2020 IT 21
Global Conference
Digital New Deal
Technology Essentials
디지털 뉴딜 기술 핵심

Session 6-4

엣지 기반 분산 클라우드 기술 동향

윤주상 교수 (동의대학교)

[요약문]

본 강연에서는 엣지 기반 분산 클라우드 구현을 위해 개발 중인 지능형 분산-협업 엣지컴퓨팅 협업 모델 및 관련 기술 개발 현황을 소개한다. 특히, 클라우드-엣지 및 엣지 간 지능형 협업 모델의 개발 방향 및 요소 기술에 대해서 논의한다.

최근 지능형 분산-협업 엣지컴퓨팅 기술 개발 방향은 컴퓨팅 오프로딩 및 서비스 이동 제공 시 클라우드-엣지 및 엣지 간 협업 및 연계 운영이 가능한 딥러닝 기반 지능형 기술을 개발 중이다. 본 강연에서는 이와 관련된 다양한 기술 개발 사례를 중심으로 관련 기술을 소개 하며 주로 딥러닝 기술이 엣지 컴퓨팅 협업 모델에 어떻게 적용되는지를 소개한다. 또한, 클라우드-엣지 및 엣지 간 협업 모델에 대해서는 분산-협업 시 필요한 엣지컴퓨팅 협업 요소 기술을 정의하고 관련 기능을 소개한다.

[발표자 약력]

1999년 고려대 전기전자전파공 학사
2008년 고려대학교 전자컴퓨터공학 박사
2002년 ETRI 연구원
2012년~현재 TTA 국제표준전문가
2015년~현재 TTA 사물인터넷 네트워킹 그룹 부의장
관심분야 : 지능형 사물인터넷, 클라우드 컴퓨팅, 포그/엣지 컴퓨팅, 5G 등

엣지 기반 분산 클라우드 기술 동향

동의대학교
윤주상
joosang.youn@gmail.com
2020.09.25

내용

- 엣지 컴퓨팅 소개
- 엣지 컴퓨팅 도입 배경 및 필요성
- 엣지 기반 분산 클라우드 기술
- 지능형 분산 협업 기술 동향
- 결론

Computing 기술의 변화

Cloud
(2000 – 2015)

Edge / Fog
(2016 – ?)

배경

- 사물인터넷 서비스 증가로 인한 폭발적인 데이터 증가
  - 데이터 활용을 위한 컴퓨팅 서비스 요구 증가

- 인터넷 기반 응용 서비스 환경 변화
  - 컴퓨팅 기반 응용 서비스 증가
  - 지연에 민감한 이동 응용 서비스 증가
  - 자원 제약적 모바일 디바이스 (IoT 디바이스)의 컴퓨팅 요구 사항 증가
    - Computation offloading
    - Service migration
    - Network Control & Resource Allocation

  - → Reduces energy consumption for local processing and prolongs b attery life

엣지 컴퓨팅 도입의 필요성

- Cloud computing(CC)
  - Easy-to-use platform
  - 사물인터넷 디바이스는 데이터 저장 및 프로세징을 위한 CC에 의존적임
- 하지만, 최근 CC 사용에 있어 한계점이 발견됨
  - Real-time, low latency, mobile applications 서비스 제공에 있어 지연 발생
- 엣지 컴퓨팅 장점
  - 초지연 서비스 제공 가능
  - 네트워크 자원 사용률 감소

클라우드-엣지 컴퓨팅 협업 기술

- 솔루션 구분

구분	분류	특징	배포 방식	지원 Platform
MS, Azure IoT Edge	Computing	클라우드에서 ML 모델 학습 후 배포 머신 러닝, 스트림 데이터 질의 / 필터링 클라우드 기반 Edge 기기 관리	Microservice	Linux, Mac, Windows
AWS, Greengrass	Computing	AWS 기능(Lambda, Shadow)을 Edge에서 실행 클라우드에서 ML 모델 학습 후 배포	Proprietary (MQTT 통신)	Linux, RPi
Google IoT Edge, Apple 등	Computing	클라우드에서 ML 모델 학습 후 배포, 딥러닝 기반 AI 서비스 지향, H/W(NPU, GPU, 전용 AI Processor 등) 활용	SDK	Android iOS,macOS,tvOS
ARM, MBed Edge	Collecting	Edge와 연결된 기기의 Protocol Translation	Microservice	Linux, RPi

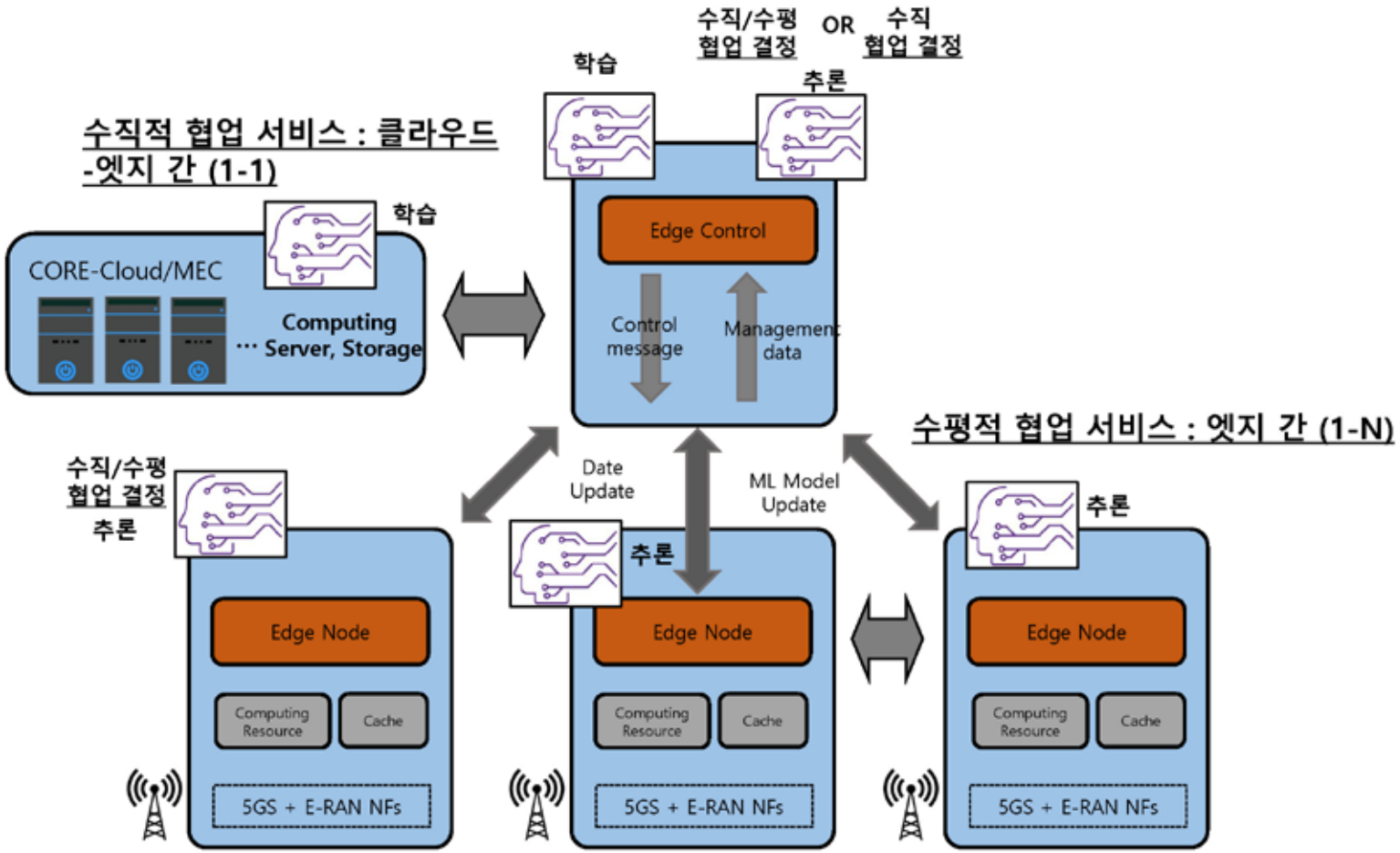
엣지 컴퓨팅 기술의 변화

- Edge computing **Power 증가**
- **Intelligent** Service **in Edge** 요구
- 5G 융합서비스 개발 촉진
  - 초저지연 서비스 증가
  - Local cache 서비스 증가
- **Finding insights in data**

