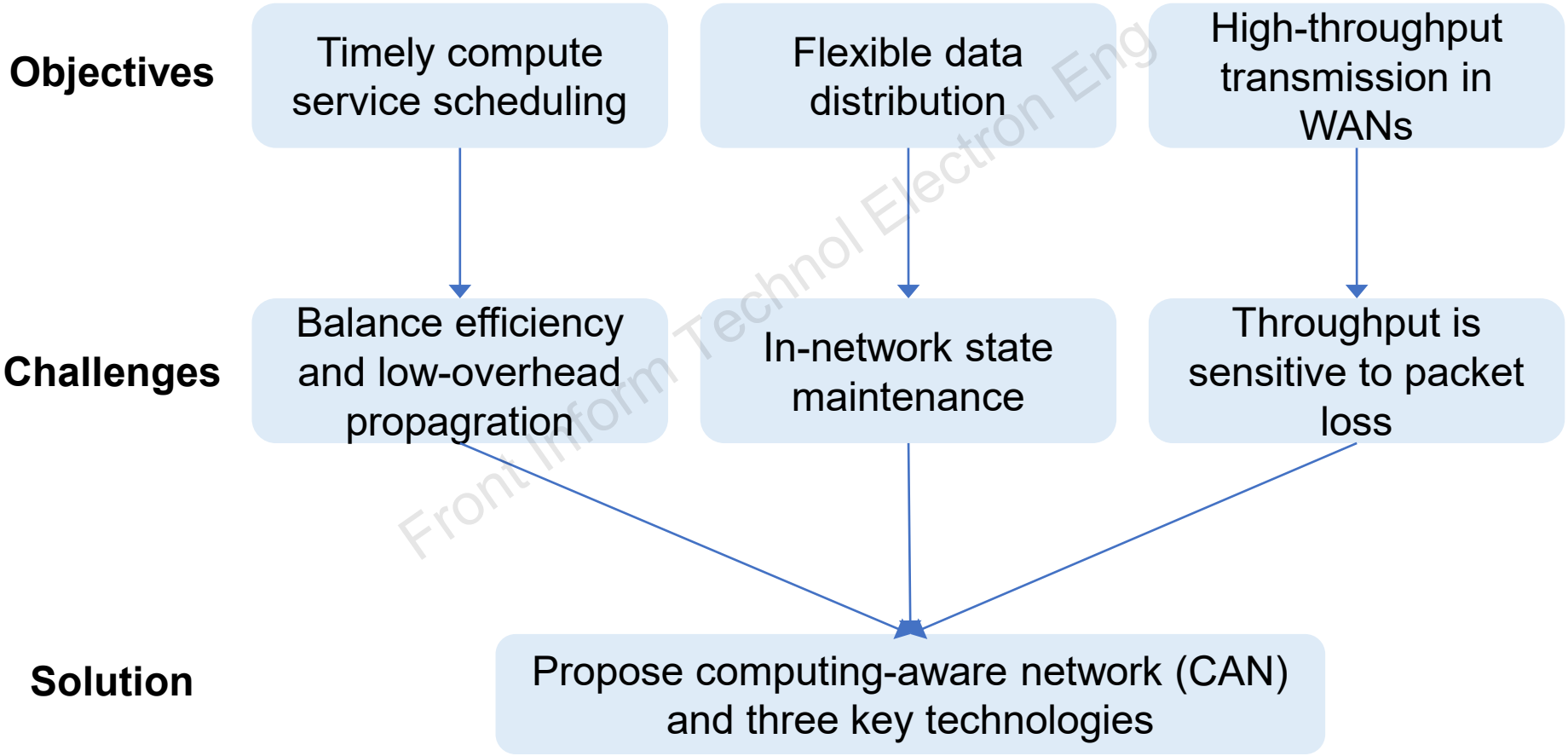
Background

- Diverse and contradictory optimization goals drive the convergence of computing and network, including critical service requirements, high resource utilization, and low energy consumption
- Data center level convergence has been relatively well investigated, while wide-area grade computing and network convergence still lacks further research
- Over-the-top solutions primarily focus on application orchestration over multi-clouds
- There is a need to study the convergence of the underlying network and computing resources over wide area networks (WANs)

