



(19) 대한민국특허청(KR)
(12) 등록특허공보(B1)

(45) 공고일자 2022년05월19일
(11) 등록번호 10-2400158
(24) 등록일자 2022년05월16일

(51) 국제특허분류(Int. Cl.)
H04L 41/00 (2022.01) H04L 47/00 (2022.01)
H04L 47/80 (2022.01)
(52) CPC특허분류
H04L 41/0896 (2022.05)
H04L 41/083 (2022.05)
(21) 출원번호 10-2020-0170471
(22) 출원일자 2020년12월08일
심사청구일자 2020년12월08일
(56) 선행기술조사문헌
KR1020200085005 A*
KR1020200109303 A*
*는 심사관에 의하여 인용된 문헌

(73) 특허권자
인하대학교 산학협력단
인천광역시 미추홀구 인하로 100(용현동, 인하대학교)
(72) 발명자
김영진
인천광역시 연수구 송도과학로27번길 55, 201동 1412호(송도동, 롯데캐슬 캠퍼스타운)
임정아
인천광역시 미추홀구 인하로47번길 113, 204호(용현동)
(74) 대리인
양성보

전체 청구항 수 : 총 3 항

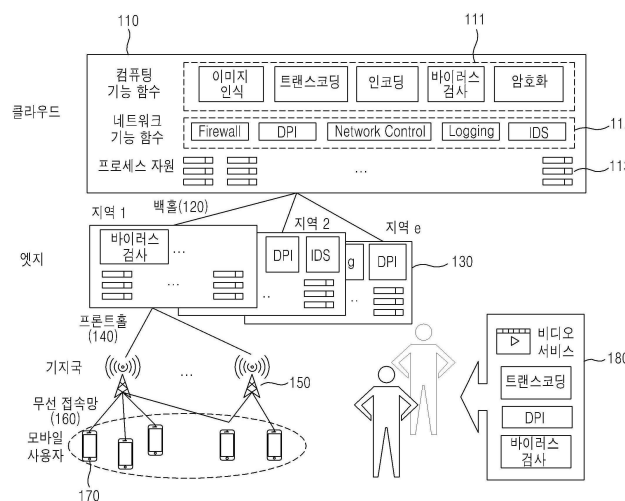
심사관 : 윤태섭

(54) 발명의 명칭 계층적 5G 네트워크 구조에서 서비스 체이닝을 위한 동적 자원 할당 방법 및 장치

(57) 요약

계층적 5G 네트워크 구조에서 서비스 체이닝을 위한 동적 자원 할당 방법 및 장치가 제시된다. 일 실시예에 따른 컴퓨터 장치를 통해 구현되는 계층적 5G 네트워크 구조에서 서비스 체이닝을 위한 동적 자원 할당 방법은, 사용자 단말을 통해 사용자의 서비스 요청에 따라, 클라우드 서버 및 엣지 서버 중 적어도 어느 하나 이상의 서버에서 상기 서비스에 필요한 함수를 처리하는 단계; 및 상기 클라우드 서버, 상기 엣지 서버 및 상기 사용자 단말을 연결하는 백홀, 프론트홀 링크 및 무선접속망 중 적어도 어느 하나 이상을 통해, 상기 클라우드 서버 및 엣지 서버 중 적어도 어느 하나 이상의 서버에서 처리된 상기 서비스가 사용자에게 전달되는 단계를 포함하여 이루어질 수 있다.

대표도



(52) CPC특허분류

H04L 47/782 (2022.05)

H04L 47/80 (2022.05)

이 발명을 지원한 국가연구개발사업

과제고유번호	1711118368
과제번호	2020R1F1A1065638
부처명	과학기술정보통신부
과제관리(전문)기관명	한국연구재단
연구사업명	기본연구
연구과제명	[Ezbaro] 클라우드/엣지 인프라 가상화 기반 다중 자원 관리 기술 연구
기 여 율	1/10
과제수행기관명	인하대학교
연구기간	2020.06.01 ~ 2021.02.28

이 발명을 지원한 국가연구개발사업

과제고유번호	1711116910
과제번호	2020-0-01389-001
부처명	과학기술정보통신부
과제관리(전문)기관명	정보통신기획평가원
연구사업명	인공지능융합연구센터지원사업(국고)
연구과제명	[Ezbaro][정부] 인공지능융합연구센터지원(1차년도)
기 여 율	9/10
과제수행기관명	인하대학교 산학협력단
연구기간	2020.04.01 ~ 2020.12.31

명세서

청구범위

청구항 1

컴퓨터 장치를 통해 구현되는 계층적 5G 네트워크 구조에서 서비스 체이닝을 위한 동적 자원 할당 방법에 있어서,

사용자 단말을 통해 사용자의 서비스 요청에 따라, 클라우드 서버 및 엣지 서버 중 적어도 어느 하나 이상의 서버에서 상기 서비스에 필요한 함수를 처리하는 단계; 및

상기 클라우드 서버, 상기 엣지 서버 및 상기 사용자 단말을 연결하는 백홀, 프론트홀 링크 및 무선접속망 중 적어도 어느 하나 이상을 통해, 상기 클라우드 서버 및 엣지 서버 중 적어도 어느 하나 이상의 서버에서 처리된 상기 서비스가 사용자에게 전달되는 단계

를 포함하고,

상기 서비스에 필요한 함수를 처리하는 단계는,

상기 클라우드 서버에서 서비스 경로로 각 네트워크 함수 또는 컴퓨팅 함수를 상기 클라우드 서버 및 엣지 서버 중 어느 서버에서 처리할지 결정하는 서비스 체이닝 단계;

상기 클라우드 서버에서 프로세스 자원 또는 네트워크 자원의 사용을 결정하여 프로세스 또는 네트워크 스케줄링을 제어하는 스케줄링 단계;

상기 클라우드 서버의 결정에 따라 각 상기 엣지 서버가 동작되며, 각 상기 엣지 서버에서 프로세스 스케줄링을 제어하는 엣지 서버의 스케줄링 단계; 및

상기 엣지 서버에서 기지국에서 상기 사용자에게 서비스를 제공할 때 무선접속망을 빔 활성화 상태가 되도록 하고, 처리가 완료된 서비스를 어느 사용자에게 줄 것을 결정하는 사용자 스케줄링 단계

를 포함하고,

상기 클라우드 서버는 프로세스 스케줄링을 결정하며, 처리된 트래픽이 상기 클라우드 서버에서 상기 엣지 서버로 연결되는 상기 백홀의 경로의 네트워크 큐에 쌓이며, 쌓인 상기 트래픽은 상기 클라우드 서버의 네트워크 스케줄링 결정에 따라, 상기 클라우드 서버에서 서비스가 처리되는 경로는 상기 엣지 서버에서 사용자로 연결되는 상기 프론트홀 링크와 상기 무선접속망의 네트워크 큐에 쌓이고, 상기 엣지 서버에서 서비스가 처리되는 경로 및 상기 클라우드 서버와 상기 엣지 서버 모두에서 서비스가 처리되는 경로는 상기 엣지 서버의 경로의 프로세싱 큐에 쌓이며, 상기 엣지 서버의 경로의 프로세싱 큐에 쌓인 트래픽은 상기 엣지 서버의 프로세스 스케줄링 결정으로 처리되어, 상기 엣지 서버에서 사용자로 연결되는 상기 프론트홀 링크와 상기 무선접속망의 네트워크 큐에 쌓이는 것

을 특징으로 하는, 동적 자원 할당 방법.

청구항 2

삭제

청구항 3

삭제

청구항 4

제1항에 있어서,

네트워크와 서비스 환경이 실시간으로 변화하는 상황에서 상기 클라우드 서버, 상기 엣지 서버 및 상기 무선접속망의 전 계층을 고려해 동적인 네트워크 및 프로세싱 자원을 할당하는 것

을 특징으로 하는, 동적 자원 할당 방법.

청구항 5

계층적 5G 네트워크 구조에서 서비스 체이닝을 위한 동적 자원 할당 장치에 있어서,

사용자 단말을 통해 사용자의 서비스 요청에 따라, 서비스 경로로 각 네트워크 함수 또는 컴퓨팅 함수를 클라우드 서버 및 엣지 서버 중 어느 서버에서 처리할지 결정하고, 프로세스 자원 또는 네트워크 자원의 사용을 결정하여 프로세스 또는 네트워크 스케줄링을 제어하는 클라우드 서버; 및

상기 클라우드 서버의 결정에 따라 동작되며, 엣지 서버의 프로세스 스케줄링을 제어하고, 기지국에서 상기 사용자에게 서비스를 제공할 때 무선접속망을 빔 활성화 상태가 되도록 하고, 처리가 완료된 서비스를 어느 사용자에게 줄 것을 결정하는 엣지 서버

를 포함하고,

상기 클라우드 서버는,

상기 클라우드 서버에서 서비스 경로로 각 네트워크 함수 또는 컴퓨팅 함수를 상기 클라우드 서버 및 엣지 서버 중 어느 서버에서 처리할지 결정하는 서비스 체이닝 후, 상기 클라우드 서버에서 프로세스 자원 또는 네트워크 자원의 사용을 결정하여 프로세스 또는 네트워크 스케줄링을 제어하며,

상기 엣지 서버는,

클라우드 서버의 결정에 따라 각 상기 엣지 서버가 동작되며, 각 상기 엣지 서버에서 프로세스 스케줄링을 제어하고, 상기 엣지 서버에서 기지국에서 상기 사용자에게 서비스를 제공할 때 무선접속망을 빔 활성화 상태가 되도록 하고, 처리가 완료된 서비스를 어느 사용자에게 줄 것을 결정하며,

상기 클라우드 서버는 프로세스 스케줄링을 결정하며, 처리된 트래픽이 상기 클라우드 서버에서 상기 엣지 서버로 연결되는 백홀의 경로의 네트워킹 큐에 쌓이며, 쌓인 상기 트래픽은 상기 클라우드 서버의 네트워크 스케줄링 결정에 따라, 상기 클라우드 서버에서 서비스가 처리되는 경로는 상기 엣지 서버에서 사용자로 연결되는 프론트홀 링크와 상기 무선접속망의 네트워킹 큐에 쌓이고, 상기 엣지 서버에서 서비스가 처리되는 경로 및 상기 클라우드 서버와 상기 엣지 서버 모두에서 서비스가 처리되는 경로는 상기 엣지 서버의 경로의 프로세싱 큐에 쌓이며, 상기 엣지 서버의 경로의 프로세싱 큐에 쌓인 트래픽은 상기 엣지 서버의 프로세스 스케줄링 결정으로 처리되어, 상기 엣지 서버에서 사용자로 연결되는 상기 프론트홀 링크와 상기 무선접속망의 네트워킹 큐에 쌓이는 것

을 특징으로 하는, 동적 자원 할당 장치.