→Where? **Cloud --> Edge**

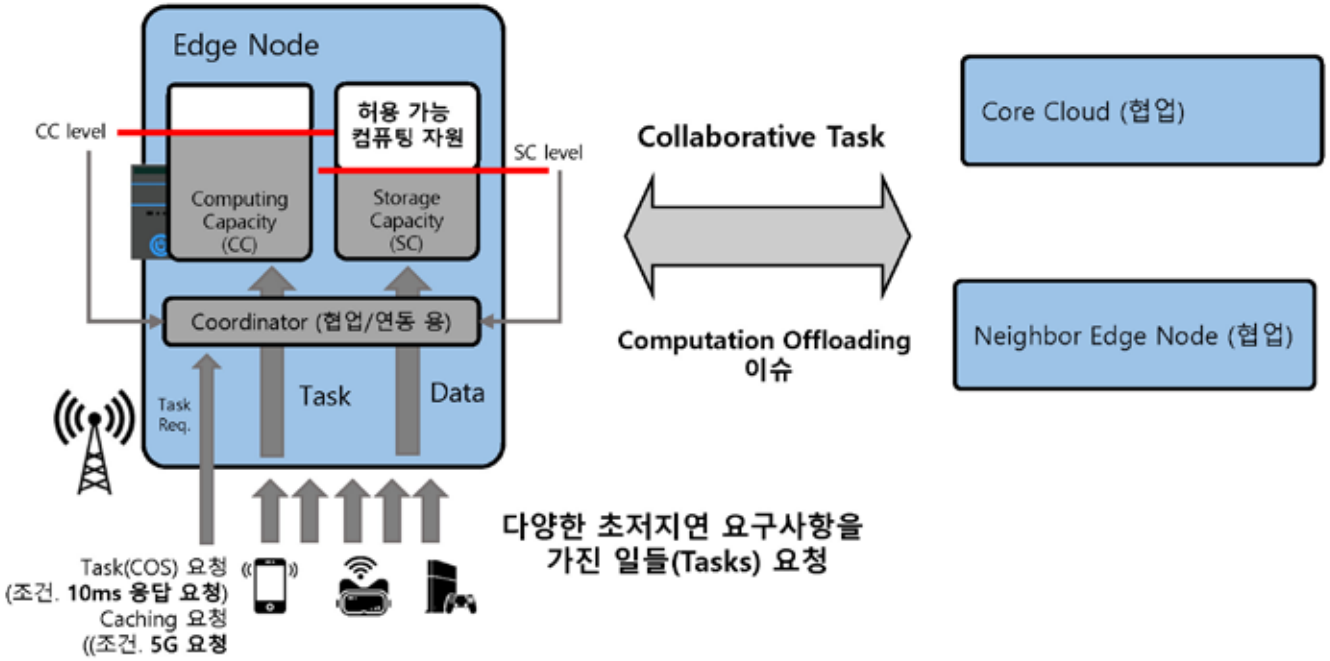
→ 현재 클라우드-엣지 간 협업 시나리오 및 기술은 존재하나 엣지 간 협업 기술 개발은 진행형
→ **엣지-클라우드 및 엣지 간 협업 기술 요구**
→ **이를 통해 분산 클라우드 인프라 구축**

분산-엣지 협업 서비스


수직/수평
협업 결정
OR
수직
협업 결정
학습
추론
수직적 협업 서비스 : 클라우드
-엣지 간 (1-1)
학습
CORE-Cloud/MEC
Computing
… Server, Storage
Edge Control
Control
message
Management
data
수평적 협업 서비스 : 엣지 간 (1-N)
수직/수평
협업 결정
추론
Date
Update
ML Model
Update
추론
추론
Edge Node
Computing
Resource
Cache
5GS + E-RAN NFs
Edge Node
Computing
Resource
Cache
5GS + E-RAN NFs
Edge Node
Computing
Resource
Cache
5GS + E-RAN NFs

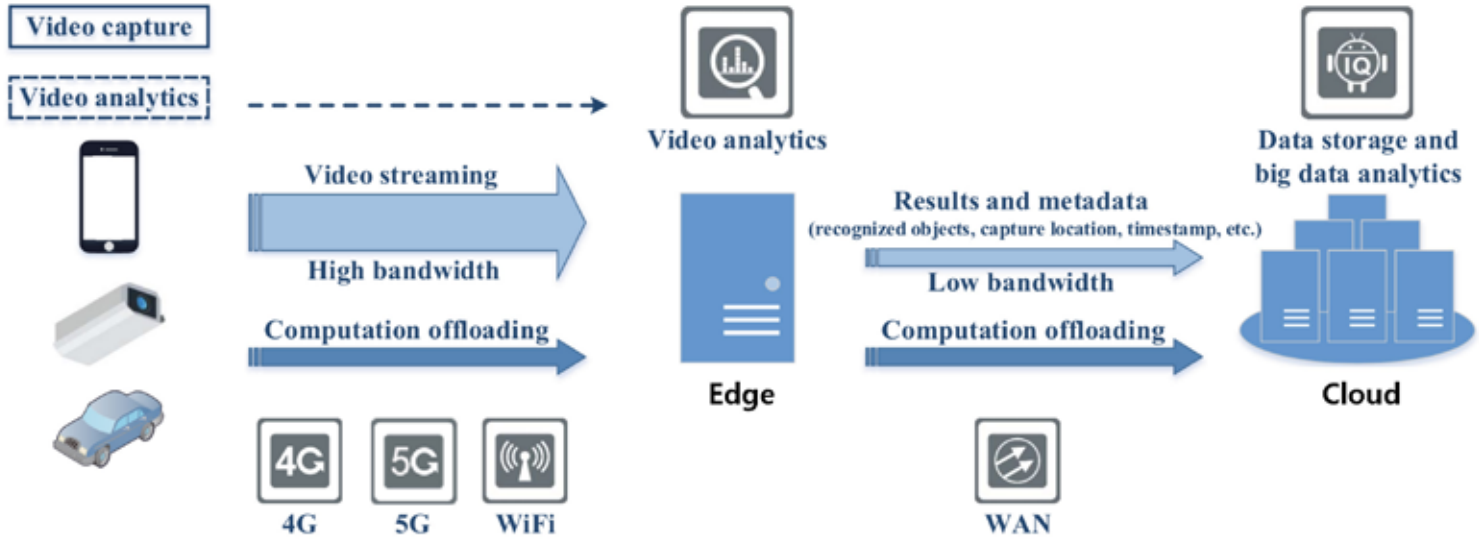
협업이 필요한 분산-엣지 서비스 시나리오

- COS, DCS, DSS 가 필요한 서비스 시나리오 정의



용합 서비스 관점에서 엣지-클라우드 간 협업 서비스 (예)

