모바일/엣지/클라우드 3 계층 MEC 시스템에서의 서비스 캐싱 연구

함동호, 김영진*, 곽정호
대구경북과학기술원, 인하대학교*

dhham97@dgist.ac.kr, yj.kim@inha.ac.kr jeongho.kwak@dgist.ac.kr

A Service Caching Architecture for Mobile/Edge/Cloud System

Dongho Ham, Yeongjin Kim*, Jeongho kwak
DGIST, Inha University*

요 약

서비스 캐싱은 최종 사용자 근처에 위치한 MEC (Mobile Edge Computing)서버로부터 자원을 빌려 모바일 단말의 하드웨어적 한계를 극복할 수 있는 핵심 기술로 등장하고 있다. 이 기술은 가장 인기가 높은 컨텐츠를 캐싱해야 성능을 극대화할 수 있는 컨텐츠 캐싱과 본질적으로 다르나 최근 많은 관심을 받기 시작 한 서비스 캐싱 연구의 대부분이 이점을 간과하고 있다. 일반적인 서비스는 컴퓨팅 자원도 함께 사용해야 하기에, 너무 많은 사용자들의 서비스 요청이 하나의 MEC 서버에 몰리게 되면 해당 서비스를 해당 서버에 캐싱 하는 것이 사용자들의 지연 관점에서 최선의 선택이 아닐 수 있다. 이 논문에서, 우리는 모바일/엣지/클라우드의 3 계층 서비스 캐싱 구조에서 MEC 서버 와 모바일 단말의 에너지-지연을 최소화하기 위하여 동적 서비스 캐싱 및 오프로딩 결정을 포함한 모바일의 자원 할당 문제를 다룬다. 이를 위해, 우리는 대기열 (queue)을 기반으로 하는 3 계층 서비스 캐싱 시스템을 모델링 하였고, 통계 기반 동적최적화 프레임워크를 시스템에 적용하여 매 시간슬롯 마다 제어 파라미터를 결정하는 현실적인 문제로 유도하였다.

Ⅰ. 서 론

최근 모바일 기기들의 성능진화와 네트워크 속도, 클라우드 컴퓨팅의 발전은 사용자들로 하여금 언제 어디서든 고성능의 컴퓨팅을 활용할 수 있게 하였다. 다시 말해서, 기존 모바일 기기의 컴퓨팅, 메모리, 스토리지, 배터리 자원의 제약을 극복하고, 여러가지 6G 서비스들이 요구하는 자원들을 충족시킬 수 있게 되었다. 이렇게 서비스에 활용될 자원의 시공간적 제약이 줄어들면서 증강현실 (AR), 메타버스, 그리고 상호작용 온라인 게임과 같은 자원 집약적인 서비스들이 모바일에서 제공될 수 있게 되었다. 이러한 서비스들은 대부분 많은 컴퓨팅, 스토리지 자원을 소모할 뿐만 아니라 실시간으로 모바일과 클라우드 간에 대량의 데이터들이 교환된다는 특징이 있다. 하지만, 모바일 단말이 가진 적은 컴퓨팅 및 스토리지 자원만으로는 이러한 서비스들을 이용하기에 한계가 있다.

일반적으로 모바일 단말들의 CPU 및 GPU 클럭 속도는 서버보다 낮고, 모바일 단말의 CPU/GPU 코어의 갯수는 서버보다 더 적기 때문에 서비스를 처리하는데 시간이 더 오래 걸린다. 게다가 모바일 단말들은 대부분의 시간을 유선으로 연결되어 있지 않은 배터리에 의존하기 때문에 에너지 소모에 민감하며, 따라서 서비스 워크로드의 상황에 맞추어 CPU/GPU 클럭의 조절을 하는 DVFS (Dynamic Voltage and Frequency Scaling) 기능을 가지고 있다. 이러한 모바일 단말의 한계를 극복하기 위하여 풍부한 프로세싱/스토리지 자원을 가진 클라우드 컴퓨팅으로 바로 서비스의 워크로드를 오프로딩하는 방안이 제안되었다. 하지만 이것 역시 서비스 성능을 향상시키는 관점에서 영구적인 해결책이 되지는 못하는데 그 이유는 클라우드 컴퓨팅 서버와 사용자의 물리적 거리가 멀기 때문에 데이터 전송률 이 떨어지거나 전송지연이 많이 발생하여 서비스의 질 (예: FPS)에 직접적인 영향을 주기 때문이다.