발명의 설명

기술 분야

[0001] 아래의 실시예들은 계층적 5G 네트워크 구조에서 서비스 체이닝을 위한 동적 자원 할당 방법 및 장치에 관한 것으로, 더욱 상세하게는 계층적 5G 네트워크 구조에서 사용자가 요청하는 서비스를 처리하고, 적은 딜레이와 함께 사용자에게 높은 품질의 서비스를 제공하도록 서비스 경로를 선택하는 동적 자원 할당 방법 및 장치에 관한 것이다.

배경 기술

[0002]가상화 기술이 도입됨에 따라 네트워킹 자원, 프로세싱 자원, 스토리지 자원을 클러스터링하여 함께 사용하는 서비스들이 등장하게 되었다. 소프트웨어 정의 네트워킹(SDN)과 네트워크 기능 가상화(NFV) 기술이 가상화 기술에 해당하는데, 이러한 기술들을 사용하여 네트워크 환경에서 필요로 하는 deep packet inspection(DPI), intrusion detection system(IDS), firewall 등의 네트워크 기능 함수를 일반 서버에서 가상 머신이나 컨테이너 기술을 통하여 가상화 할 수 있게 되었다.

[0003]가상화 기술이 도입되기 이전에는 이러한 네트워크 기능 함수 하나 하나가 하드웨어 기반으로 구현되었기 때문에 해당 네트워크 기능을 필요로 하는 서비스의 경우 반드시 그 기능을 지원하는 하드웨어 추가적으로 구입을 하여 그 하드웨어를 반드시 거쳐가야만 했다. 가상화 기술은 네트워크 기능 함수뿐만 아니라 바이러스 검사,

데이터 처리에 해당하는 컴퓨팅 기능 함수 등 다양한 가상 함수의 사용과 네트워크, 프로세스, 스토리지 자원의 공동 관리가 가능해졌다.

[0004] 따라서 인프라 관리자는 필요한 경우에 따라 서비스에 필요한 네트워크 기능 함수나 컴퓨팅 기능 함수를 많은 비용 없이 일반 서버에 추가할 수 있게 되었다. 덕분에 네트워크 관리자는 유연한 자원 관리를 통해서 네트워크, 프로세스, 스토리지 자원을 효율적으로 사용하여 다양한 서비스를 제공할 수 있게 되었다. 예를 들어, 동영상 스트리밍 서비스를 제공하는 경우 트랜스코딩, DPI 등 네트워크와 프로세싱 자원을 공동으로 최적화하여 사용자에게 높은 품질의 서비스를 제공한다.

[0005] 한국등록특허 10-2156439호는 이러한 클라우드-엣지 시스템 및 이의 데이터 처리 방법에 관한 것으로, 클라우드-엣지에서의 파이프라인 구조의 분석 모델을 통해 데이터를 처리할 수 있는 시스템 및 방법에 관한 기술을 기재하고 있다.

선행기술문헌

특허문헌

[0006] (특허문헌 0001) 한국등록특허 10-2156439호

발명의 내용

해결하려는 과제

[0007] 실시예들은 계층적 5G 네트워크 구조에서 서비스 체이닝을 위한 동적 자원 할당 방법 및 장치에 관하여 기술하며, 보다 구체적으로 클라우드-엣지-무선접속망으로 이루어진 계층적 5G 네트워크 구조의 클라우드부터 백홀, 엣지, 프론트홀, 무선접속망까지 네트워크 전 계층의 동적 자원을 관리하는 기술을 제공한다.

[0008] 실시예들은 네트워크와 서비스 환경이 실시간으로 변화하는 상황에서 전 계층(클라우드, 엣지, 무선접속망)을 고려해 동적인 네트워크, 프로세싱 자원을 할당해 줌으로써 모든 사용자의 서비스 만족도를 최대화하면서도 서비스 지연을 낮출 수 있는 계층적 5G 네트워크 구조에서 서비스 체이닝을 위한 동적 자원 할당 방법 및 장치를 제공하는데 있다.

과제의 해결 수단

[0009] 일 실시예에 따른 컴퓨터 장치를 통해 구현되는 계층적 5G 네트워크 구조에서 서비스 체이닝을 위한 동적 자원 할당 방법은, 사용자 단말을 통해 사용자의 서비스 요청에 따라, 클라우드 서버 및 엣지 서버 중 적어도 어느 하나 이상의 서버에서 상기 서비스에 필요한 함수를 처리하는 단계; 및 상기 클라우드 서버, 상기 엣지 서버 및 상기 사용자 단말을 연결하는 백홀, 프론트홀 링크 및 무선접속망 중 적어도 어느 하나 이상을 통해, 상기 클라우드 서버 및 엣지 서버 중 적어도 어느 하나 이상의 서버에서 처리된 상기 서비스가 사용자에게 전달되는 단계를 포함하여 이루어질 수 있다.

[0010] 상기 서비스에 필요한 함수를 처리하는 단계는, 상기 클라우드 서버에서 서비스 경로로 각 네트워크 함수 또는 컴퓨팅 함수를 상기 클라우드 서버 및 엣지 서버 중 어느 서버에서 처리할지 결정하는 서비스 체이닝 단계; 및 상기 클라우드 서버에서 프로세스 자원 또는 네트워크 자원의 사용을 결정하여 프로세스 또는 네트워크 스케줄링을 제어하는 스케줄링 단계를 포함하여 이루어질 수 있다.

[0011] 상기 서비스에 필요한 함수를 처리하는 단계는, 상기 클라우드 서버의 결정에 따라 각 상기 엣지 서버가 동작되며, 각 상기 엣지 서버에서 프로세스 스케줄링을 제어하는 엣지 서버의 스케줄링 단계; 및 상기 엣지 서버에서 기지국에서 상기 사용자에게 서비스를 제공할 때 무선접속망을 빔 활성화 상태가 되도록 하고, 처리가 완료된 서비스를 어느 사용자에게 줄 것을 결정하는 사용자 스케줄링 단계를 더 포함하여 이루어질 수 있다.

[0012] 네트워크와 서비스 환경이 실시간으로 변화하는 상황에서 상기 클라우드 서버, 상기 엣지 서버 및 상기 무선접속망의 전 계층을 고려해 동적인 네트워크 및 프로세싱 자원을 할당할 수 있다.

[0013] 다른 실시예에 따른 계층적 5G 네트워크 구조에서 서비스 체이닝을 위한 동적 자원 할당 장치는, 사용자 단말을 통해 사용자의 서비스 요청에 따라, 서비스 경로로 각 네트워크 함수 또는 컴퓨팅 함수를 상기 클라우드 서버

및 엣지 서버 중 어느 서버에서 처리할지 결정하고, 프로세스 자원 또는 네트워크 자원의 사용을 결정하여 프로세스 또는 네트워크 스케줄링을 제어하는 클라우드 서버; 및 상기 클라우드 서버의 결정에 따라 동작되며, 엣지 서버의 프로세스 스케줄링을 제어하고, 기지국에서 상기 사용자에게 서비스를 제공할 때 무선접속망을 빔 활성화 상태가 되도록 하고, 처리가 완료된 서비스를 어느 사용자에게 줄 것을 결정하는 엣지 서버를 포함하여 이루어질 수 있다.

발명의 효과

[0014] 실시예들에 따르면 네트워크와 서비스 환경이 실시간으로 변화하는 상황에서 전 계층(클라우드, 엣지, 무선접속망)을 고려해 동적인 네트워크, 프로세싱 자원을 할당해 줌으로써 모든 사용자의 서비스 만족도를 최대화하면서도 서비스 지연을 낮출 수 있는 계층적 5G 네트워크 구조에서 서비스 체이닝을 위한 동적 자원 할당 방법 및 장치를 제공할 수 있다.

[0015] 실시예들에 따르면 사용자가 서비스를 받기 위한 과정에서 발생하는 end-to-end 서비스 지연은 줄이고, 사용자의 서비스 만족도는 높일 수 있다.

도면의 간단한 설명

[0016] 도 1은 일 실시예에 따른 계층적 5G 네트워크 구조를 설명하기 위한 도면이다.

도 2a는 일 실시예에 따른 서비스 체이닝을 위한 동적 자원 할당 방법을 나타내는 흐름도이다.

도 2b는 일 실시예에 따른 서비스 체이닝을 위한 동적 자원 할당 방법의 함수를 처리하는 방법을 나타내는 흐름도이다.

도 3은 일 실시예에 따른 계층적 5G 네트워크 구조의 큐와 제어 모델을 설명하기 위한 도면이다.

도 4는 일 실시예에 따른 정의한 문제를 풀기 위한 알고리즘 유도의 수학적 원리를 설명하기 위한 도면이다.

도 5는 일 실시예에 따른 균등한 사용자 분포에서의 제안된 알고리즘과 변형한 알고리즘의 비교를 나타내는 도면이다.

도 6은 일 실시예에 따른 불균등한 사용자 분포에서의 제안된 알고리즘과 변형한 알고리즘의 비교를 나타내는 도면이다.

발명을 실시하기 위한 구체적인 내용

[0017] 이하, 첨부된 도면을 참조하여 실시예들을 설명한다. 그러나, 기술되는 실시예들은 여러 가지 다른 형태로 변형될 수 있으며, 본 발명의 범위가 이하 설명되는 실시예들에 의하여 한정되는 것은 아니다. 또한, 여러 실시예들은 당해 기술분야에서 평균적인 지식을 가진 자에게 본 발명을 더욱 완전하게 설명하기 위해서 제공되는 것이다. 도면에서 요소들의 형상 및 크기 등은 보다 명확한 설명을 위해 과장될 수 있다.