- Edge-cloud 연계 비디오 분석 운영 시나리오
  - Latency-sensitive Applications over Edge-Cloud Computing



비디오 분석을 위한 협업 운영 솔루션 (기존)

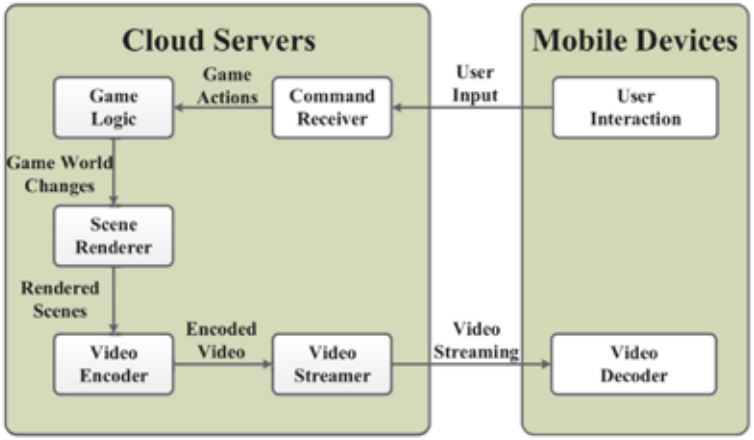
- 기존 솔루션

Research	Proposed solutions	Edge devices	Edge nodes	Offloading granularity	What's to be offloaded
VideoEdge [158]	A hierarchical framework for video analytics and large-scale video query	IoT cameras	Private clusters	Full	The pipeline of computer vision components, such as video decoder, object detector, and associator
GigaSight [73]	A 3-tier architecture for first-person video storage, sharing, and content searching with privacy removing	Head-up displays, like Google Glass	Cloudlets	Tasks-Components	Video privacy removing, video tagging and indexing, and content searching
Wang *et al.* [159]	A privacy-aware and scalable live video analytics for face recognition	Devices with cameras, e.g., mobile phones	Cloudlets	Tasks-Components	Video privacy removing, the execution of analytics algorithms, e.g., DNN-based feature extraction and classification
Wang *et al.* [160]	A bandwidth-efficient drone video analytics framework employing adaptive policies to reduce video transmission	Drones	Cloudlets	Tasks-Components	The execution of computer vision algorithms
LAVEA [161]	A latency-aware 3-tier video analytics architecture enabling offloading tasks inter-edge collaboration	Smartphones and wearable devices	Container-based cloudlets	Tasks-Components	The execution of computer vision algorithms

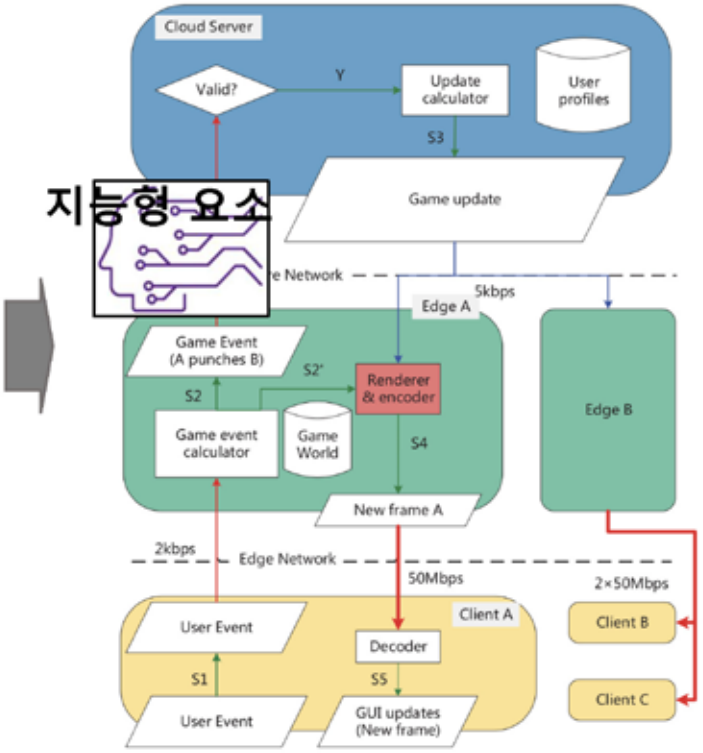
지능형 요소 개발 (딥러닝을 활용한 솔루션 도출)

서비스 연동 지능형 운영 방안

- 예)