Front Inform Technol Electron Eng

Motivation

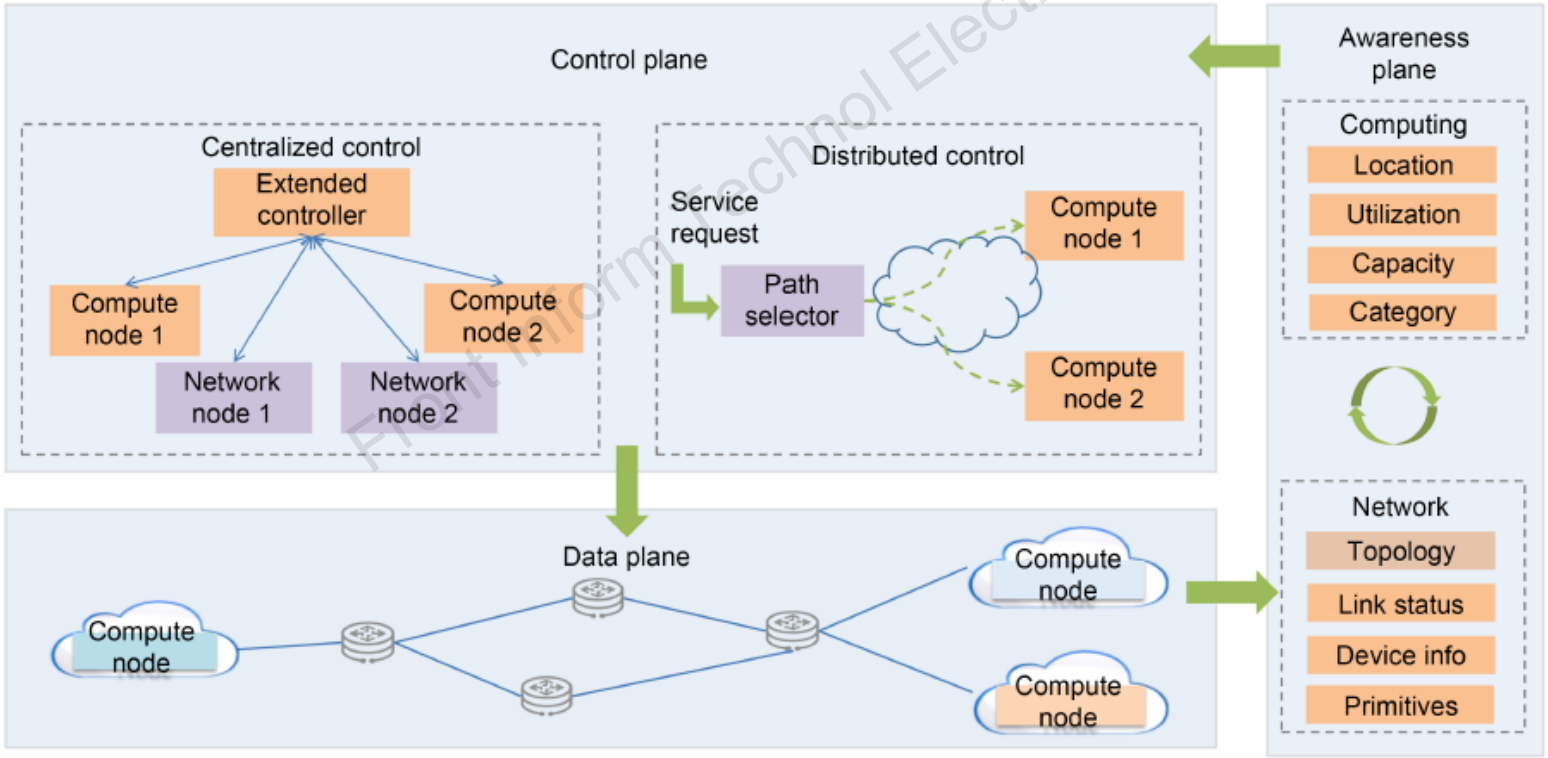
- Computing and network convergence in WANs still meets problems:



Front Inform Technol Electron Eng

CAN architecture

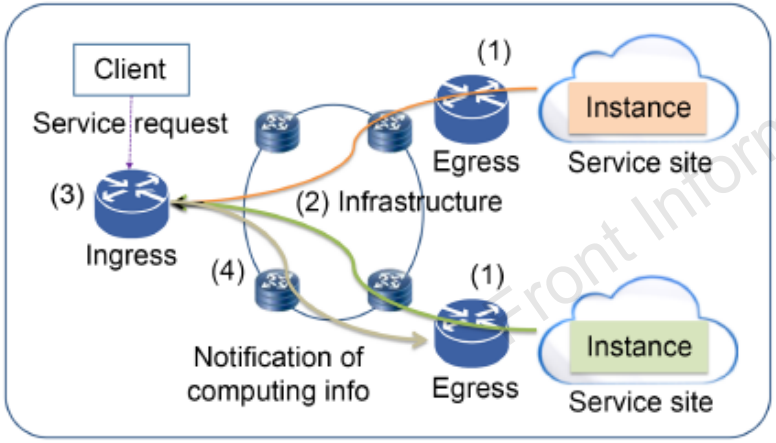
- CAN is defined as the integrated interconnection, joint awareness, and hybrid control of computing and network resources.
- It introduces the awareness plane to collect, manage, and synthesize computing as well as network information, which facilitates wide-area coordination of networking and computing



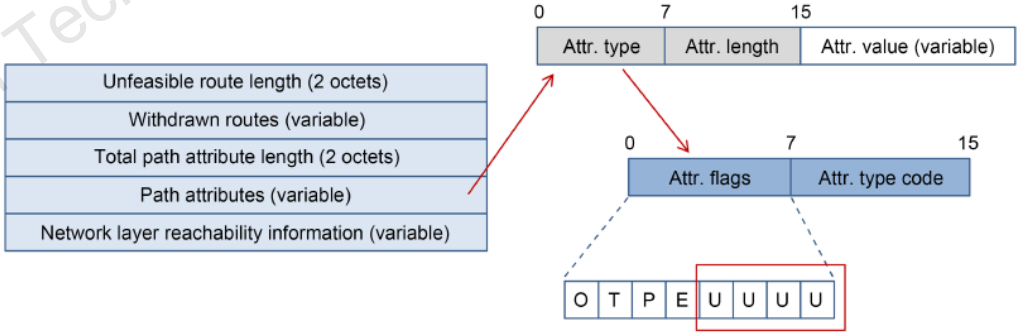
Computing-aware network (CAN) architecture

Computing-aware traffic steering (CATS)

- Service scheduling across multiple edge nodes usually faces the problem that the closest edge node may not be the best
- CATS achieves the optimal path selection through comprehensively considering computing and network information
- CATS based on distributed control adopts the "on-path" decision-making mechanism to meet the low latency requirements



CATS based on distributed control



BGP extension for metric distribution

Front Inform Technol Electron Eng

Elastic broadcast

- Elastic broadcast is designed for flexible wide-area one-to-many data delivery
- It allows in-network nodes to stitch, tailor, and duplicate data, which could save bandwidth occupancy and reduce data copies at endpoints
- Bit index explicit replication (BIER) extensions are adopted to avoid much state maintenance overhead in the network

