(19) 대한민국특허청(KR)
(12) 공개특허공보(A)

(11) 공개번호 10-2025-0076314
(43) 공개일자 2025년05월29일

(51) 국제특허분류(Int. Cl.)
H04L 47/56 (2022.01) *H04L 47/625* (2022.01)
H04L 47/628 (2022.01) *H04L 47/783* (2022.01)
H04L 47/83 (2022.01) *H04W 72/542* (2023.01)
(52) CPC특허분류
H04L 47/56 (2022.05)
H04L 47/626 (2013.01)
(21) 출원번호 10-2023-0163850
(22) 출원일자 2023년11월22일
심사청구일자 2023년11월22일

(71) 출원인
서울대학교산학협력단
서울특별시 관악구 관악로 1 (신림동)
(72) 발명자
이경한
서울특별시 관악구 관악로 1 서울대학교 301동 1006호
진성현
서울시 강남구 선릉로 221, 도곡렉슬아파트 409동 604호
김세래
서울특별시 마포구 성미산로 29안길 12 2층
(74) 대리인
박정환

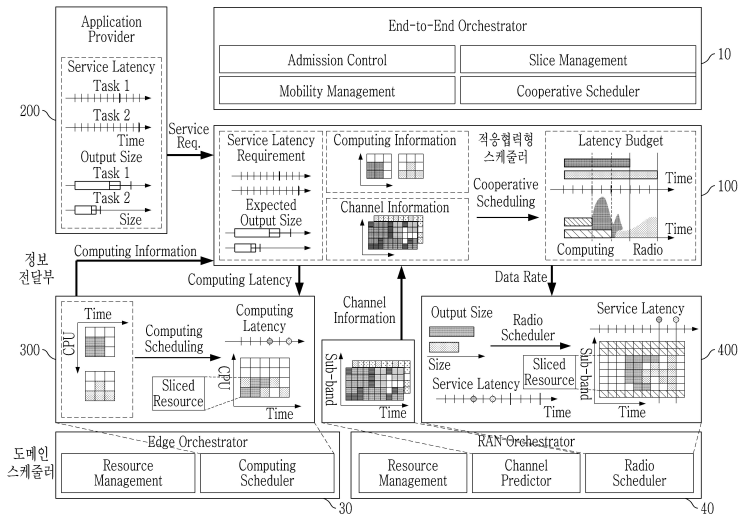
전체 청구항 수 : 총 12 항

(54) 발명의 명칭 **응용 수준 성능 보장을 위한 무선 및 컴퓨팅 자원의 협력 스케줄링 방법**

(57) 요 약

응용 수준 성능 보장을 위한 협력 스케줄링 방법은, 응용서비스의 업무 별 요구사항에 대한 제1 정보를 수신하는 단계; 제1 도메인 스케줄러로부터 상기 응용서비스의 컴퓨팅과 관련된 제2 정보를 수신하는 단계; 제2 도메인 스케줄러로부터 사용자 별 채널상태와 관련된 정보 및 네트워크 트래픽과 관련된 정보 중 적어도 어느 하나를 포함하는 제3 정보를 수신하는 단계; 상기 제1 정보, 상기 제2 정보 및 상기 제3 정보에 기초하여 상기 업무 별로 컴퓨팅 자원과 네트워크 자원에 대한 가치 또는 비용을 각각 산출하는 단계; 및 상기 산출된 가치 또는 비용에 기초하여 상기 응용서비스의 전체 지연시간을 컴퓨팅 지연시간과 네트워크 지연시간에 분배하여 할당하는 단계를 포함할 수 있다.