[0019] 최근에는 클라우드 서버, 엣지 서버, 엔드 디바이스까지 전 계층을 통괄하는 계층적 네트워크 구조 속에서 네트워크, 컴퓨팅 기능 함수를 포함한 서비스를 제공하는 기술이 주목을 받고 있다.

[0020] 본 실시예들은 단일 클라우드 서버, 여러 개의 엣지 서버, 기지국, 그리고 무선접속망까지 한 번에 고려할 수 있는 계층적 5G 이동통신망 네트워크 구조를 고려한다. 계층적 5G 네트워크 구조는 단일 클라우드, 일정 지역을 담당하는 엣지, 클라우드와 엣지를 연결하는 유선 단의 백홀 링크, 각 지역의 기지국, 엣지와 기지국을 연결하는 유선단의 프론트홀 링크, 각 지역의 모바일 사용자, 그리고 각 기지국과 모바일 사용자간의 무선접속망으로 구성되어 있다. 클라우드와 엣지 서버는 서비스를 처리할 수 있는 프로세싱 자원(CPU, GPU)이 존재하고, 백홀/프론트홀/무선접속망 링크에는 네트워크 자원(유/무선단의 가능한 대역폭)이 존재한다. 이러한 구조 속에서 모바일 사용자가 이동통신망에 서비스를 요청하게 되면, 클라우드와 엣지에서 서비스에 필요한 함수를 처리하고, 백홀/프론트홀 링크, 무선접속망을 통해 서비스가 전달된다.

[0021] 이러한 계층적 5G 네트워크 구조에서 클라우드는 최상위층에 존재하며, 대량의 프로세싱 자원을 통하여 전체 시스템을 통합 최적화하는데 효과적이다. 하지만, 클라우드에서 지역별 상세 특성까지 피드백 받아가며 자원을 관리하는 것은 비효율적일 수 있다. 이러한 경우, 계층적인 네트워크 구조 내 엣지의 자원을 활용하면 각 지역의 세밀한 자원 제어에 용이하다. 하지만, 엣지는 클라우드보다 적은 프로세싱 자원을 보유한 약점이 존재하기 때문에 클라우드와 엣지의 프로세싱 자원을 동시에 효율적으로 관리해야만 사용자에게 높은 서비스 품질을 제공

할 수 있다. 뿐만 아니라, 클라우드나 엣지에서 풍부한 프로세싱 자원으로 사용자가 요청한 서비스를 처리한다고 하여도, 백홀/프론트홀/무선접속망 링크의 네트워크 자원을 효율적으로 관리하지 못하면 사용자는 낮은 품질의 서비스를 제공받게 된다. 따라서 가능한 서비스의 경로 중에서 네트워크 자원과 프로세싱 자원을 동시에 효율적으로 사용을 하면서도 동시에 높은 서비스 만족도를 달성해야만 한다.

- [0022] 기존 연구에서는 클라우드-엣지의 계층적 구조에서의 서비스 경로 문제만을 다루었을 뿐, 클라우드-엣지-무선접속망 계층적 구조에서의 서비스 경로 문제는 다루지 않았다. 하지만 다양한 서비스를 빠르게 제공해야 하는 현재 네트워크 상황에서 서비스 처리와 자원관리는 동시에 이루어지는 것은 자연스러운 일이다.
- [0023] 본 실시예에서는 계층적 5G 네트워크 구조에서 사용자가 요청하는 서비스를 처리하고, 적은 딜레이와 함께 사용자에게 높은 품질의 서비스를 제공하도록 서비스 경로를 선택하는 기술을 제안한다.
- [0024] 실시예들은 네트워크와 서비스 환경이 실시간으로 변화하는 상황에서 전 계층(클라우드, 엣지, 무선접속망)을 고려해 동적인 네트워크, 프로세싱 자원을 할당해 줌으로써 모든 사용자의 서비스 만족도를 최대화하면서도 서비스 지연을 낮출 수 있다. 이 기법은 서비스를 지원하기 위하여 매 순간 클라우드에서 (1) 서비스 체이닝(서비스 경로로 각 네트워크/컴퓨팅 함수를 클라우드/엣지에서 처리할지 결정), (2) 프로세스/네트워크 스케줄링(프로세스/네트워크 자원의 사용을 결정)을 제어하고, 각 엣지에서 (1) 프로세스 스케줄링, (2) 무선접속망의 빔 활성화(기지국에서 사용자에게 서비스를 제공할 때 활성화) 및 사용자 스케줄링(처리가 완료된 서비스를 어느 유저에게 줄 것을 결정)을 결정한다.
- [0026] 도 1은 일 실시예에 따른 계층적 5G 네트워크 구조를 설명하기 위한 도면이다.
- [0027] 도 1에 도시된 바와 같이, 클라우드(클라우드 서버, 110), 엣지(엣지 서버, 130), 무선 통신망(120, 140, 160)으로 이루어진 계층적 5G 네트워크 환경에서의 적은 딜레이와 함께 사용자에게 높은 품질의 서비스를 제공하도록 서비스 경로를 선택하고 동적 자원을 할당할 수 있다. 여기서, 클라우드(110)는 컴퓨팅 함수(111), 네트워크 함수(112) 및 프로세스 자원(1130)을 포함할 수 있다.
- [0028] 복수개의 엣지(130)는 하나의 클라우드(110)에 백홀 링크(120)로 연결되어 있고, 각 엣지(130)에는 프론트홀 링크(140)로 연결된 여러 기지국(150)이 존재한다. 각 엣지(130)가 관할하는 지역 내 사용자 단말(170)은 해당 엣지(130) 내 기지국(150) 무선접속망(160)을 통해 서비스를 제공받는다.
- [0029] 이 때, 무선접속망(160)은 CoMP가 고려된 상황으로, 사용자(170)가 기지국(150)의 영향을 받을 수 있는 거리에 위치했을 경우, 여러 개의 기지국(150)에서 서비스를 받을 수 있고, 사용자(170)가 서비스를 받을 때 한 기지국(150) 내에서 여러 개의 빔을 할당 받을 수 있다.
- [0030] 사용자 서비스(180)는 네트워크 기능 함수와 컴퓨팅 기능 함수의 연속적 형태로 정의된다. 사용자(170)가 요청한 서비스 내의 각 함수는 클라우드(110) 혹은 엣지(130)의 프로세싱을 통해 처리된다.
- [0032] 도 2a는 일 실시예에 따른 서비스 체이닝을 위한 동적 자원 할당 방법을 나타내는 흐름도이다.
- [0033] 도 2a를 참조하면, 일 실시예에 따른 컴퓨터 장치를 통해 구현되는 계층적 5G 네트워크 구조에서 서비스 체이닝을 위한 동적 자원 할당 방법은, 사용자 단말을 통해 사용자의 서비스 요청에 따라, 클라우드 서버 및 엣지 서버 중 적어도 어느 하나 이상의 서버에서 서비스에 필요한 함수를 처리하는 단계(S110), 및 클라우드 서버, 엣지 서버 및 사용자 단말을 연결하는 백홀, 프론트홀 링크 및 무선접속망 중 적어도 어느 하나 이상을 통해, 클라우드 서버 및 엣지 서버 중 적어도 어느 하나 이상의 서버에서 처리된 서비스가 사용자에게 전달되는 단계(S120)를 포함하여 이루어질 수 있다.
- [0034] 도 2b는 일 실시예에 따른 서비스 체이닝을 위한 동적 자원 할당 방법의 함수를 처리하는 방법을 나타내는 흐름도이다.
- [0035] 도 2b를 참조하면, 단계(S110)는, 클라우드 서버에서 서비스 경로로 각 네트워크 함수 또는 컴퓨팅 함수를 클라우드 서버 및 엣지 서버 중 어느 서버에서 처리할지 결정하는 서비스 체이닝 단계(S111), 및 클라우드 서버에서 프로세스 자원 또는 네트워크 자원의 사용을 결정하여 프로세스 또는 네트워크 스케줄링을 제어하는 스케줄링 단계(S112)를 포함하여 이루어질 수 있다.
- [0036] 또한, 클라우드 서버의 결정에 따라 각 엣지 서버가 동작되며, 각 엣지 서버에서 프로세스 스케줄링을 제어하는 엣지 서버의 스케줄링 단계(S113), 및 엣지 서버에서 기지국에서 사용자에게 서비스를 제공할 때 무선접속망을 빔 활성화 상태가 되도록 하고, 처리가 완료된 서비스를 어느 사용자에게 줄 것을 결정하는 사용자 스케줄링 단

계(S114)를 더 포함하여 이루어질 수 있다.