Workflows of cloud gaming with
the types of video streaming



Hybrid Architecture,
to reduce latency and network traffic

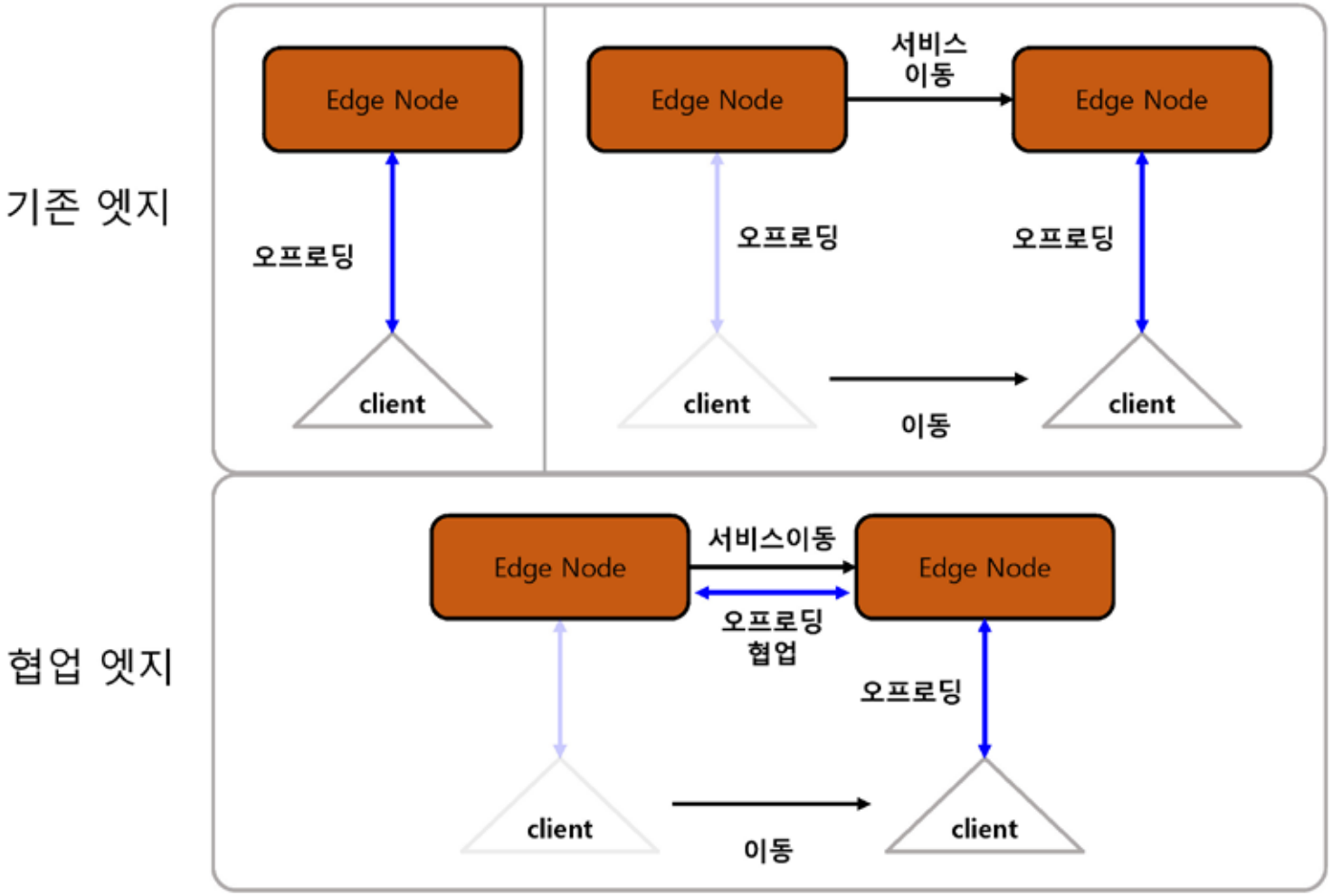
서비스 결정 필요 → 지능형 요소 설계

Distributed AI Service

- 인공지능 서비스에서 고민 중
  - 엣지 추론 – 클라우드 학습 (일반적인 모델)
  - 엣지 추론**+학습** – 클라우드 학습 (제안하는 모델)
    - Multi-task 상황에서 추론 서비스 지연 발생.. 등
    - 5G 융합서비스 중 고민 중

- 서비스 시나리오 발굴 및 정의 후 Edge-Edge(E-E), Edge-Cloud(E-C) 간 서비스 연동을 위한 지능형 운영 방안 제안

엣지 간 협업 모델


기존 엣지
Edge Node
오프로딩
client
Edge Node
서비스
이동
Edge Node
오프로딩
오프로딩
client
이동
client
협업 엣지
Edge Node
서비스이동
Edge Node
오프로딩
협업
오프로딩
client
이동
client

엣지 서버 간 협업 기술의 중요성

- 엣지 컴퓨팅 서버 기반 응용 개발 중
  - 드론 응용, cyber–physical–social systems (CPSSs)
  - Traffic violation tracking cameras
  - Drone services for delivery
- 클라우드-엣지 간 역할 분담이 이루어지고 있음
  - 지능형 서비스
    - 클라우드-학습, 엣지-추론, 학습
  - 실시간 서비스 및 결정 서비스 등 엣지 서버 역할로 다운 중
- 클라우드 단절 서비스 제공을 위한 엣지 역할
- 엣지 기술은 분산 방식으로 개발이 용이함

엣지 기술 관련 이슈

- 실시간 이동 응용 서비스 경우 여전히 지연 및 에너지 컴퓨팅 이슈가 존재함
  - 엣지 네트워크 내 자원의 동적 특성으로 실행 시간이 불안정함
  - 엣지 컴퓨팅 서버의 효률적 오프로딩 서비스 제공 방법 필요
- 오프로딩 서비스 제공시 오프로딩 수준 정의

- 협업 관련 이슈
  - 협업이 가능한 분산 엣지 네트워킹 기술 부족
    - 엣지 간 정보 전달을 위한 인터페이스 및 프로토콜 부재
  - 오프로딩 수준에 따른 협업 모델 필요
  - 협업을 위한 엣지 서버 탐색 방법 부재
  - 협업 목적에 따른 협업 방법 및 정책 부족
  - 이외에도 다양한 협업 이슈가 존재함

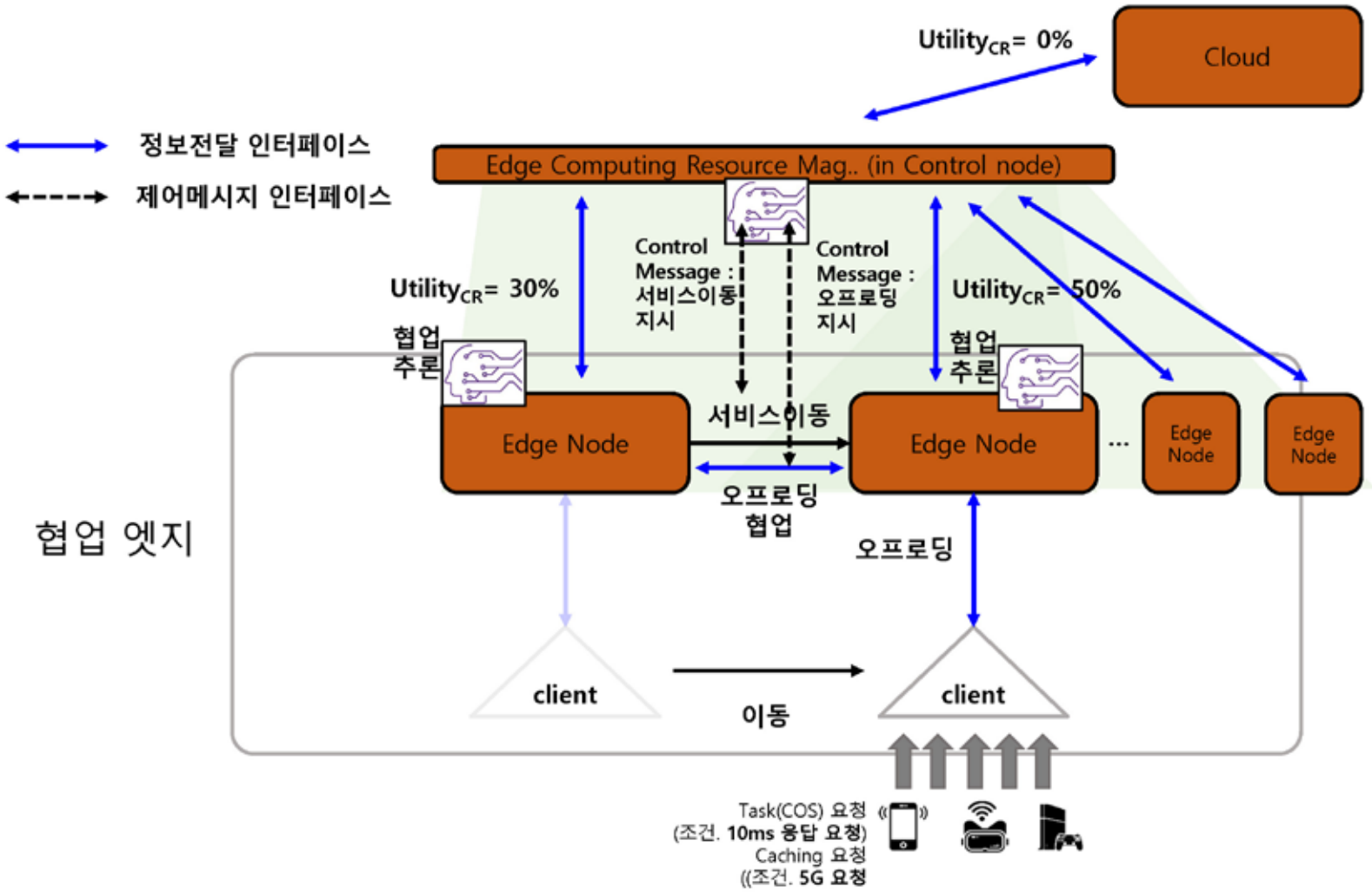
지능형 분산-엣지 컴퓨팅 서비스 연구

- 기존 연구
  - 인접한 엣지 서버와 협업이 없음
  - 자원 할당 및 관리를 위한 policy-based reinforcement learning techniques 기술 적용
    - 현재 상태가 반영되지 않는 모델 적용
    - 일반적인 오프로딩 시나리오 적용