대 표 도



(52) CPC특허분류
H04L 47/628 (2022.05)
H04L 47/783 (2022.05)
H04L 47/83 (2022.05)
H04W 72/542 (2023.01)

이 발명을 지원한 국가연구개발사업

과제고유번호	1711193202
과제번호	2022-0-00420-002
부처명	과학기술정보통신부
과제관리(전문)기관명	정보통신기획평가원
연구사업명	6G핵심기술개발
연구과제명	6G 종단간 초정밀 네트워킹을 위한 핵심기술 개발
기 여 율	1/1
과제수행기관명	서울대학교 산학협력단
연구기간	2023.01.01 ~ 2023.12.31

명 세 서

청구범위

청구항 1

응용 수준 성능 보장을 위한 무선 및 컴퓨팅 자원의 협력 스케줄링 방법에 있어서,

응용서비스의 업무 별 요구사항에 대한 제1 정보를 수신하는 단계;

제1 도메인 스케줄러로부터 상기 응용서비스의 컴퓨팅과 관련된 제2 정보를 수신하는 단계;

제2 도메인 스케줄러로부터 사용자 별 채널상태와 관련된 정보 및 네트워크 트래픽과 관련된 정보 중 적어도 어느 하나를 포함하는 제3 정보를 수신하는 단계;

상기 제1 정보, 상기 제2 정보 및 상기 제3 정보에 기초하여 상기 업무 별로 컴퓨팅 자원과 네트워크 자원에 대한 가치 또는 비용을 각각 산출하는 단계; 및

상기 산출된 가치 또는 비용에 기초하여 상기 응용서비스의 전체 지연시간을 컴퓨팅 지연시간과 네트워크 지연시간에 분배하여 할당하는 단계를 포함하는, 협력 스케줄링 방법.

청구항 2

제 1항에 있어서,

상기 할당된 컴퓨팅 지연시간을 상기 제1 도메인 스케줄러로 전송하는 단계; 및

상기 할당된 네트워크 지연시간에 해당하는 데이터 전송량 혹은 데이터 전송률을 상기 제 2 도메인 스케줄러에게 전송하는 단계를 더 포함하는, 협력 스케줄링 방법.

청구항 3

제 1항에 있어서,

상기 제 1 정보는 상기 응용서비스의 업무 별 요구 지연시간 및 상기 업무 별 예상 출력데이터 크기 중 적어도 어느 하나를 포함하는, 협력 스케줄링 방법.

청구항 4

제 1항에 있어서,

상기 컴퓨팅 자원의 가치 혹은 비용을 산출하는 단계는,

연결된 서버의 종류, 상기 업무 별 종류 및 시간에 따라 변하는 트래픽 양 중 적어도 어느 하나에 더 기초하여 산출하는, 협력 스케줄링 방법.

청구항 5

제 1항에 있어서,

상기 네트워크 자원의 가치 혹은 비용을 산출하는 단계는,

상기 사용자 별 채널 추정의 정확도, 상기 사용자의 채널 상황 및 시간에 따라 변하는 트래픽 양 중 적어도 어느 하나에 더 기초하여 산출하는, 협력 스케줄링 방법.

청구항 6

제 1항에 있어서,

상기 제2 정보는 컴퓨팅 자원 요구량, 메모리 사용량 및 응용서버의 트래픽 양 중 적어도 어느 하나를 포함하는, 협력 스케줄링 방법.

청구항 7

제 1항에 있어서,

상기 제3 정보는 예측되는 채널 정보를 더 포함하는, 협력 스케줄링 방법.

청구항 8

응용 수준 성능 보장을 위한 무선 및 컴퓨팅 자원의 협력 스케줄링을 수행하는 장치에 있어서,

응용서비스의 업무 별 요구사항에 대한 제1 정보를 수신하고,

제1 도메인 스케줄러로부터 상기 응용서비스의 컴퓨팅과 관련된 제2 정보를 수신하며,

제2 도메인 스케줄러로부터 사용자 별 채널상태와 관련된 정보 및 네트워크 트래픽과 관련된 정보 중 적어도 어느 하나를 포함하는 제3 정보를 수신하는 통신부; 및

상기 제1 정보, 상기 제2 정보 및 상기 제3 정보에 기초하여 상기 업무 별로 컴퓨팅 자원과 네트워크 자원에 대한 가치 또는 비용을 각각 산출하고,

상기 산출된 가치 또는 비용에 기초하여 상기 응용서비스의 전체 지연시간을 컴퓨팅 지연시간과 네트워크 지연시간에 분배하여 할당하는 프로세서를 포함하는, 협력 스케줄링 장치.