- [0037] 실시예들에 따르면 네트워크와 서비스 환경이 실시간으로 변화하는 상황에서 클라우드 서버, 엣지 서버 및 무선 접속망의 전 계층을 고려해 동적인 네트워크 및 프로세싱 자원을 할당할 수 있다.
- [0038] 일 실시예에 따른 서비스 체이닝을 위한 동적 자원 할당 방법은 일 실시예에 따른 서비스 체이닝을 위한 동적 자원 할당 장치를 예를 들어 설명할 수 있다. 일 실시예에 따른 서비스 체이닝을 위한 동적 자원 할당 장치는 클라우드 서버 및 엣지 서버를 포함하여 이루어질 수 있다.
- [0040] 아래에서 일 실시예에 따른 서비스 체이닝을 위한 동적 자원 할당 방법의 각 단계를 보다 상세히 설명한다. 일 실시예에 따른 서비스 체이닝을 위한 동적 자원 할당 방법은 서비스 체이닝을 위한 동적 자원 할당 장치를 통해 수행될 수 있다.
- [0041] 단계(S110)에서, 사용자 단말을 통해 사용자의 서비스 요청에 따라, 클라우드 서버 및 엣지 서버 중 적어도 어느 하나 이상의 서버에서 서비스에 필요한 함수를 처리할 수 있다.
- [0042] 보다 구체적으로, 서비스 체이닝 단계(S111)에서, 클라우드 서버는 사용자 단말을 통해 사용자의 서비스 요청에 따라, 서비스 경로로 각 네트워크 함수 또는 컴퓨팅 함수를 클라우드 서버 및 엣지 서버 중 어느 서버에서 처리할지 결정할 수 있다.
- [0043] 스케줄링 단계(S112)에서, 클라우드 서버는 프로세스 자원 또는 네트워크 자원의 사용을 결정하여 프로세스 또는 네트워크 스케줄링을 제어할 수 있다.
- [0044] 엣지 서버의 스케줄링 단계(S113)에서, 엣지 서버는 클라우드 서버의 결정에 따라 동작되며, 엣지 서버의 프로세스 스케줄링을 제어할 수 있다.
- [0045] 사용자 스케줄링 단계(S114)에서, 엣지 서버는 기지국에서 사용자에게 서비스를 제공할 때 무선접속망을 빔 활성화 상태가 되도록 하고, 처리가 완료된 서비스를 어느 사용자에게 줄 것을 결정할 수 있다.
- [0046] 단계(S120)에서, 클라우드 서버, 엣지 서버 및 사용자 단말을 연결하는 백홀, 프론트홀 링크 및 무선접속망 중 적어도 어느 하나 이상을 통해, 클라우드 서버 및 엣지 서버 중 적어도 어느 하나 이상의 서버에서 처리된 서비스가 사용자에게 전달될 수 있다.
- [0047] 실시예들은 클라우드-엣지-무선접속망으로 이루어진 계층적 5G 네트워크 구조에서 서비스 체이닝을 위해 동적 자원을 효율적이게 할당하는 알고리즘을 제공한다. 이로 인해 사용자가 서비스를 받기 위한 과정에서 발생하는 end-to-end 서비스 지연은 줄이고, 사용자의 서비스 만족도는 높이는 알고리즘을 제공한다.
- [0048] 실시예들은 클라우드-엣지-무선접속망으로 이루어진 네트워크 구조가 발전함에 따라서 한정된 프로세스, 네트워크 자원을 효율적으로 사용할 수 있는 알고리즘을 제안함으로써, 실시간으로 네트워크 환경이 변화하는 상황에서 전 계층(클라우드, 엣지, 무선접속망)을 아울러 프로세스와 네트워크 자원을 상황에 맞게 효율적으로 할당할 수 있다.
- [0050] 도 3은 일 실시예에 따른 계층적 5G 네트워크 구조의 큐와 제어 모델을 설명하기 위한 도면이다.
- [0051] 도 3을 참조하면, 계층적 5G 네트워크 구조에서 나타내는 큐와 제어 모델을 나타내며, 표현된 변수들은 다음 표 1과 같이 정리할 수 있다.

[0052] [표 1]

$\mathcal{P}$	모든 사용자의 서비스 경로의 집합
$\mathcal{P}^e$	엣지 e 를 지나는 모든 서비스 경로의 집합
$\hat{\mathcal{P}}$	클라우드에서 모든 기능 함수를 처리하는 서비스 경로의 집합
$\check{\mathcal{P}}$	엣지에서 모든 기능 함수를 처리하는 서비스 경로의 집합
$\tilde{\mathcal{P}}$	클라우드와 엣지 모두에서 기능 함수를 분담하여 처리하는 서비스 경로의 집합
$\hat{p}$	클라우드에서만 요청한 서비스가 처리되는 경로
$\check{p}$	엣지에서만 요청한 서비스가 처리되는 경로
$\tilde{p}$	클라우드와 엣지 두가지 모두에서 서비스가 처리되는 경로
ρ_p^c	Processing density – 클라우드 c 에서 경로 p 를 위해 처리하는 단위 인풋 전송 속도 당 프로세싱 처리 로드 (cycles/bit). 같은 프로세싱 자원에서 수행되는 기능이 경로마다 다를 수 있기 때문에 달라지는 파라미터
ρ_p^e	Processing density – 엣지 e 에서 경로 p 를 위해 처리하는 단위 인풋 전송 속도 당 프로세싱 처리 로드 (cycles/bit)
σ_p^c	Bit conversion ratio – 클라우드 c 에서 경로 p 에 대하여 단위 인풋 전송 속도 당 네트워크 자원을 사용하는 네트워킹 로드 (bits/bit).
σ_p^e	Bit conversion ratio – 엣지 e 에서 경로 p 에 대하여 단위 인풋 전송 속도 당 네트워크 자원을 사용하는 네트워킹 로드 (bits/bit).
$L^c(t) \in \{L_{min}^c, \dots, L_{max}^c\}$	클라우드에서 사용가능한 가능한 프로세싱 리소스 블록 수
$L^e(t) \in \{L_{min}^e, \dots, L_{max}^e\}$	엣지에서 사용가능한 가능한 프로세싱 리소스 블록 수
s^c	클라우드의 단일 리소스 블록의 컴퓨팅 능력 (in cycle/sec)
s^e	엣지의 단일 리소스 블록의 컴퓨팅 능력 (in cycle/sec)
$R^e(t) \in [R_{min}^e, R_{max}^e]$	클라우드에서 엣지 e 로 연결된 백홀 링크의 대역폭 (in bits/sec)
N^e	백홀의 가능한 대역폭을 블록으로 나누는 개수
$\mathcal{B}^e$	엣지 e 에 존재하는 기지국의 집합
$R^b(t) \in [R_{min}^b, R_{max}^b]$	엣지에서 기지국 b 로 연결된 프론트홀 링크 대역폭 (in bits/sec)

[0053]

$y_{b,i}(t) \in \mathbb{C}^M$	기지국 b 의 빔들로부터 사용자 i 까지의 신호 세기
$\omega_i(t)$	사용자 i 채널에서의 간섭 세기
$\delta_i(t)$	사용자 i 채널에서의 열 잡음
$r_i(t)$	$D \log_2 \left(1 + \frac{\sum_{b \in \mathcal{B}} y_{b,i}(t) w_{b,i}(t)}{\omega_i(t) + \delta_i^2(t)} \right)$ 로, 빔 활성화 및 사용자 스케줄링에 따른 사용자 i 의 달성 가능한 쓰루풋. D 는 무선 접속망의 대역폭
$Q_p^c(t)$	클라우드 c 의 경로 p 의 프로세싱 큐
$W_p^e(t)$	클라우드에서 엣지 e 로 연결되는 백홀 링크의 경로 p 의 네트워크 큐
$Q_p^e(t)$	엣지 e 의 경로 p 의 프로세싱 큐
$Z_i^e(t)$	엣지 e 에서 사용자 i 로 연결되는 프론트홀 링크와 무선접속망의 네트워크 큐
$a_p(t) \in [0, \alpha_{i,max}(t)]$	사용자의 서비스 체이닝 옵션 별 트래픽 주입량
$\theta_p^c(t) \in \{0,1\}$	프로세스 스케줄링. 클라우드 c 에서 경로 p 에 대해 프로세스를 할지 결정한다. 1일 때 스케줄링 함을 뜻한다.
$\theta_p^e(t) \in \{0,1\}$	네트워크 스케줄링. 백홀에서 경로 p 에 대해 클라우드 c 로부터 엣지 e 로 전송을 할 지 결정한다. 1일 때 스케줄링 함을 뜻한다.
$\theta_p^e(t) \in \{0,1\}$	프로세스 스케줄링. 엣지 e 에서 경로 p 에 대해 프로세스를 할지 결정한다. 1일 때 스케줄링 함을 뜻한다
$w_{b,i}(t) \in \{0,1\}^M$	빔 활성화 및 사용자 스케줄링. 기지국 b 에서 사용자 i 에 빔 M 을 활성화하여 데이터를 전송할지 결정한다. 1일 때 스케줄링 함을 뜻한다.

[0054]

[0055]

본 실시예에서 제어하고자 하는 변수는 표 1에서 빨간색 변수와 같다. 결정되는 변수에 따른 트래픽의 이동을 살펴보면, 클라우드에서 사용자의 서비스 체이닝 경로를 결정하고, 해당 경로의 큐에 트래픽 $a_p(t)$ 를 넣어준다. 이 때, $\hat{p}$ 와 $\tilde{p}$ 경로는 $Q_p^c(t)$ 에, $\check{p}$ 경로는 $W_p^e(t)$ 에 트래픽이 쌓인다. 클라우드는 $\theta_p^c(t)$ 을 결정하고, 이 때 처리된 트래픽은 $W_p^e(t)$ 에 쌓인다. $W_p^e(t)$ 에 쌓인 트래픽은 클라우드의 $\theta_p^e(t)$ 결정에 따라서 $\hat{p}$ 경로는 $Z_i^e(t)$ 에, $\check{p}$ 와 $\tilde{p}$ 경로는 $Q_p^e(t)$ 에 쌓인다.

[0056]

$Q_p^e(t)$ 에 쌓인 트래픽은 엣지의 $\theta_p^e(t)$ 결정으로 처리되며 $Z_i^e(t)$ 에 쌓인다. 마지막으로 엣지별 무선접속망에서 기지국 b 에서 사용자 i 에 대해 $w_{b,i}(t)$ 를 결정해 프로세싱이 완료된 최종 데이터를 사용자에게 전송한다.

[0057]

아래에서는 문제 정의 및 알고리즘 유도 방식에 대해 설명한다.

[0058]

먼저, 문제의 목적 함수 (P1)를 다음 식과 같이 표현할 수 있다.

[0059]

[수학식 1]

[0060]

$$(P1): \max_{(\alpha, \theta, o, w)} \sum_{i \in I} U_i \left(\sum_{p \in P_i} \bar{a}_p \right)$$

[0061]

또한, 제약사항 (C1), (C2), (C3), (C4), (C5), 및 (C5)를 다음 식들과 같이 표현할 수 있다.