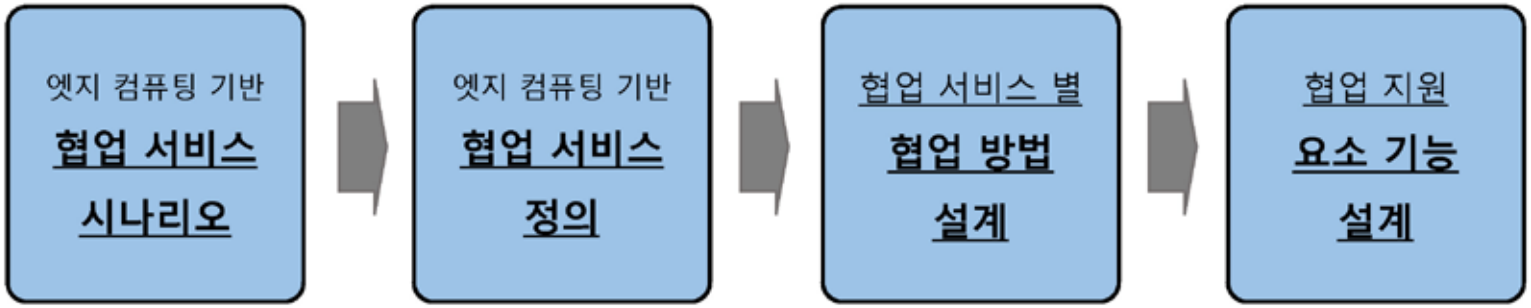
- 고려 사항
  - 분산 엣지 네트워크 환경 고려
  - 지연과 에너지 측면을 동시 고려
  - 현재 상태가 반영되는 심층 강화 학습 모델 적용
  - 협업 엣지 디스커버리 및 선택

→ 분산 엣지 네트워크 환경에서 DRL 기반 자원 할당 및 관리 기술

지능형 분산-엣지 협업 모델 (간단한)



협업 모델 개발 시나리오


엣지 컴퓨팅 기반
협업 서비스
시나리오
엣지 컴퓨팅 기반
협업 서비스
정의
협업 서비스 별
협업 방법
설계
협업 지원
요소 기능
설계

협업이 필요한 컴퓨팅 서비스 정의

- 클라우드-분산 엣지 간 컴퓨팅 서비스 (엣지 기반 융합 서비스 아님)
  - Computing Resource Provisioning
    - Computing Offloading Service (COS)
    - Decentralized/Distributed Cache/Storage Service (DCS, DSS)
    - Service Migration (SM)
    - Distributed AI service
      - Task offloading service와 유사
  - Network Resource Provisioning
    - 이동 디바이스의 서비스 품질 보장을 위해 엣지 서버간 네트워크 자원 보장 등을 협의 할 수 있는 협업 기술 필요

Computation Offloading & Granularity

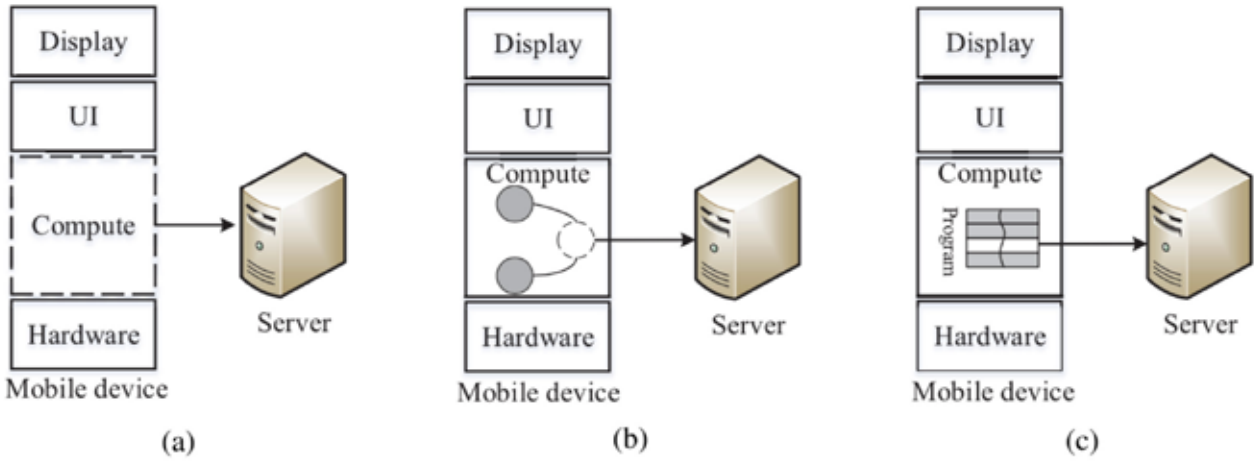


Fig. 4. ***Granularity of computation offloading. Generally, mobile devices can offload computation at (a) full offloading, (b) task/component, and (c) method/thread levels.***

- 협업 시 Granularity 설정이 필요

협업 서비스 수준 정의

- <컴퓨팅 오프로딩 제공 방법 별로 협업 서비스 수준 정의>

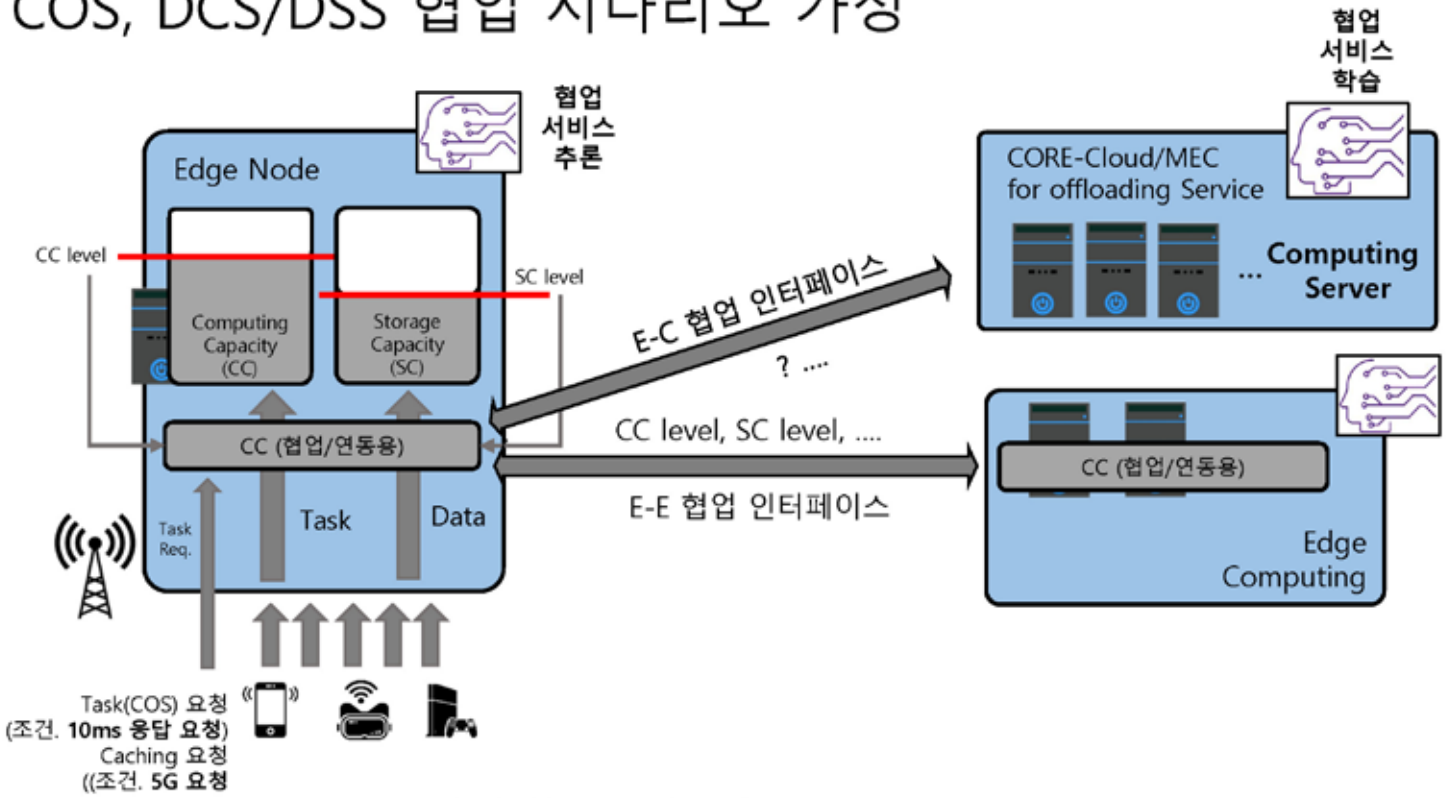
- Computation Offloading (CO-X)
  - **Application Partitioning**
    - **Partitioning** for multi-offloading in distributed edge node
      - Tasks can distribute edge-cloud or between edges
    - Partitioning granularity
  - **Task Allocation**
    - The runtime decision of task placement and scheduling associated with the resource management
      - if considering multiuser scenarios, multi-edge node collaborated resource allocations and load balancing are two challenges
  - **Task Execution**
    - VM-based Execution, Container-based Execution
    - Serverless architecture
    - Process-level isolation