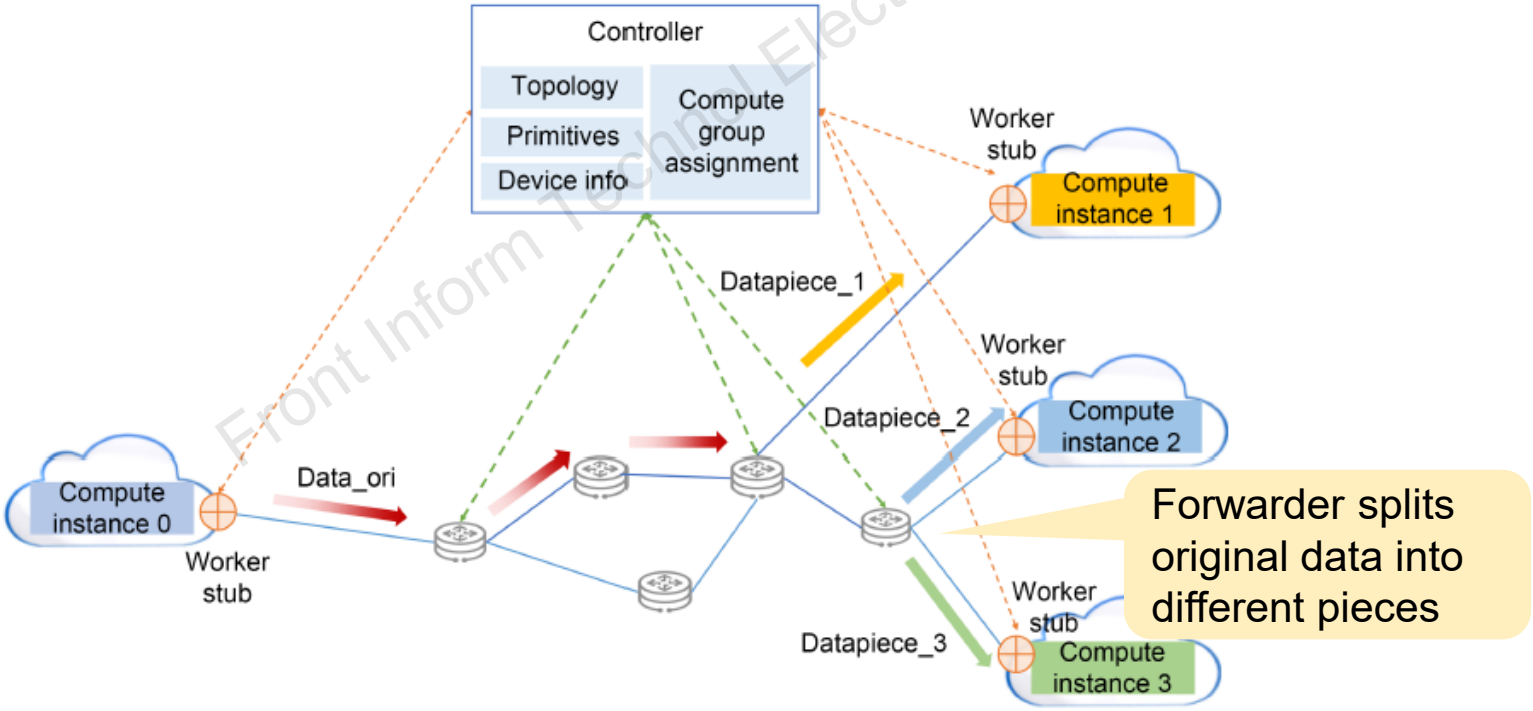
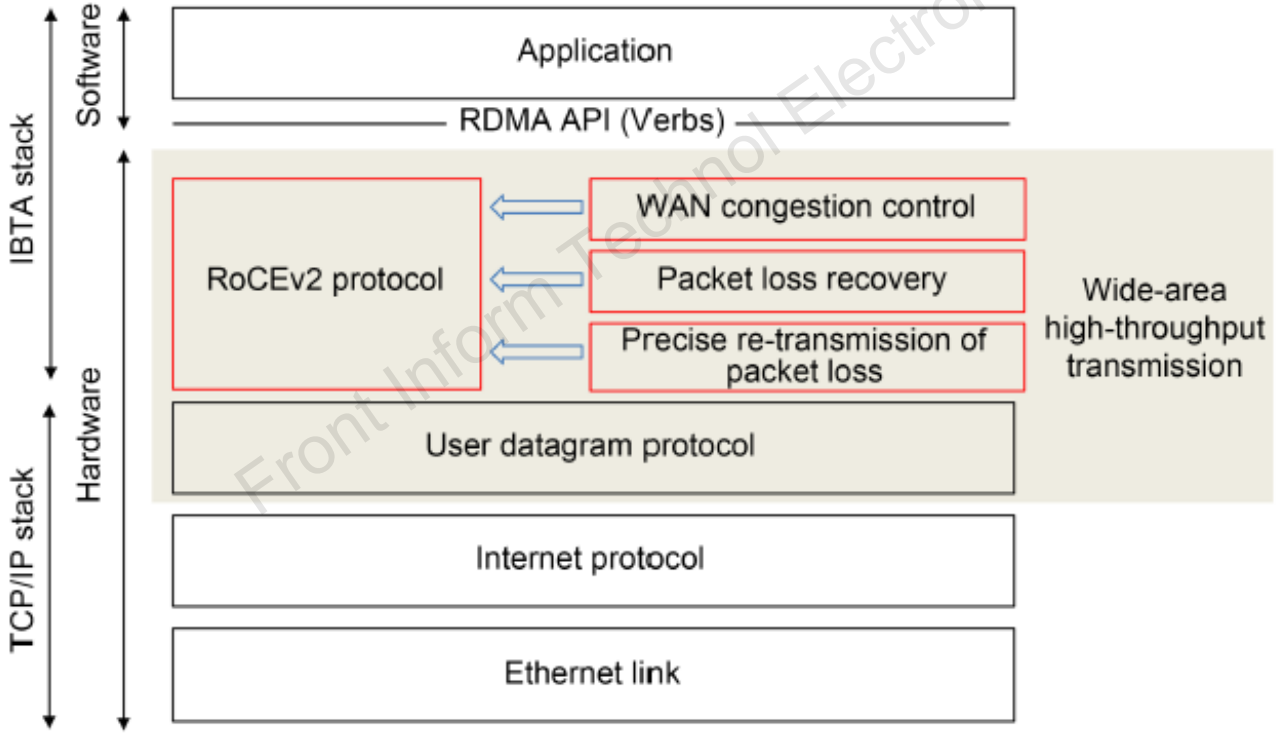