[0062]

[수학식 2]

[0063]

$$(C1): 0 \leq \sum_{p \in \mathcal{P}} a_p(t) \leq \alpha_{i,max}(t), \forall i \in I$$

[0064] (C2): $\sum_{p \in \mathcal{P}} \theta_p^c(t) \leq L^c(t) \text{ and } \sum_{p \in \mathcal{P}} \theta_p^e(t) \leq L^e(t), \forall e \in \mathcal{E}$

[0065] (C3): $\sum_{p \in \mathcal{P}} o_p^e(t) \leq N^e(t), \forall e \in \mathcal{E}$

[0066] (C4): $\sum_{i \in I} 1\{[w_{b,i}(t)]\} r_i(t) \leq R^b(t), \forall b \in \mathcal{B}^e, \forall e \in \mathcal{E}$

[0067] (C5): $\sum_{p \in \mathcal{P}} \bar{a}_p \geq a_{i,min}, \forall i \in I$

[0068] (C6): Mean-rate stability of all queues.

[0069] 클라우드-엣지-무선접속망으로 이루어진 계층적 5G 네트워크 구조에서 사용자들의 서비스 만족도를 최대화하기 위한 목적을 갖는다. 이 때, 목적함수 내의 $U_i(\cdot)$ 은 유틸리티 함수로, 사용자 i 의 평균 쓰루풋을 사용해 서비스 만족도의 정도를 나타낼 수 있다. 제약사항 (C1)은 단위시간 t 에서의 사용자 i 의 최대 처리율의 제약, (C2)는 클라우드와 엣지의 프로세싱 자원의 가용 용량 제약사항, (C3)와 (C4)는 각각 백홀과 프론트홀의 네트워크 자원의 가용 용량 제약사항, (C5)는 사용자 i 의 최소 평균 쓰루풋의 제약, (C6)은 모든 서비스는 유한한 시간 내에 처리되어야 한다는 제약사항이다. 본 문제를 최적화하기 위해 변수 사용자의 서비스 체이닝 옵션별 트래픽 주입량 $\mathbf{a}(t)$, 프로세스 스케줄링 $\boldsymbol{\theta}(t)$, 네트워크 스케줄링 $\mathbf{o}(t)$, 빔 활성화 및 사용자 스케줄링 $\mathbf{w}(t)$ 를 제어한다.

[0070] 문제 (P1)은 시간평균의 유틸리티 함수를 최대화하는 문제이다. 이 때 유틸리티 함수는 비선형 함수이기 때문에, 장시간적으로 풀어내야 한다. 따라서 문제를 매 단위시간마다 풀어갈 수 있도록 변형하기 위해서, 우선 $\mathbf{a}$ 에 대한 제약사항을 추가한다. 이후에 보조 변수의 추가와 jensen's inequality를 이용해 사용해 유틸리티 함수의 시간평균을 최대화하는 문제로 변형할 수 있다. 이는 매 단위시간마다 하위 문제를 풀어나가는 문제로, 본래의 문제보다 훨씬 다루기 쉽다. 따라서 최종적으로 본래 문제의 최적성을 유지한 채 원래의 장기적 관점 문제를 매 단위시간 마다 문제를 풀어나가는 알고리즘을 제공한다.

[0071] 도 4는 일 실시예에 따른 정의한 문제를 풀기 위한 알고리즘 유도의 수학적 원리를 설명하기 위한 도면이다.

[0072] 결과적으로, 최적으로 풀 (P3)(430)하는 알고리즘은 도 4에 요약되어 있을 뿐만 아니라 최적으로 풀 (P1)(410)도 해결하므로 이하에서는 문제 (P3)(430)를 만족시키기 위해 RACER 알고리즘을 제공한다.

[0074] 아래에서는 제안된 알고리즘을 설명한다.

[0075] 최종적으로 매 단위시간마다 하위 문제를 풀어나가는 문제로 알고리즘을 제작하였으며, 이 하위 문제는 각각의 물리적 계층(클라우드, 엣지, 무선접속망)마다의 문제로 분할하여 풀어낸다. 알고리즘은 매시간 사용자의 큐의 정보만을 이용해 각 계층을 독립적으로 제어한다. 알고리즘을 도출할 때 추가한 보조 변수의 제어를 위해 가상 큐 $F_i(t)$ 와 $G_i(t)$ 를 추가한다. 각 계층에서의 변수 제어는 다음 식과 같이 나타낼 수 있다.

[0076] 보조변수 α 의 제어를 다음 식과 같이 표현할 수 있다.

[0077] [수학식 3]

[0078]
$$\alpha_i(t) = \left[(U'_i)^{-1} \left(\frac{G_i(t)}{V} \right) \right]_{\alpha_{i,min}}^{\alpha_{i,max}}$$

[0079] 여기서, 유틸리티의 증가분과 가상 큐 G 의 안정화에 따라 위의 식을 통해 제어한다.

[0080] 사용자의 서비스 체이닝 옵션 별 트래픽 주입량의 제어를 다음 식과 같이 표현할 수 있다.

[0081] [수학식 4]

$$[0082] \begin{cases} a_{p^*}(t) = a_{i,max}(t), & \text{if } p^* = \operatorname{argmin}_{p \in \mathcal{P}_i} \psi_{i,p}(t) \text{ and } \psi_{i,p^*}(t) < 0, \\ a_p(t) = 0, & \text{otherwise.} \end{cases}$$

$$[0083] \psi_{i,p}(t) = \begin{cases} Q_p^c(t) - F_i(t) - G_i(t) & \text{if } p \in \{\tilde{p}_i, \tilde{p}_i\} \\ W_p^s(t) - F_i(t) - G_i(t) & \text{if } p = \tilde{p}_i \end{cases}$$

[0084] 여기서, 사용자 i 의 가능한 세가지 경로 중, 첫 번째 큐에 해당하는 큐의 $\psi_{i,p}(t)$ 값이 가장 적은 큐에 트래픽을 주입한다. 큐 길이가 가상 큐의 한계점보다 작을 때 트래픽을 주입할 수 있다.

[0085] 클라우드에서의 프로세스 스케줄링 $\theta^c(t)$ 을 다음 식과 같이 표현할 수 있다.

[0086] [수학식 5]

$$[0087] \min_{\theta^c(t)} \sum_{e \in \mathcal{E}} \sum_{p \in \mathcal{P} \cup \tilde{\mathcal{P}}} \theta_p^c(t) \phi_p^c(t) s^c$$

$$[0088] \text{subject to. } \sum_{p \in \tilde{\mathcal{P}}} \theta_p^c(t) \leq L^c(t)$$

$$[0089] \phi_p^c(t) = \frac{W_p^s(t) - Q_p^c(t)}{\rho_p^c}$$

[0090] 여기서, $\phi_p^c(t)$ 의 값이 음수인 값 중 가장 작은 값부터 프로세싱 자원을 한 개씩 할당 받아서 처리하게 된다. 이 때, 스케줄링될 수 있는 경로는 해당 단위시간의 프로세싱 자원 내에서만 처리될 수 있다. 또한, 백홀의 네트워킹 큐와 클라우드의 프로세싱 큐 길이의 차가 같을 경우에, ρ_p^c 가 작을수록 프로세싱 자원을 먼저 할당 받게 된다.

[0091] 백홀에서의 네트워크 스케줄링 $\theta^o(t)$ 을 다음 식과 같이 표현할 수 있다.

[0092] [수학식 6]

$$[0093] \min_{\theta^o(t)} \sum_{e \in \mathcal{E}} \sum_{p \in \mathcal{P}^e} \frac{R^e(t)}{N^e} o_p^e(t) \varphi_p^e(t)$$

$$[0094] \text{subject to. } \sum_{p \in \mathcal{P}^e} o_p^e(t) \leq N^e$$

$$[0095] \varphi_p^e(t) = \begin{cases} (Q_p^e(t) - W_p^s(t)) / \sigma_p^c & \text{if } p \in \tilde{\mathcal{P}}^e \cup \tilde{\mathcal{P}}^e \\ (Z_{i_p}^e(t) - W_p^s(t)) / \sigma_p^c & \text{if } p = \tilde{\mathcal{P}}^e \end{cases}$$

[0096] 여기서, $\varphi_p^e(t)$ 의 값이 음수인 값 중 가장 작은 값부터 네트워크 자원을 한 개씩 할당 받아 처리하게 된다. 이 때, 스케줄링될 수 있는 경로는 해당 단위시간의 네트워크 자원 내에서만 처리될 수 있다. 또한, $\varphi_p^e(t)$ 의 계산하는 분자의 값이 같을 경우에, σ_p^c 가 작을수록 네트워크 자원을 먼저 할당 받게 된다.

[0097] 엣지에서의 프로세스 스케줄링 $\theta^e(t)$ 을 다음 식과 같이 표현할 수 있다.

[0098] [수학식 7]

$$\min_{\theta^e(t)} \sum_{p \in \mathcal{P} \cup \tilde{\mathcal{P}}^e} \theta_p^e(t) \phi_p^e(t) s^e$$

[0099]

$$\text{subject to. } \sum_{p \in \mathcal{P}^e} \theta_p^e(t) \leq L^e(t)$$

[0100]

$$\phi_p^e(t) = \frac{Z_{i_p}^e(t) - Q_p^e(t)}{\rho_p^e}$$

[0101]

[0102] 여기서, $\phi_p^e(t)$ 의 값이 음수인 값 중 가장 작은 값부터 프로세싱 자원을 한 개씩 할당 받아서 처리하게 된다. 이 때, 스케줄링될 수 있는 경로는 해당 단위시간의 프로세싱 자원 내에서만 처리될 수 있다. 또한, 프론트홀 링크/무선접속망의 네트워크 큐와 엣지의 프로세싱 큐 길이의 차가 같을 경우에, ρ_p^e 가 작을수록 프로세싱 자원을 먼저 할당 받게 된다.

[0103] 무선접속망에서의 빔 활성화 및 사용자 스케줄링 $w^e(t)$ 을 다음 식과 같이 표현할 수 있다.

[0104] [수학식 8]

$$\max_{w^e(t)} \left\{ h(w^e(t)) = \sum_{i \in I} r_i(t) \frac{Z_i^e(t)}{\sigma_i^e} \right\}$$

[0105]

$$\sum_{i \in I} \{[w_{b,i}(t)]\} r_i(t) \leq R^b(t), \quad \forall b \in \mathcal{B}^e, \forall e \in \mathcal{E}$$

[0106]

[0107] 이를 풀어내기 위해서 greedy 기반의 알고리즘을 제안한다. 알고리즘은 다음과 같다.