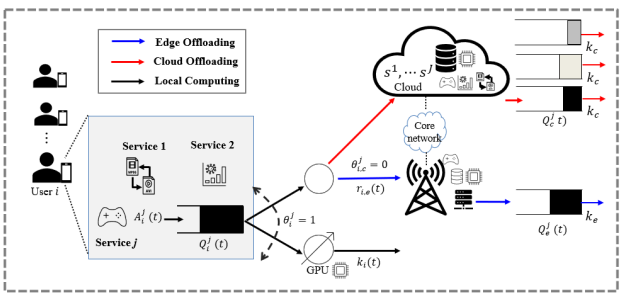

그림 1. 모바일/엣지/클라우드의 3 계층 시스템

Ⅱ. 본론

이러한 모바일 컴퓨팅과 클라우드 컴퓨팅의 단점을 상쇄하기 위한 방법으로 컴퓨팅 및 스토리지 자원을 서비스 사용자와 물리적으로 가까운 위치에 두는 모바일 엣지 컴퓨팅 (MEC) 아키텍처가 제안되었다[1]. 사용자와 가까운 위치에 MEC 서버를 배치하게 되면, 클라우드 컴퓨팅 서버보다 물리적으로 사용자와 훨씬 가까운 위치에 MEC 서버가 있기 때문에 네트워킹 지연을 줄일 수 있고, 모바일 단말보다 더 많은 컴퓨팅 및 스토리지 자원을 가지고 있기 때문에 프로세싱 지연을 줄일 수 있다. 하지만 이러한 MEC 아키텍처도 다음과 같은 문제가 존재한다. MEC 서버는 클라우드와는 다르게 사용자들의 요구에 따라 여러 서버에 자원을 분산시켜야 하기 때문에, 항상 풍부한 자원을 사용할 수는 없다. 따라서 각 MEC 서버에 어떤 서비스를 캐싱할 것인지는 MEC 서버의 스토리지 컴퓨팅 자원 측면에서 중요한 문제이다.

서비스 캐싱은 반복적으로 요청되는 컨텐츠를 MEC 서버에 캐싱함으로써 컨텐츠의 전달지연을 줄이는 컨텐츠 캐싱 기술과 유사한 측면이 있다. 하지만 서비스는 컨텐츠와는 다르게 서비스를 실행하기 위한 컴퓨팅 자원을 활용해야 하기 때문에 무조건 인기있는 서비스를 MEC 서버에 캐싱하는 것이 정답이 아닐 수 있다. 예를 들어, 인기있는 서비스라고 할지라도 모바일 단말이 MEC 서버로 오프로딩하지 않으면 MEC 서버의 활용률이 떨어질 수 있고, 반대로 너무 많은 서비스의 워크로드들이 MEC 서버로 오프로딩되면 MEC 서버의 처리지연이 길어져 서비스 성능이 떨어질 수 있다. 따라서 무선채널의 환경, 서비스 요청량 및 코드 오프로딩과의 관계 등을 종합적으로 고려하여 동적으로 서비스를 캐싱할 필요가 있다.

우리는 그림 1 과 같이 하나의 클라우드, 하나의 MEC 서버가 연결되어 있는 하나의 기지국 그리고 여러 명의 사용자들을 고려하는 실시간 3 계층 시스템 아키텍쳐를 제안한다. 서비스 제공자는 MEC 서버에 서비스 캐싱을 통해 엣지 컴퓨팅을 기회적으로 제공한다. 사용자는 자신의 모바일 기기를 통해 직접 서비스를 실행하거나 무선으로 연결된 엣지 혹은 유/무선으로 연결된 클라우드를 통해 서비스를 실행한다. 클라우드에서는 모든 서비스를, 엣지에서는 MEC 서버의 자원 상황에 따라 일부 서비스를 캐시한다. 엣지 서버에 사용자가 요구하는 서비스가 캐시되어 있다면 사용자는 서비스의 실행 공간 선택지로 세 가지 (로컬, 엣지, 클라우드)가 존재하며, 그렇지 않은 경우에는 선택지가 두 가지(로컬, 클라우드)로 줄어든다.

Ⅲ. 결론

이 논문에서 우리는 여러 명의 사용자와 하나의 MEC 서버와 하나의 클라우드 서버가 있는 실시간 클라우드-지원 MEC 시스템을 고려한다. 사용자는 서비스를 실행하는 총 세 가지의 방법을 가지고 있다. (i)로컬 컴퓨팅: 사용자의 모바일기기을 사용하여 직접 로컬 컴퓨팅을 진행한다. (ii)클라우드컴퓨팅: 사용자의 컴퓨팅 작업을 클라우드로 포워딩한 뒤,작업의 결과를 전달받는다. (iii)엣지 컴퓨팅: 사용자의 컴퓨팅 작업을 엣지에서 대신한다. 이 경우는 기지국에 사용자가 요구하는 서비스가 캐시되어 있을 때만 가능하다.

ACKNOWLEDGMENT

이 논문은 2022 년도 정부 (과학기술정보통신부)의 재원으로 정보통신기획평가원 (No.RS-2022-00155915 인공지능융합혁신인재양성(인하대학교), 2021-0-02201 사용자 프라이버시를 보존하는 비디오 캐싱을 위한 연합 학습 시스템, No.2022-0-00448 인간처럼 회상이 가능한 인공 신경망 지속학습 플랫폼 개발), (No.2022-0-00704, 초고속 이동체 지원을 위한 3D-NET 핵심 기술 개발) 및 한국연구재단 (No. 2020R1F1A1065638)의 지원을 받아 수행된 연구임.

참 고 문 헌

[1] J. Xu, L. Chen, and P. Zhou, “Joint service caching and task offloading for mobile edge computing in dense networks,” in proc of IEEE INFOCOM, Honolulu, USA, Apr. 2018, pp. 207-215.