Illustration of elastic broadcast

Wide-area high-throughput transmission

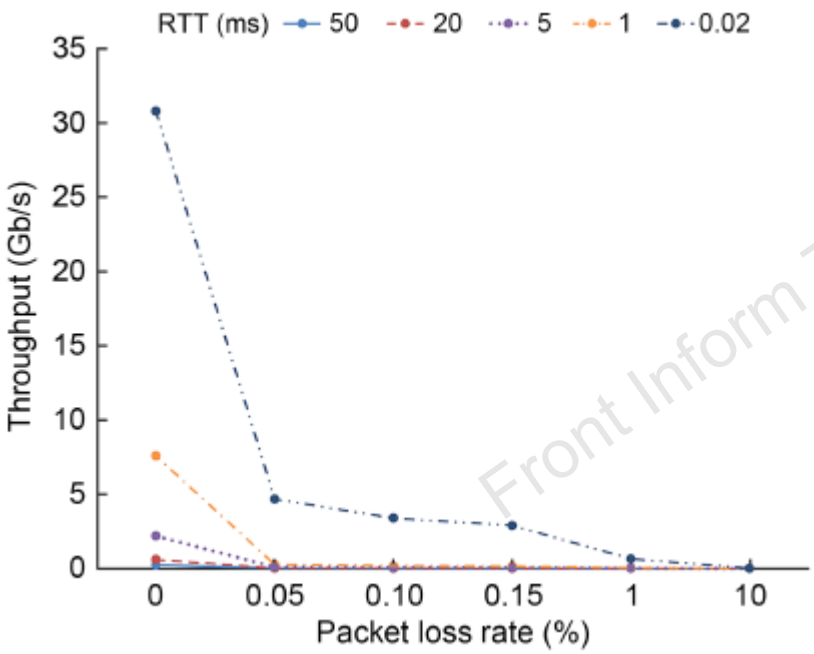
- Challenges of wide-area high-throughput transmission come from long-distance delays, packet loss, and server limitations
- Three key designs are fast packet loss recovery, precise re-transmission of packet loss, and optimized congestion control based on one-way delay