[0108] [표 2]

빔 활성화 및 사용자 스케줄링

Every time slot t and for edge e ,

Input: $Z_i(t)$, $y_{b,i}(t)$, σ_i^e , ω_i , $\delta_i(t)$, $\forall i \in I^e$, $\forall b \in \mathcal{B}^e$

Output: $w^e(t) = \{w_{b,i}(t) | \forall b \in \mathcal{B}^e, \forall i \in I^e\}$

(Initialization)

1: Define vectors $w_{sel}^e = \mathbf{0}$ and $w_{cand}^e = \mathbf{1}$

2: Find one beam-user pair w that maximize $h(\cdot)$

3: Add w to w_{sel}^e

4: Remove w and other beam-user pairs related to the beam used by w from w_{cand}^e

(Iteration)

5: **while** $w_{cand}^e \neq \mathbf{0}$

6: Find one beam-user pair w from w_{cand}^e that maximizes increment of $h(\cdot)$ with offset w_{sel}^e

7: **if** w with w_{sel}^e satisfies (C4)

8: Do step 3 and 4

9: **else**

10: Remove w from w_{cand}^e

11: **end if**

12: **end while**

13: $w^e(t) = w_{sel}^e$

[0109]

[0111] 본 실시예에서는 제안한 알고리즘의 성능을 검증하기 위해 MATLAB을 이용해 시뮬레이션을 진행한다. (a) 모든 엣지에 동일한 사용자가 있는 경우, (b) 사용자가 다르게 분포되어 있는 경우, 두 가지의 상황에서 시뮬레이션을 진행한다.

[0112] 각 계층의 제어가 얼마나 중요한 사항인지 확인하기 위해서 모든 계층 중 하나의 계층만을 제안한 알고리즘이 아닌 다른 방식으로 동작하도록 변형해 제안한 알고리즘과 비교하였다.

[0113] 변형된 알고리즘은 다른 계층의 동작은 제안 알고리즘과 동일한 채, (1) RACER-SO: 사용자의 가능한 서비스 경로가 한 개만 갖도록 사전에 설정, (2) RACER-CRR: 클라우드의 프로세스 스케줄링이 라운드 로빈 형식으로 동작, (3) RACER-ERR: 엣지의 프로세스 스케줄링이 라운드 로빈 형식으로 동작, (4) RACER-BRR: 백홀의 네트워크 스케줄링이 라운드 로빈 형식으로 동작, (5) RACER-RRS: 무선접속망의 빔 활성화 및 사용자 스케줄링이 랜덤하게 동작, (6) RACER-RNCP: 무선접속망에서 CoMP가 고려되지 않은 상황에서 제안한 알고리즘과 동일하게 동작, (7) RACER-RLQ: 무선접속망에서 사용자의 네트워크 큐 길이가 긴 사용자가 스케줄링 되도록 동작, (8) RACER-RDR: 무선접속망에서 사용자의 신호가 큰 사용자가 스케줄링 되도록 동작으로, 총 8가지이다. 여기서, 본 실시예에서 제안한 알고리즘은 RACER이다.

[0114] 도 5 및 도 6은 제안된 알고리즘과 변형한 알고리즘의 평균 큐 길이 별 평균 쓰루풋을 나타낸다. 보다 구체적으로, 도 5는 일 실시예에 따른 균등한 사용자 분포에서의 제안된 알고리즘과 변형한 알고리즘의 비교를 나타내는 도면이다. 또한, 도 6은 일 실시예에 따른 불균등한 사용자 분포에서의 제안된 알고리즘과 변형한 알고리즘

의 비교를 나타내는 도면이다.

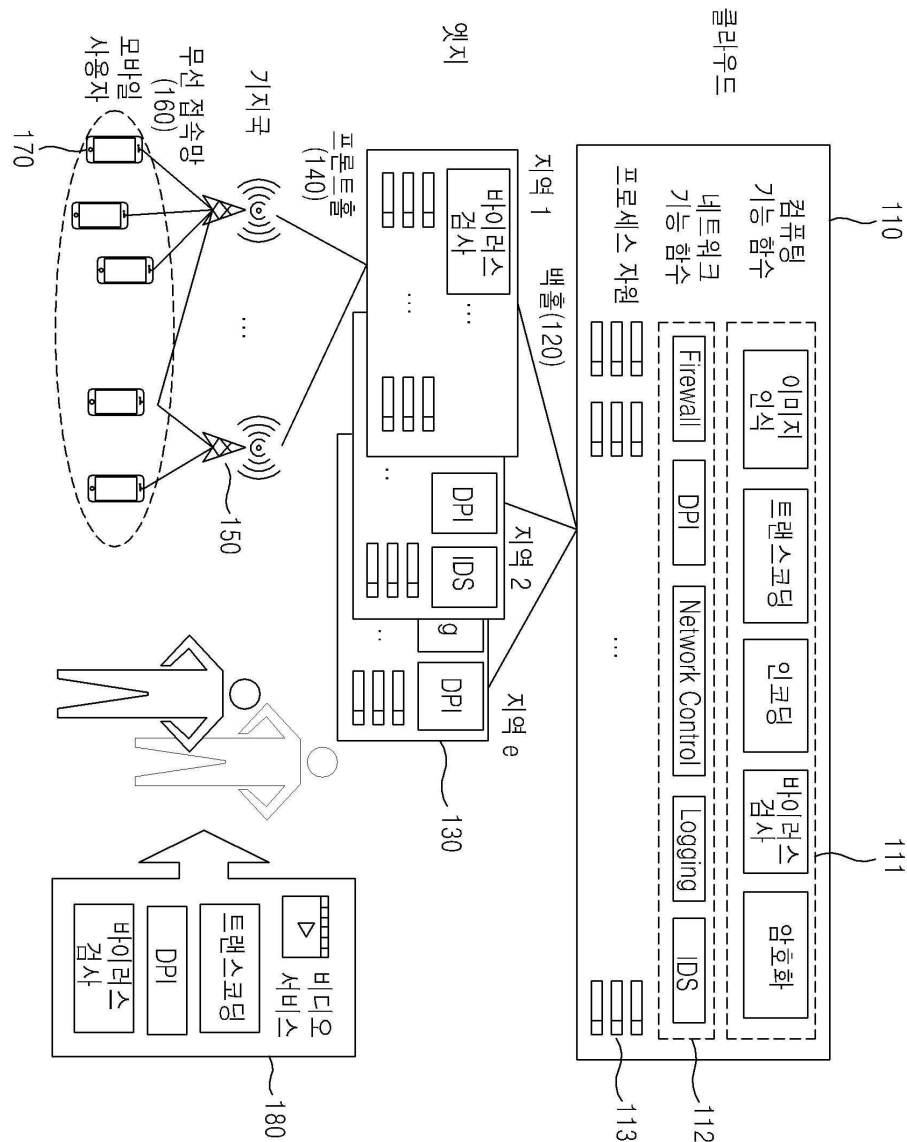
- [0115] 도 5를 참조하면, 모든 엣지에 동일한 수의 사용자가 랜덤하게 있는 경우의 결과를 나타내고, 도 6을 참조하면 각 엣지에 다른 수의 사용자가 랜덤하게 분포되어 있는 경우의 결과를 나타낸다.
- [0116] 다음 결과를 통해서 두 가지의 분석할 수 있다. (1) 제안 알고리즘은 적은 양의 지연(평균 큐 길이)을 허용함으로써 사용자 쓰루풋을 크게 증가시킬 수 있다. (2) 제안 알고리즘은 같은 평균 큐의 길이를 갖는 하에서 정적 알고리즘들보다 더 많은 트래픽을 처리할 수 있다. 이러한 성능 차이는 사용자가 균등하게 분포한 경우보다 불균등한 분포일 때 더 확연히 나타난다. 이는 제안한 알고리즘이 변화하는 환경에 적응적으로 자원을 할당할 수 있기 때문이다.
- [0117] 해당 시뮬레이션을 통해서 본 실시예에서 제안한 알고리즘에서 어느 한 계층이라도 효율적이게 동작하지 않으면 성능이 잘 나오지 않음을 확인했다. 이는 제안한 알고리즘이 클라우드-엣지-무선접속망의 동적 자원 할당을 효율적이게 통합 제어함을 의미한다.
- [0118] 이상과 같이, 실시예들은 클라우드-엣지-무선접속망으로 이루어진 5G 네트워크 인프라 구조에서 동적 자원 관리 기술을 제공할 수 있다. 특히, 사용자들의 서비스 지연을 감소시키며 서비스 만족도는 높일 수 있는 동적 자원 관리 방법을 제공할 수 있고, 네트워크와 프로세싱 자원을 함께 관리하는 방법을 제공할 수 있다. 또한, 네트워크 기능 함수와 컴퓨팅 기능 함수를 함께 지원하는 자원 관리 방법을 제공할 수 있다.
- [0120] 이상에서 설명된 장치는 하드웨어 구성요소, 소프트웨어 구성요소, 및/또는 하드웨어 구성요소 및 소프트웨어 구성요소의 조합으로 구현될 수 있다. 예를 들어, 실시예들에서 설명된 장치 및 구성요소는, 예를 들어, 프로세서, 컨트롤러, ALU(arithmetic logic unit), 디지털 신호 프로세서(digital signal processor), 마이크로컴퓨터, FPA(field programmable array), PLU(programmable logic unit), 마이크로프로세서, 또는 명령(instruction)을 실행하고 응답할 수 있는 다른 어떠한 장치와 같이, 하나 이상의 범용 컴퓨터 또는 특수 목적 컴퓨터를 이용하여 구현될 수 있다. 처리 장치는 운영 체제(OS) 및 상기 운영 체제 상에서 수행되는 하나 이상의 소프트웨어 애플리케이션을 수행할 수 있다. 또한, 처리 장치는 소프트웨어의 실행에 응답하여, 데이터를 접근, 저장, 조작, 처리 및 생성할 수도 있다. 이해의 편의를 위하여, 처리 장치는 하나가 사용되는 것으로 설명된 경우도 있지만, 해당 기술분야에서 통상의 지식을 가진 자는, 처리 장치가 복수 개의 처리 요소(processing element) 및/또는 복수 유형의 처리 요소를 포함할 수 있음을 알 수 있다. 예를 들어, 처리 장치는 복수 개의 프로세서 또는 하나의 프로세서 및 하나의 컨트롤러를 포함할 수 있다. 또한, 병렬 프로세서(parallel processor)와 같은, 다른 처리 구성(processing configuration)도 가능하다.
- [0121] 소프트웨어는 컴퓨터 프로그램(computer program), 코드(code), 명령(instruction), 또는 이들 중 하나 이상의 조합을 포함할 수 있으며, 원하는 대로 동작하도록 처리 장치를 구성하거나 독립적으로 또는 결합적으로(collectively) 처리 장치를 명령할 수 있다. 소프트웨어 및/또는 데이터는, 처리 장치에 의하여 해석되거나 처리 장치에 명령 또는 데이터를 제공하기 위하여, 어떤 유형의 기계, 구성요소(component), 물리적 장치, 가상 장치(virtual equipment), 컴퓨터 저장 매체 또는 장치에 구체화(embody)될 수 있다. 소프트웨어는 네트워크로 연결된 컴퓨터 시스템 상에 분산되어서, 분산된 방법으로 저장되거나 실행될 수도 있다. 소프트웨어 및 데이터는 하나 이상의 컴퓨터 판독 가능 기록 매체에 저장될 수 있다.
- [0122] 실시예에 따른 방법은 다양한 컴퓨터 수단을 통하여 수행될 수 있는 프로그램 명령 형태로 구현되어 컴퓨터 판독 가능 매체에 기록될 수 있다. 상기 컴퓨터 판독 가능 매체는 프로그램 명령, 데이터 파일, 데이터 구조 등을 단독으로 또는 조합하여 포함할 수 있다. 상기 매체에 기록되는 프로그램 명령은 실시예를 위하여 특별히 설계되고 구성된 것들이거나 컴퓨터 소프트웨어 당업자에게 공지되어 사용 가능한 것일 수도 있다. 컴퓨터 판독 가능 기록 매체의 예에는 하드 디스크, 플로피 디스크 및 자기 테이프와 같은 자기 매체(magnetic media), CD-ROM, DVD와 같은 광기록 매체(optical media), 플롭티컬 디스크(floptical disk)와 같은 자기-광 매체(magneto-optical media), 및 롬(ROM), 램(RAM), 플래시 메모리 등과 같은 프로그램 명령을 저장하고 수행하도록 특별히 구성된 하드웨어 장치가 포함된다. 프로그램 명령의 예에는 컴파일러에 의해 만들어지는 것과 같은 기계어 코드뿐만 아니라 인터프리터 등을 사용해서 컴퓨터에 의해서 실행될 수 있는 고급 언어 코드를 포함한다.
- [0123] 이상과 같이 실시예들이 비록 한정된 실시예와 도면에 의해 설명되었으나, 해당 기술분야에서 통상의 지식을 가진 자라면 상기의 기재로부터 다양한 수정 및 변형이 가능하다. 예를 들어, 설명된 기술들이 설명된 방법과 다른 순서로 수행되거나, 및/또는 설명된 시스템, 구조, 장치, 회로 등의 구성요소들이 설명된 방법과 다른 형태

