
Hung-Yu Wei

National Taiwan University

Part I: Overview of Fog Computing and Networking
Why Cloud and Fog?

• Cloud
  • Centralized paradigm
  • Pooling
    • Efficiency of resource utilization
    • Scalability

• Fog
  • Closer to the edge
  • Lower latency
Fog Networking

“...a network architecture that uses one or a collaborative multitude of **end-user clients** or **near-user edge devices** to carry out a substantial amount of **storage** (rather than stored primarily in cloud data centers), **communication** (rather than routed over backbone networks), and **control, configuration, measurement, and management** (rather than controlled primarily by network gateways such as those in LTE core) “

Mung Chiang, Princeton University

---

Two Paradigms

- **Computing**
  - Cloud Computing
  - Fog Computing
- **Networking**
  - Cloud Networking
  - Fog Networking
- **RAN (Radio Access Network)**
  - Cloud RAN
  - Fog RAN
Part II: Serving Low-Latency Applications with Fog RAN

Acknowledgement to Prof. Ai-Chun Pang

“Fog Computing” and “Fog Network”

• Fog Computing
  • Cloud computing is extended to the edge of the network.
    • To create a highly virtualized platform that provides compute, storage, and networking services between end-devices and traditional cloud data centers

• Fog Network
  • Use one or a collaborative multitude of end-user clients or near-user edge devices to carry out
    • storage, communications, control, configuration, measurement and management

• F-RAN (Fog Radio Access Network)
  • NGMN Taiwan Team proposed this idea to the Advisory Session in the NGMN Forum in Bonn, Germany on June 3-4, 2014.
Birth of “F-RAN”

• Follow the same direction with “computing is extended to the edge of the network”

• Act as an aggregator connecting IoT end-devices and cellular network

• Not only responsible for communication (protocol/signaling), but also application services (data processing)

• In 5G era, a new business model is rising.
  • Telecom. operator cooperates with application/service provider for better quality of IoT services.

F-RAN

• A customized design for IoT applications for
  • Heterogeneous communication
  • Real-time computation
  • Local storage
  • Application specific functionalities

• Benefits
  • Reduce latency
  • Increase throughput
  • Leverage locality
  • Relieve backhaul load
IoT Generation

• Internet of Things (IoT) will grow fast from 2013 to 2018.
• Current cellular Infrastructure is hard to deal with incredibly increasing IoT traffics.

Categorization of IoT Applications

Massive (large scale)  Critical (low latency)

- Ultra-long range
- Low protocol overhead
- Scalable Access
- Capillary Networks & short-range radio
- Ultra reliable
- Very low latency
- Very high availability

Sensors, actuators
Capillary networks
Future Low-latency IoT Applications

• Tactile Internet
  • “The latency of communication systems becomes low enough to enable a round-trip delay from terminals through the network back to terminals of approximately 1ms”

• Applications
  • Health care, education and sport, traffic, robotics manufacturing, augmented reality.


Ultra Low Latency Applications in Wireless 5G

• 5G vision for ultra low latency
  • “To support such latency-critical applications, 5G should allow for an application end-to-end latency of 1ms or less.” (Ericsson 5G White Paper, 2015)
  • “Zero latency gigabit experience” (Nokia 5G White Paper, 2014)
  • “Provide consistent experience under diverse scenarios with ultra high data rate, ultra low latency and massive connections” (ITU-2020 5G Vision, 2014)

• F-RAN in wireless 5G can achieve low latency via:
  • Do computing for latency-critical IoT applications at the edge of the network.
Augmented Reality (AR)

• Definition:
  • “A technology which allows virtual graphics imagery to exactly overlay physical object in real time”

• Mobile AR with Wireless 5G
  • Widespread availability, portability, and built-in good quality cameras of smart phones
  • Exciting AR applications (e.g., navigation, public security, real-time translation)
  • 10ms end-to-end latency is required.

Examples of AR Applications


- Microsoft HoloLens, 2015
- Google Translate, 2015
- IBM Research, 2013
Flowchart for a Typical AR Service

Tracking module calculates the relative pose of the camera in real time for the correct location and orientation of the virtual components.

“Typical CPU hardware on mobile devices is not able to handle the computational/memory demands of tracking algorithm”

- The tracking needs to be done outside mobile devices with low latency.
- F-RAN can be a good solution.