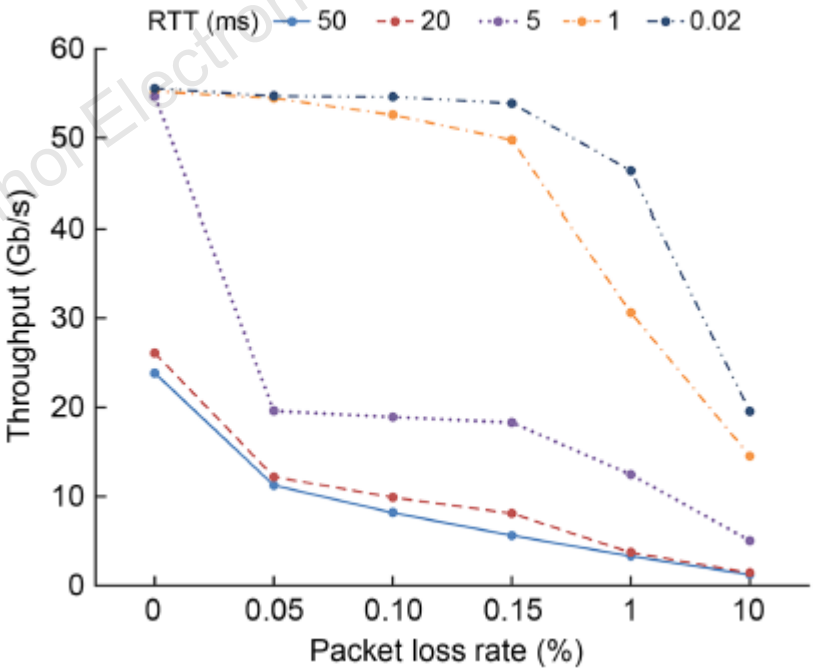
Wide-area high-throughput transmission protocol stack

Wide-area high-throughput transmission (Cont'd)

- Wide-area high-throughput transmission simulations
- Compare throughput between TCP-based solution and the proposed solution, under multiple test portfolios, by changing the packet loss rate and RTT