Application Partitioning
Task Allocation
Task Execution

협업 요청 시 고려사항 (협업 수준)

협업 모델 도출

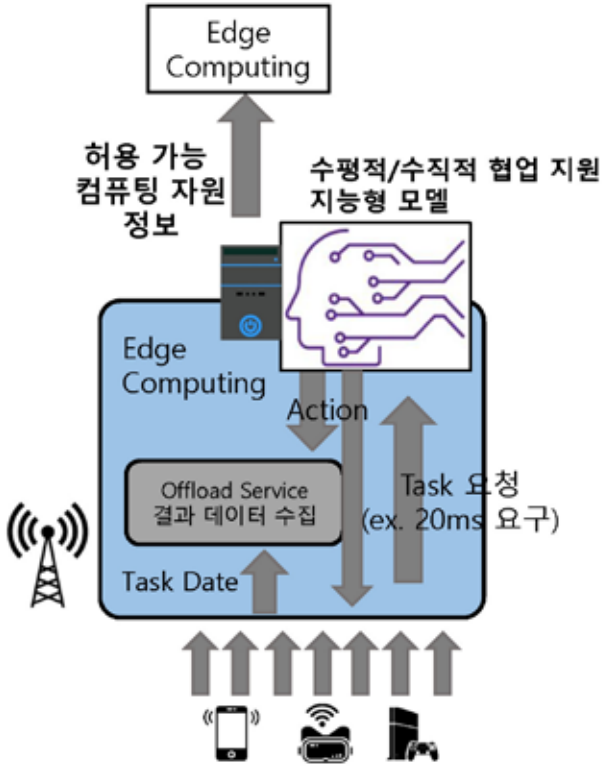
- COS, DCS/DSS 협업 시나리오 가정

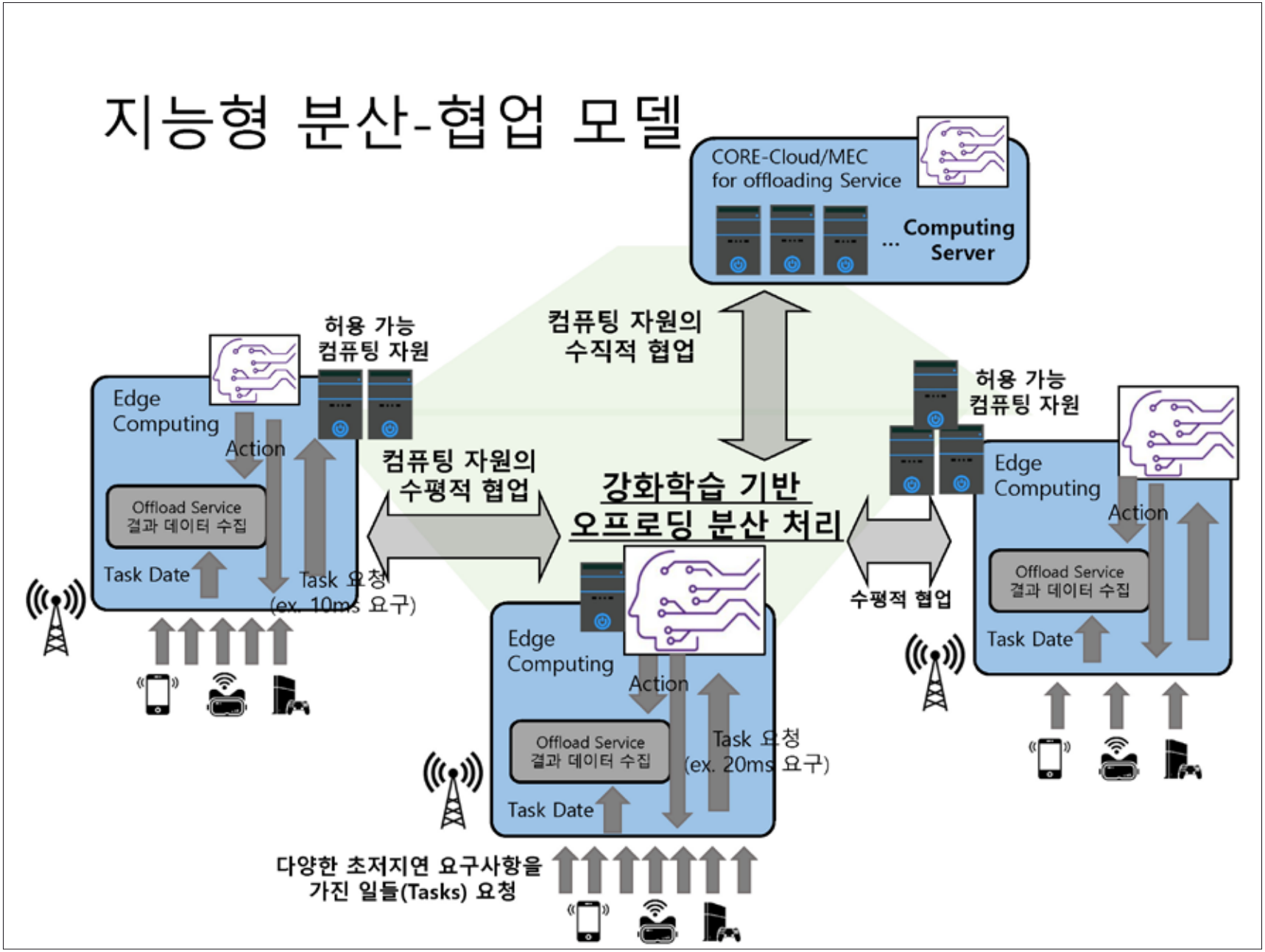


CC: Collaborative Coordinator

엣지 내 협업 요소 기술 도출

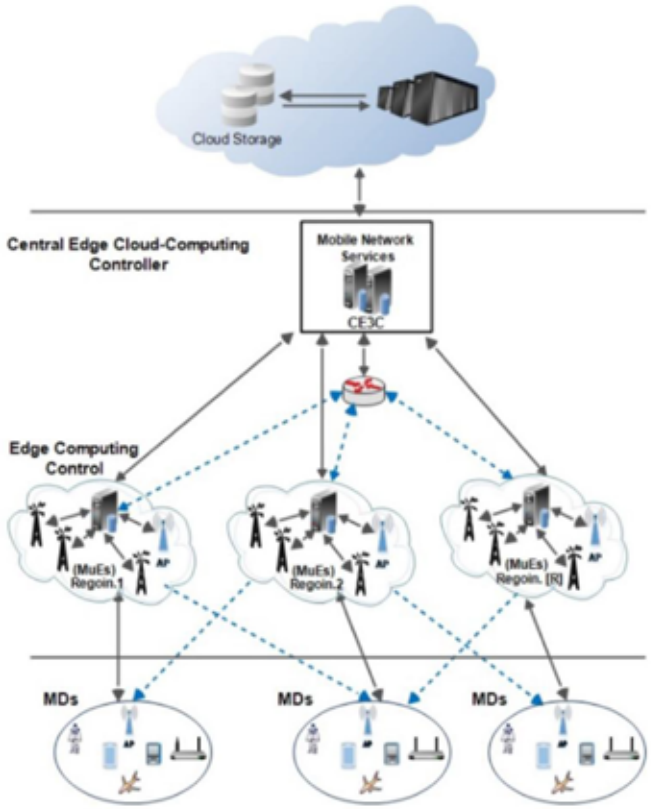
- COS, DCS/DSS 협업 시나리오 가정
- 요소 기술 안
  - Policy for collaborative triggering
  - Task scheduling in distributed computing manner
  - Collaborative computation resource allocation
    - Intelligent RA based DRL algorithm for optimal offloading action with distributed edge node or cloud
      - Training element & process
      - Learning based Action Prediction Method
      - Adaptive Allocation Strategy
      - Optimal resource allocation scheme




지능형 분산-협업 모델
CORE-Cloud/MEC
for offloading Service
Computing
Server
컴퓨팅 자원의
수직적 협업
허용 가능
컴퓨팅 자원
Edge
Computing
Action
Offload Service
결과 데이터 수집
Task Date
Task 요청
(ex. 10ms 요구)
컴퓨팅 자원의
수평적 협업
강화학습 기반
오프로딩 분산 처리
허용 가능
컴퓨팅 자원
Edge
Computing
Action
Offload Service
결과 데이터 수집
Task Date
수평적 협업
Edge
Computing
Action
Offload Service
결과 데이터 수집
Task 요청
(ex. 20ms 요구)
Task Date
다양한 초저지연 요구사항을
가진 일들(Tasks) 요청

협업 지원 **지능형** 컴퓨팅 오프로딩 연구 동향

- High-level overview of co mputation offloading in M EC model
  - Communication Model
  - Task & Offloading Model
  - Computation Model
  - Edge Processing Time
    - Local Processing Time
    - Edge Processing Time