AR Service with F-RAN

- The computing capability of an F-RAN node (FN) is approximate to that of a small/femto AP.
  - Single FN is not enough for the computing-intensive tracking algorithm.
  - Multiple FNs need to do the application-layer computing collaboratively.

Distributed Computing

1. Receive Captured Images
2. **Assign** Tracking Tasks
3. Do Pose Tracking
4. **Summarize** Results
5. Send back for Rendering

Cost-Performance Tradeoff

**Accuracy** (Performance)
- The result of pose tracking
- Low accuracy leads to **virtual-object misplacement.**

**Computing** (Cost)
- Computing resource of each FN is limited.

**Communication** (Cost)
- Tracking data sent via wireless links among FNs
- Distance among FNs affects transmission rate (i.e., latency)
Factors Affecting the Tradeoff

- **Number of cFN**
  - More cFNs
  - ▶ Higher *comm. cost* for mFN but lower *computing cost* for each cFN

- **Amount of Assigned Tasks for each cFN**
  - More tasks for each cFN
  - ▶ Higher *accuracy* with higher *computing cost*

- **Attributes of cFNs**
  - Loading of cFNs ▶ *Computing cost*
  - Distances between mFN and cFNs ▶ *Comm. cost*

- **Amount of Information Sharing**
  - Information (pose tracking data for AR service) shared from mFN to each cFN
  - Especially complex for AR service

Technical Challenges

- **How to decide which FNs to be the cFNs**
  - Number of cFN
  - Attributes of cFNs

- **How to split the task for distributed computing**
  - Amount of Assigned Tasks for each cFN
  - Amount of Information Sharing
Summary

• In wireless 5G, many emerging IoT services and applications (e.g., augmented reality) require **ultra-low latency**.

• We have proposed F-RAN and tried to push application-layer computing from cloud data centers to mobile edge for ultra-low latency IoT applications.
  • Properly handling **performance-cost tradeoff** is the key.

• However, there will be diverse application requirements in wireless 5G.
  • Low-latency access co-exists with non-real-time access
  • Flexible operation + networking slicing

Part III: Radio Access Network Design for Low Latency 5G
Technical Opportunities in 5G

- Growth Opportunities
  - IoT, M2M, Internet of Everything ...
  - LTE have been adding **MTC** (Machine Type Communications) capability

- MTC in 5G
  - Native support for MTC

### Design 5G to provide native support in MTC

5G Machine Type Communications for Internet of Everything

- Cellular wireless infrastructure for MTC
  - Long transmission range
  - Low cost equipment
  - Improve quality of life

- Trends in Wide-Area IoT
  - Proprietary LPWA (Low-Power, Wide-Area) Networks
    - LoRA, Sigfox, ...
  - 3GPP
    - Cellular IoT
    - Narrow-band LTE → Narrow-band IoT

### 5G MTC could re-shape telecom economy
Everyone Talks about MTC/IoT

- A diverse sets of applications and requirements

- Vertical markets
  - Growth opportunities
  - $

- Two major types
  - **Mission critical MTC**
    - Low latency, high reliability
  - Massive MTC
    - Low cost, power efficient

5G Usage Scenarios of IMT-2020

5G RAN Design for Low-Latency Services

- Mission-critical and low latency services

Design Principle #1
Embrace Fog Paradigm for Low-Latency

ITU-R M.2370-0 "IMT traffic estimates for the years 2020 to 2030", July 2015
Two Paradigms

• **Cloud** computing and networking
  • Centralized pooling
  • Efficient resource utilization

• **Fog** computing and networking
  • Closer to the edge
  • Lower latency

Push Everything to Edge for Low Latency

Challenge: end-to-end latency

• Everything needs to be handled with low latency
  • Air interface
  • Baseband processing
  • Higher protocol layers
  • Computing

Handle Each Delay Component Carefully
Design for Low Latency

- End-to-end latency is composed of several segments
  - Latency in RAN network entry depends on traffic load
  - Cross-layer protocol design is needed to reduce the overall latency across all protocol layers
  - Fog computing and mobile edge computing reduce the latency in computing

ETS1 GS MEC-IEG 004, "Mobile-Edge Computing (MEC); Service Scenarios," November 2015

Challenge: Mixed of Traffic

- Diverse application requirements
  - Low-latency access co-exists with non-real-time access
  - Differentiate different types of accesses
  - Flexible operation
Design Principle #2: Differentiate Low-Latency Flow Early