TCP-based solution performance

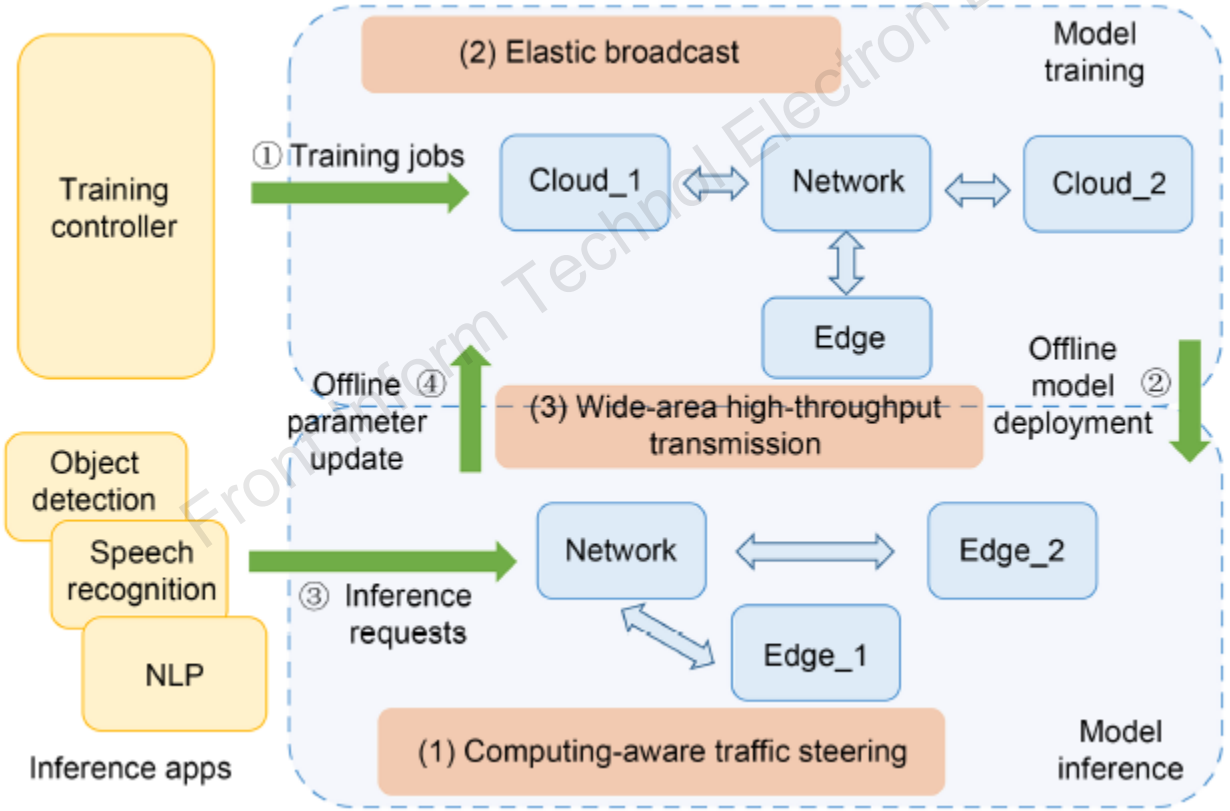


Wide-area high-throughput transmission performance

Front Inform Technol Electron Eng

CAN use cases

- Deploying AI applications over WANs will be a trend, due to inflated energy consumption and limited compute capabilities in a single data center
- CAN key designs fit well to optimize each phase of AI lifecycle, including model training, model deployment, inference, and parameter update



Centralized model training and distributed model inference based on CAN

Conclusions and future outlook

- CAN is a novel network architecture to facilitate computing and network convergence, with its core innovation in the introduction of the awareness plane
- CAN contains three key technologies, including computing-aware traffic steering (CATS), elastic broadcast, and wide-area high-throughput transmission
- Some promising research directions:
  1. Energy efficiency of CNC: explore scheduling strategies of computing and network resources from a global perspective to achieve the global optimal energy efficiency
  2. Convergence of computing, network, and applications: generate more accurate decisions through the integration of network, computing, and service information

Front Inform Technol Electron Eng

Xiaoyun WANG is a chief scientist at China Mobile. She was a recipient of multiple National Science and Technology Progress Awards, the National Innovation and Excellence Award, and the Chinese Youth Science and Technology Award. Her research interests include technology strategy, system architecture, and networking technology.

Xiaodong DUAN is a Vice President of China Mobile Research Institute (CMRI) and leader of network technology group of IMT-2030 (6G). He was a recipient of multiple National Science and Technology Progress Awards. His main research direction covers computility network, new IP technology, 5G/6G architecture, etc.

Tao SUN received his BS degree in automation in 2003 and PhD degree in control science and engineering in 2008, both from Tsinghua University, China. He is a chief expert of China Mobile. He has more than 10 years of experience on mobile network architecture design and IP technology research and standardization, and was a vice chair of 3GPP SA2 (system architecture). His research interests include 6G architecture, IP network evolution, and coordination of computing and network.

Front Inform Technol Electron Eng