    - Processing Time of Adjacent Edge Server
    - Remote Processing Time (Clo ud Server)



Problem Formulation

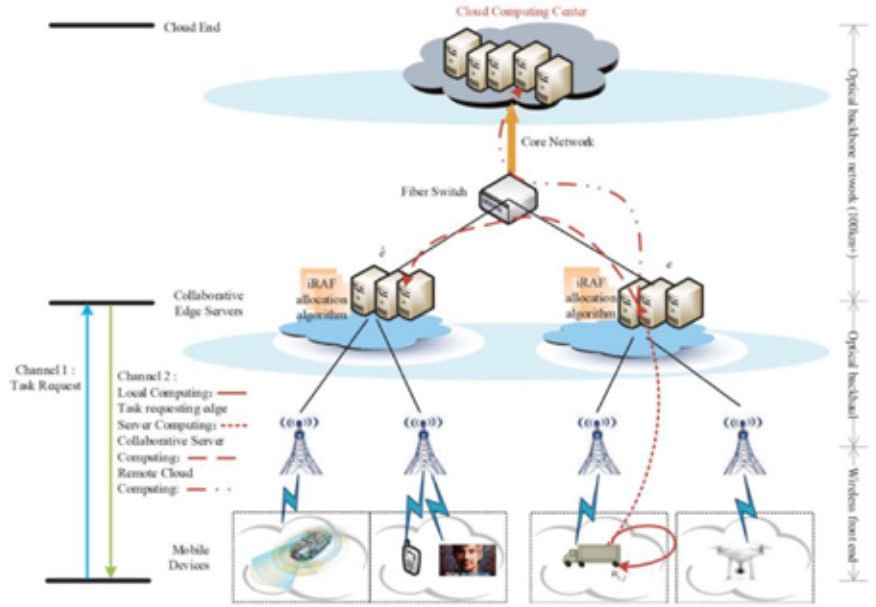
- The minimized cost is denoted by *Qmin*
  - Our objective to minimize the processing and transmission delay and reduce the power consumption
- DQN based Offloading Decision Method

Algorithm 1 OD-SARSA

1: Input: Number of MDs N and task size D_n
2: Output: efficient offloading decision, cost reduction and bandwidth allocation
3: Initialize the network parameters with upload and download bandwidth, and processing cycle number
4: Initialize the number of iterations (episodes), let I = 100
5: for iteration I < 1,2,3,…., I do
6: Select *"Action"* randomly.
7: Compute *"Current State"* according to formula No. 3
8: **if** *"Current state"* $< (St + 1)$ **then**
9: Set $r_t = 1$
10: **else if** $S_t > S_{t+1}$ **then**
11: Set $r_t = -1$
12: **else**
13: Set $r_t = 0$
14: **end if**
15: Obtain reward r_t and next state S_{t+1} after execution of a_t.
16: Set this as (S_t, a_t, r_t, S_{t+1}).
17: Compute the Q-value y_t from the target deep QL $y_t = r_{t+1} + \gamma QS_{t+1}, a_{t+1}$
18: Execute the algorithm of gradient descent to reduce $(y_t - q(s_{t+1}, a_{t+1}); \alpha)^2$
19: Update q-value: $q^*(s, a) = (1 - \alpha) q(s, a) + \alpha(R_{t+1} + \gamma q(s_{t+1}, a_{t+1}))$
20: **end for**