Diverse IoT Traffic

• Low latency in every steps to achieve end-to-end low latency
  • RACH procedures
  • Initial access
  • Computing

• Low latency service is costly

Differentiate Low Latency IoT Flows Early
Random Access Procedure in LTE

- In LTE systems, a device (UE) connect to the network by performing random access procedure
  - Termed RACH procedure in LTE
- The first step: RACH preamble transmission
  - To inform the eNB of UE’s connection request
  - A UE randomly selects one preamble sequence from a sequence pool
  - Preambles with different sequences can be successfully received simultaneously

RACH collision

- RACH collision
  - Two or more UEs selects the same sequence for preamble transmission
  - The base station might be unable to decode the transmitted preamble
  - It causes preamble transmission failure
- RACH contention resolution
  - Preambles from numerous IoT devices may cause severe RACH collision, and further affects normal H2H communications
  - Some mechanisms should be applied to resolve RACH contention
M2M/H2H Co-existence

Flexible Configuration in Fog-RAN

- **Latency requirements**
  - Ultra-low latency
  - Low latency
  - Delay tolerant

- **Two types of resources**
  - Communications
    - Fog processing
    - Cloud processing
  - Computing
    - Fog computing
    - Cloud computing
Core Network

Network Packet Classifier

Cloud BBU Pool

Fog BBU

Radio Access Classifier

Cloud Computing

Fog Computing

Ultra-low Latency

Low Latency
Design Principle #3
Be Cognitive, Be Flexible
Flexible RAN with Cognitive Operation

• Cognitive design principles
  • Observe
  • Learn
  • Decide
  • Act
• Use context to support dynamic and diverse traffic

• Challenge to support massive M2M
  • Bursty load to control plane signaling
  • Network entry

Understand the Context

• Traffic load as context information
  • Estimation of network entry load
  • The number of devices under contention

• Obtain context information with low cost
  • RACH: success, collision, empty
    • Estimation technique

• Adaptation for load control
  • Also helpful for latency reduction
  • Collision and retransmission are harmful for latency

Adaptive M2M Access: Network-Assisted Context Info

Broadcast System Context Information

Adaptive device configuration accordingly

Adaptive M2M Access: Distributed Decision at M2M Devices

Utility (successful transmission) - Cost

(utility is application-dependent)
Context-Aware H2H/M2M Adaptation


Flexible Fog RAN for Diverse Traffic

- Low latency service is costly
  - Great challenges and opportunities

- Not all services have low latency requirements
  - Flexibility in business and pricing models

- Joint design for computing and communications
  - Small cell v.s. C-RAN
  - Mobile edge computing and Fog-RAN

- RAN design issues
  - Integrated RAN architecture
  - Differentiate network access
  - Context-aware adaptation
Summary: Flexible RAN for Low-Latency and Diverse Applications

• Hybrid architecture for C-RAN and Fog RAN integration
  • Heterogeneous evolution path for C-RAN
  • Push computing and communications processing to the edge
  • Latency

• Flexible Fog RAN architecture to support diverse applications
  • Context-aware adaption

Part IV: Standard Activities
ETSI Mobile Edge Computing

Source: ETSI

IT Service at the Edge of Network

Source: ETSI
Use Case: Active Device Location Tracking

Source: ETSI

Use Case: Augmented Reality Content Delivery

Source: ETSI
Use Case: Video Analytics

Source: ETSI

Use Case: RAN-aware Content Optimization

Source: ETSI
Use Case: Distributed Content and DNS Caching

Source: ETSI

Use Case: Application-aware Performance Optimization

Source: ETSI
Use Case: Connected Vehicles

Source: ETSI

MEC Deployment Scenarios

Source: ETSI
MEC API

Mobile-edge Computing platform API

Source: ETSI

NGMN’s View on Edge Computing

Edge Server with cloud infrastructure: Cloudlet or MEC server: compute, store, network

Source: NGMN
5GMF Architecture Vision

- Ultra-Low latency “sub-msec.” demand for Delay-critical interactive services.
  Trading, Medical, IoT, Vehicle Telematics, Virtual Reality, etc.
- Mobile Edge Computing (MEC) improves QoE.
- Other Potential Approach
  - D2D networking
  - Small buffering
  - Location-/Service-awareness proximity services
  - Disaster relief emergency services from local edge servers

Source: 5GMF

Thank You

hywei@ntu.edu.tw