로 결합 또는 조합되거나, 다른 구성요소 또는 균등물에 의하여 대치되거나 치환되더라도 적절한 결과가 달성될 수 있다.

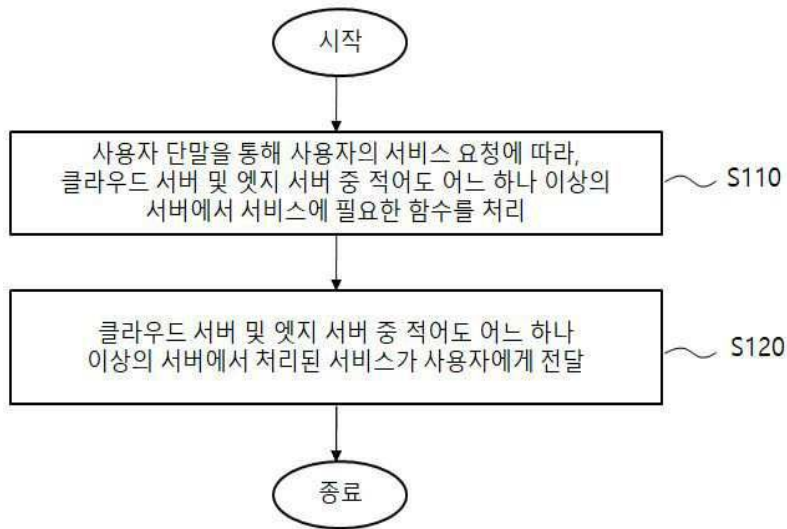
[0124] 그러므로, 다른 구현들, 다른 실시예들 및 특허청구범위와 균등한 것들도 후술하는 특허청구범위의 범위에 속한다.