협업 지원 **지능형** 오프로딩 연구 동향

- 심층강화학습 기반 지능형 협업 엣지 기술 1



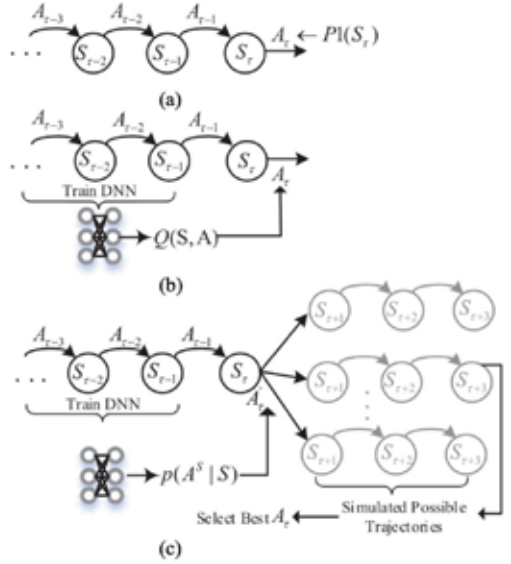


Fig. 2. MDP representation of different methods. (a) Single time slot decision. (b) Deep Q-learning. (c) Deep MCTS.

iRAF: A Deep Reinforcement Learning Approach for Collaborative Mobile E dge Computing IoT Networks, IEEE INTERNET OF THINGS JOURNAL, VOL. 6, NO. 4, AUGUST 2019

협업 지원 **지능형** 오프로딩 연구 동향

- Blockchain-empowered mobile edge computing 모델 구축
  - Deep reinforcement learning-based online computation offloading

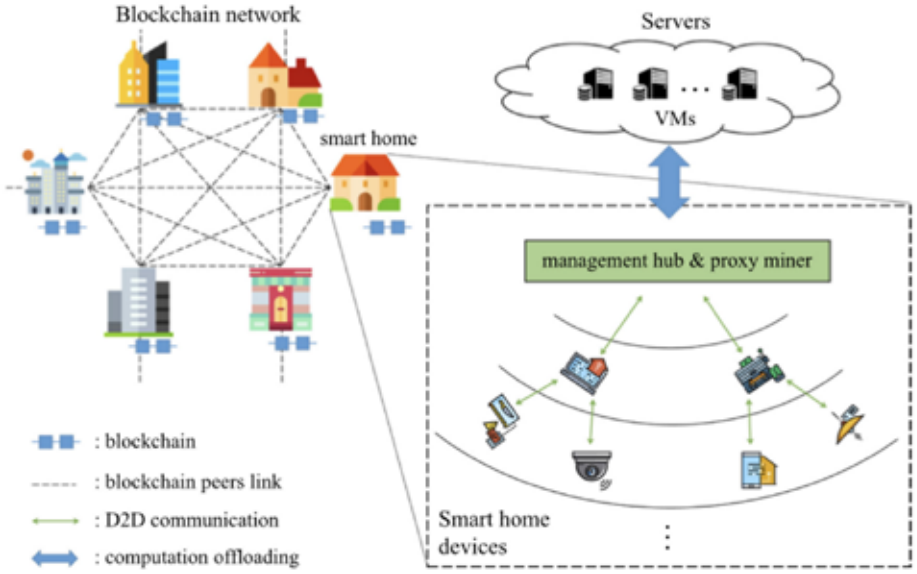
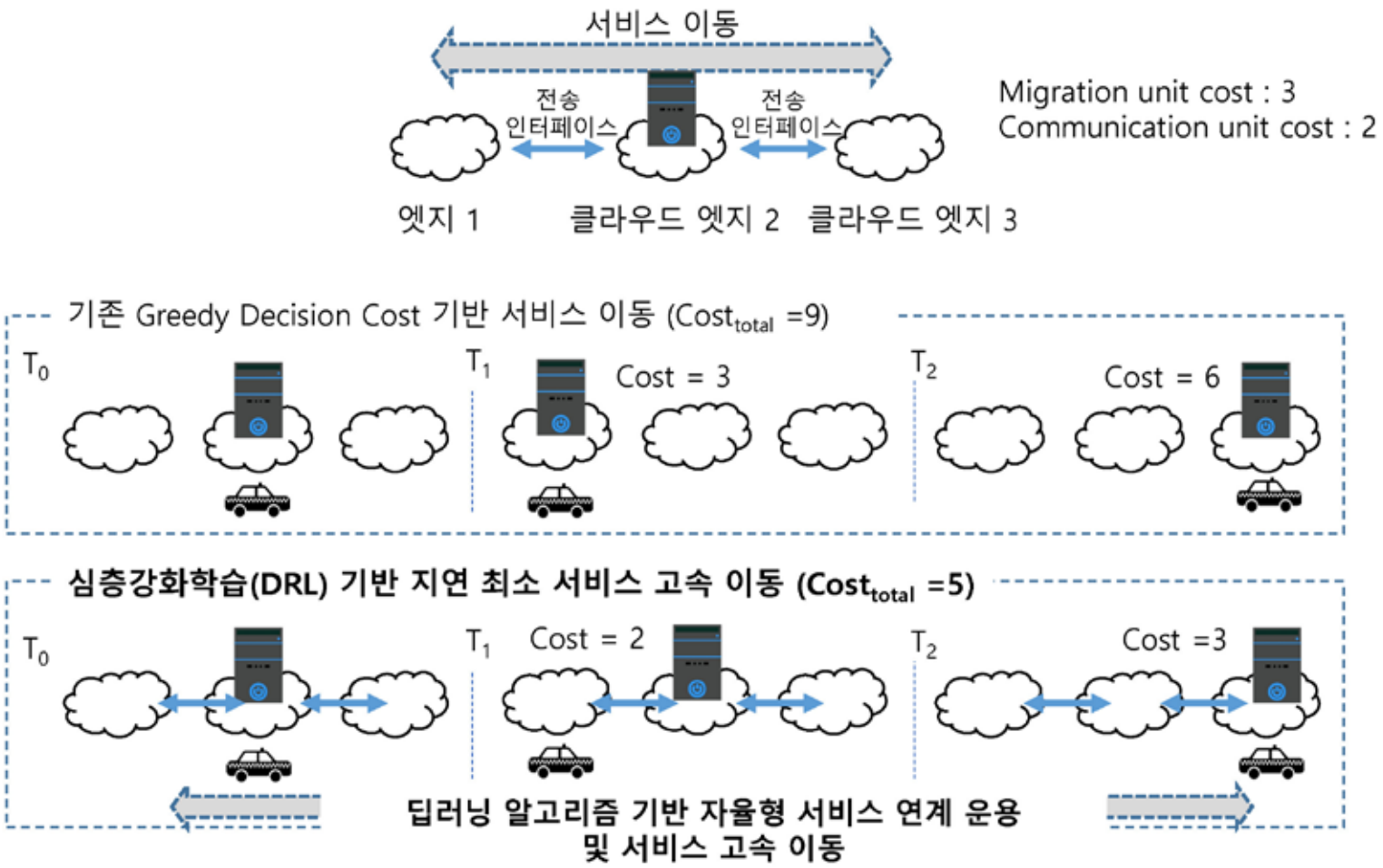


Fig. 1. Multi-hop blockchain-empowered mobile edge computing.

Online Deep Reinforcement Learning for Computation Offloading in Blockc
hain-Empowered Mobile Edge Computing, IEEE TRANSACTIONS ON VEHIC
ULAR TECHNOLOGY, VOL. 68, NO. 8, AUGUST 2019

강화학습 기반 고속 서비스 이동 기술



Conclusion

- 인터넷 기반 응용 서비스 환경 변화
  - 클라우드-분산 엣지 간 협업 모델 개발에 대한 필요성 대두
  - 이를 통해 분산 클라우드 인프라 구축을 추진 중임

- 분산 클라우딩을 위해 클라우드-엣지 간 컴퓨팅 지원 및 네트워크 자원 관리를 위한 분산-협업 기술 개발 중
  - 딥러닝 기반 분산-협업 지원 기술 개발 중