도면

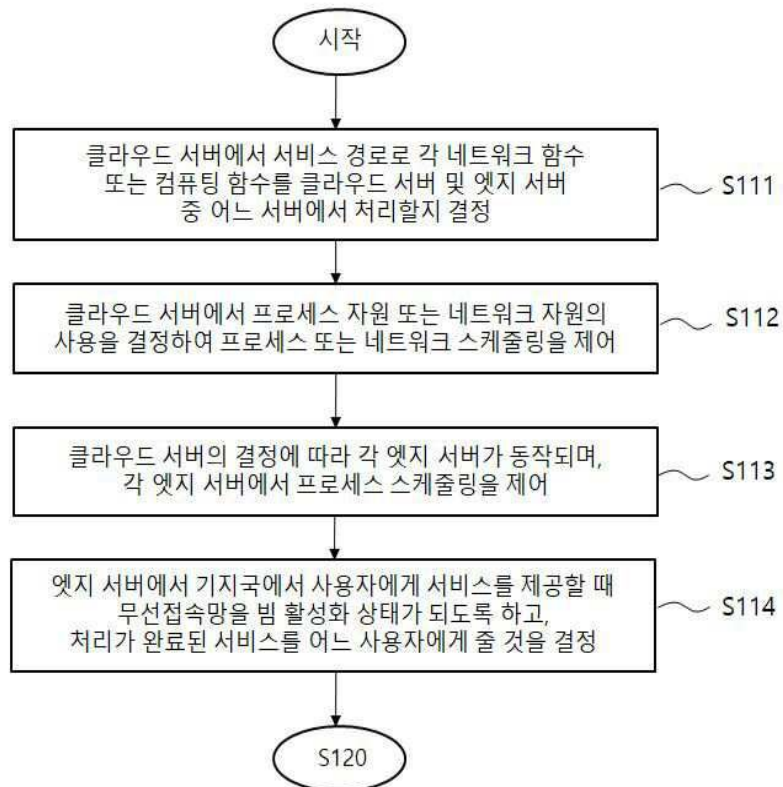
도면1



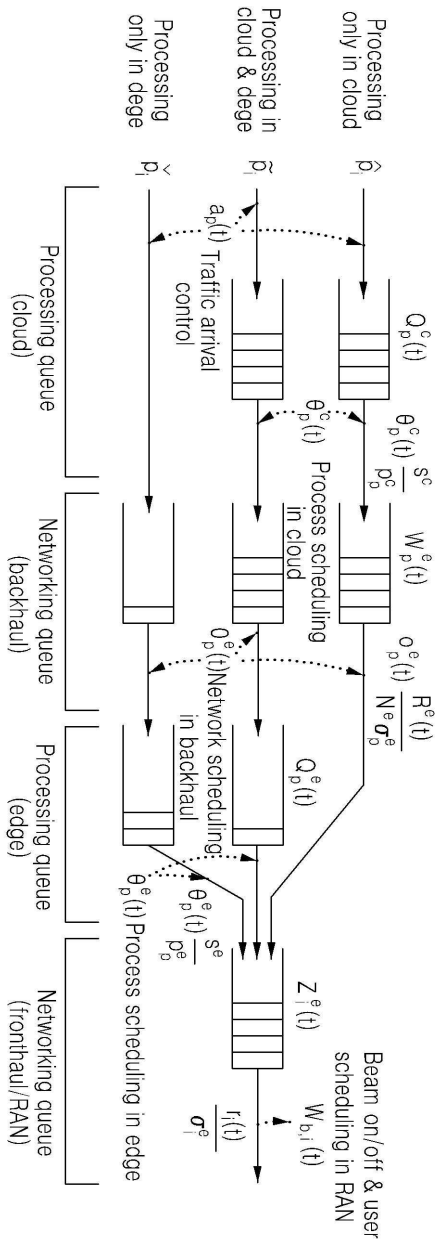
도면2a



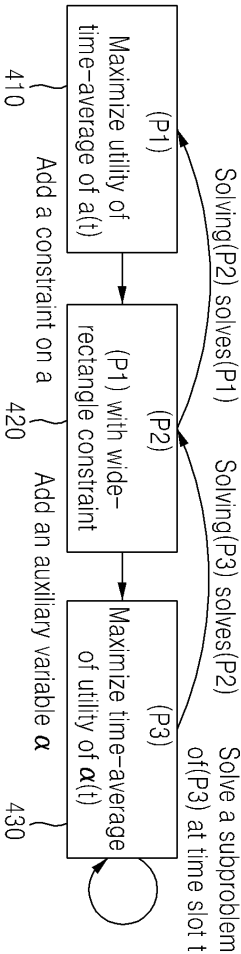
도면2b



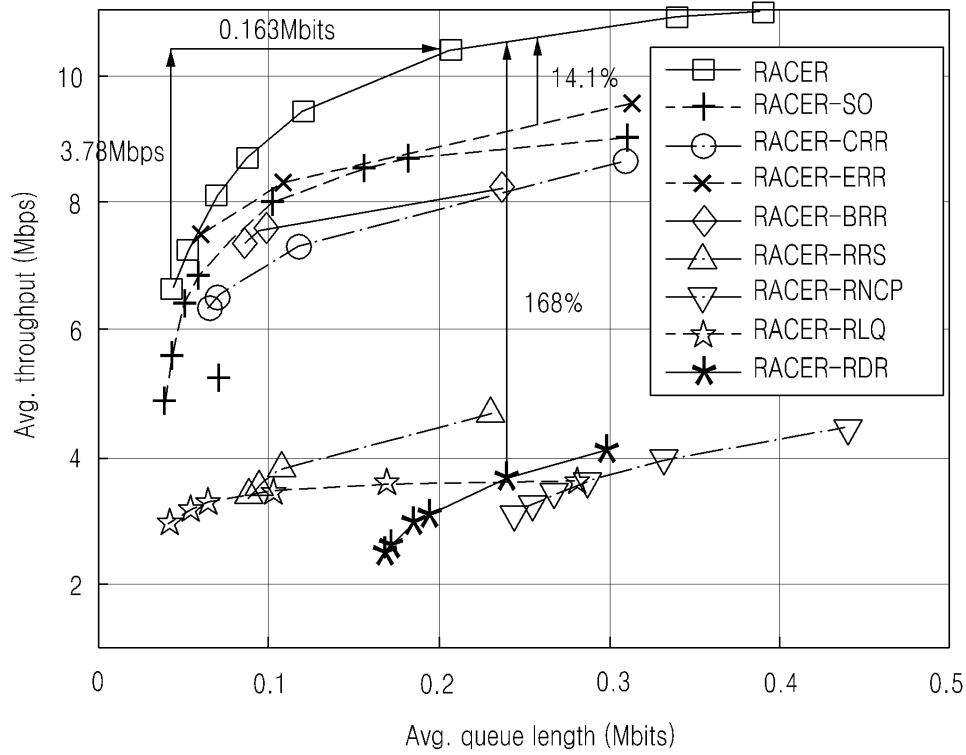
도면3



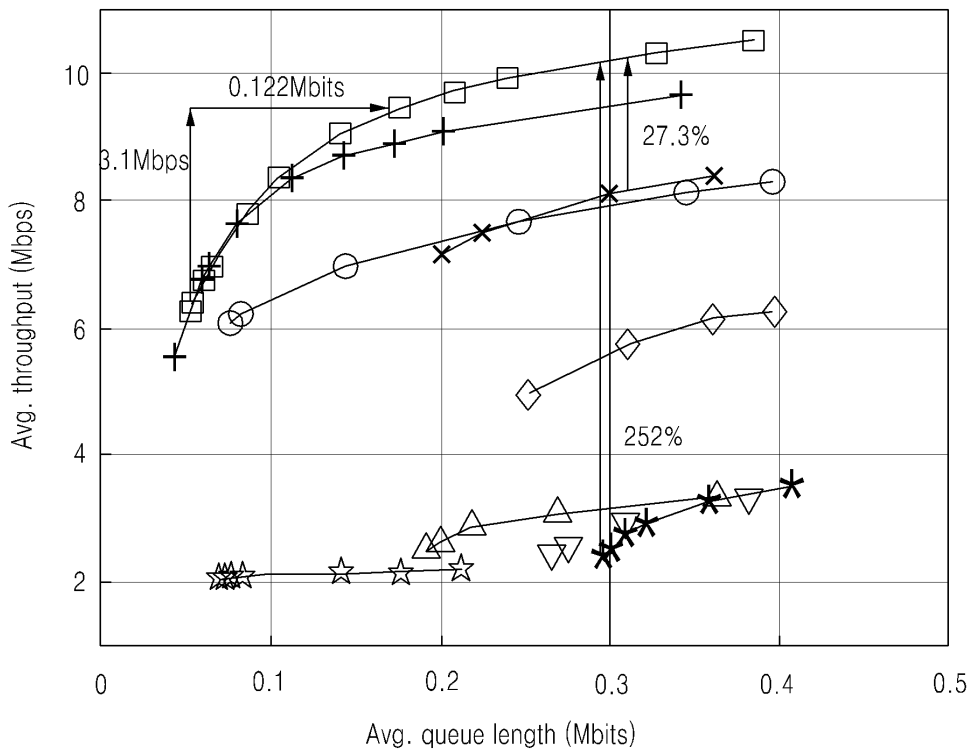
도면4



도면5



도면6



【심사관 직권보정사항】

【직권보정 1】

【보정항목】 청구범위

【보정세부항목】 청구항 5

【변경전】

계층적 5G 네트워크 구조에서 서비스 체이닝을 위한 동적 자원 할당 장치에 있어서,

사용자 단말을 통해 사용자의 서비스 요청에 따라, 서비스 경로로 각 네트워크 함수 또는 컴퓨팅 함수를 클라우드 서버 및 엣지 서버 중 어느 서버에서 처리할지 결정하고, 프로세스 자원 또는 네트워크 자원의 사용을 결정하여 프로세스 또는 네트워크 스케줄링을 제어하는 클라우드 서버; 및

상기 클라우드 서버의 결정에 따라 동작되며, 엣지 서버의 프로세스 스케줄링을 제어하고, 기지국에서 상기 사용자에게 서비스를 제공할 때 무선접속망을 빔 활성화 상태가 되도록 하고, 처리가 완료된 서비스를 어느 사용자에게 줄 것을 결정하는 엣지 서버

를 포함하고,

상기 클라우드 서버는,

상기 클라우드 서버에서 서비스 경로로 각 네트워크 함수 또는 컴퓨팅 함수를 상기 클라우드 서버 및 엣지 서버 중 어느 서버에서 처리할지 결정하는 서비스 체이닝 후, 상기 클라우드 서버에서 프로세스 자원 또는 네트워크 자원의 사용을 결정하여 프로세스 또는 네트워크 스케줄링을 제어하며,

상기 엣지 서버는,

클라우드 서버의 결정에 따라 각 상기 엣지 서버가 동작되며, 각 상기 엣지 서버에서 프로세스 스케줄링을 제어하고, 상기 엣지 서버에서 기지국에서 상기 사용자에게 서비스를 제공할 때 무선접속망을 빔 활성화 상태가 되도록 하고, 처리가 완료된 서비스를 어느 사용자에게 줄 것을 결정하며,

상기 클라우드 서버는 프로세스 스케줄링을 결정하며, 처리된 트래픽이 상기 클라우드 서버에서 상기 엣지 서버로 연결되는 상기 백홀의 경로의 네트워크 큐에 쌓이며, 쌓인 상기 트래픽은 상기 클라우드 서버의 네트워크 스케줄링 결정에 따라, 상기 클라우드 서버에서 서비스가 처리되는 경로는 상기 엣지 서버에서 사용자로 연결되는 상기 프론트홀 링크와 상기 무선접속망의 네트워크 큐에 쌓이고, 상기 엣지 서버에서 서비스가 처리되는 경로 및 상기 클라우드 서버와 상기 엣지 서버 모두에서 서비스가 처리되는 경로는 상기 엣지 서버의 경로의 프로세싱 큐에 쌓이며, 상기 엣지 서버의 경로의 프로세싱 큐에 쌓인 트래픽은 상기 엣지 서버의 프로세스 스케줄링 결정으로 처리되어, 상기 엣지 서버에서 사용자로 연결되는 상기 프론트홀 링크와 상기 무선접속망의 네트워크 큐에 쌓이는 것

을 특징으로 하는, 동적 자원 할당 장치.

【변경후】

계층적 5G 네트워크 구조에서 서비스 체이닝을 위한 동적 자원 할당 장치에 있어서,

사용자 단말을 통해 사용자의 서비스 요청에 따라, 서비스 경로로 각 네트워크 함수 또는 컴퓨팅 함수를 클라우드 서버 및 엣지 서버 중 어느 서버에서 처리할지 결정하고, 프로세스 자원 또는 네트워크 자원의 사용을 결정하여 프로세스 또는 네트워크 스케줄링을 제어하는 클라우드 서버; 및

상기 클라우드 서버의 결정에 따라 동작되며, 엣지 서버의 프로세스 스케줄링을 제어하고, 기지국에서 상기 사용자에게 서비스를 제공할 때 무선접속망을 빔 활성화 상태가 되도록 하고, 처리가 완료된 서비스를 어느 사용자에게 줄 것을 결정하는 엣지 서버

를 포함하고,

상기 클라우드 서버는,

상기 클라우드 서버에서 서비스 경로로 각 네트워크 함수 또는 컴퓨팅 함수를 상기 클라우드 서버 및 엣지 서버 중 어느 서버에서 처리할지 결정하는 서비스 체이닝 후, 상기 클라우드 서버에서 프로세스 자원 또는 네트워크 자원의 사용을 결정하여 프로세스 또는 네트워크 스케줄링을 제어하며,

상기 엣지 서버는,

클라우드 서버의 결정에 따라 각 상기 엣지 서버가 동작되며, 각 상기 엣지 서버에서 프로세스 스케줄링을 제어

하고, 상기 엣지 서버에서 기지국에서 상기 사용자에게 서비스를 제공할 때 무선접속망을 빔 활성화 상태가 되도록 하고, 처리가 완료된 서비스를 어느 사용자에게 줄 것을 결정하며,

상기 클라우드 서버는 프로세스 스케줄링을 결정하며, 처리된 트래픽이 상기 클라우드 서버에서 상기 엣지 서버로 연결되는 백홀의 경로의 네트워킹 큐에 쌓이며, 쌓인 상기 트래픽은 상기 클라우드 서버의 네트워크 스케줄링 결정에 따라, 상기 클라우드 서버에서 서비스가 처리되는 경로는 상기 엣지 서버에서 사용자로 연결되는 프론트홀 링크와 상기 무선접속망의 네트워킹 큐에 쌓이고, 상기 엣지 서버에서 서비스가 처리되는 경로 및 상기 클라우드 서버와 상기 엣지 서버 모두에서 서비스가 처리되는 경로는 상기 엣지 서버의 경로의 프로세싱 큐에 쌓이며, 상기 엣지 서버의 경로의 프로세싱 큐에 쌓인 트래픽은 상기 엣지 서버의 프로세스 스케줄링 결정으로 처리되어, 상기 엣지 서버에서 사용자로 연결되는 상기 프론트홀 링크와 상기 무선접속망의 네트워킹 큐에 쌓이는 것

을 특징으로 하는, 동적 자원 할당 장치.