청구항 9

제 8항에 있어서,

상기 통신부는,

상기 할당된 컴퓨팅 지연시간을 상기 제1 도메인 스케줄러로 전송하고,

상기 할당된 네트워크 지연시간에 해당하는 데이터 전송량 혹은 데이터 전송률을 상기 제 2 도메인 스케줄러에게 전송하는, 협력 스케줄링 장치.

청구항 10

제 8항에 있어서,

상기 프로세서는,

상기 컴퓨팅 자원의 가치 혹은 비용을 산출하는 경우, 연결된 서버의 종류, 상기 업무 별 종류 및 시간에 따라 변하는 트래픽 양 중 적어도 어느 하나에 더 기초하여 산출하는, 협력 스케줄링 장치.

청구항 11

제 8항에 있어서,

상기 프로세서는, 상기 네트워크 자원의 가치 혹은 비용을 산출하는 경우, 상기 사용자 별 채널 추정의 정확도, 상기 사용자의 채널 상황 및 시간에 따라 변하는 트래픽 양 중 적어도 어느 하나에 더 기초하여 산출하는, 협력 스케줄링 장치.

청구항 12

제 1항 내지 제 7항 중 어느 한 항에 따른 협력 스케줄링 방법을 컴퓨터에서 실행시키기 위한 프로그램을 기록한 컴퓨터-판독가능한 기록매체.

발명의 설명

기 술 분 야

[0001] 본 발명은 응용 수준 성능 보장을 위한 무선 및 컴퓨팅 자원의 협력 스케줄링 방법 및 장치에 대한 관한 것이다.

배 경 기 술

[0002] 5G의 세 가지 주요 요구 서비스 사항으로는 (1) 개선된 모바일 광대역 (Enhanced Mobile Broadband, eMBB) 서비스, (2) 다량의 머신 타입 통신 (massive Machine Type Communication, mMTC) 서비스 및 (3) 초-신뢰 및 저지연 통신 (Ultra-reliable and Low Latency Communications, URLLC) 서비스가 있다. 일부 사용 예(Use Case)는 최적화를 위해 다수의 영역들이 요구될 수 있고, 다른 사용 예는 단지 하나의 핵심 성능 지표에만 포커싱될 수 있다. 5G는 이러한 다양한 사용 예들을 유연하고 신뢰할 수 있는 방법으로 지원한다.

[0003] 5G는 또한 클라우드의 컴퓨팅 성능을 이용한 원격 업무에도 사용되며, 촉각 인터페이스가 사용될 때 우수한 사용자 경험을 유지하도록 훨씬 더 낮은 종단간(end-to-end) 지연을 요구한다. 예를 들어, 클라우드 게임 및 비디오 스트리밍은 모바일 광대역 능력에 대한 요구를 증가시키는 또 다른 핵심 요소이다. 증강 현실은 매우 낮은 지연과 순간적인 데이터 양을 필요로 한다.

[0004] 5G, 6G 통신에서의 URLLC는 주요 인프라의 원격 제어 및 자체-구동 차량(self-driving vehicle), 종단간 초정밀 네트워킹과 같은 초 신뢰/저지연 링크를 통해 산업을 변화시킬 새로운 서비스를 포함한다. 신뢰성과 지연의 수준은 6G 종단간 초정밀 네트워킹, 스마트 그리드 제어, 산업 자동화, 로봇 공학, 드론 제어 및 조정에 필수적이다. 이러한 이유로 6G 등 차세대 통신 네트워크에서는 초신뢰, 저지연, 시민감 데이터들에 대한 성능 혹은 요구지연 시간 보장을 위해 많은 자원관리 방식들이 연구되고 있다.

[0005] 몰입형 경험(immersive experiences)은 차세대 네트워크가 제공해야 할 핵심 기능이다. 가상 세계와 실제 세계의 원활한 통합을 위해서는 네트워크에 낮고 일관된 지연시간(latency)가 요구된다. 종래의 전통적인 응용서비스에서는 업무(task) 종류 하나가 서비스를 구성하였고, 각 도메인에서의 부담(burden)은 과하지 않았다. 그러나, 차세대 네트워크에서 제공할 서비스의 일 예로서 몰입형 서비스는 이종 작업(Heterogeneous task)들 서비스가 구성될 수 있다. 이미지 업무(예, 렌더링, 인코딩 등), 위치 지정 업무(예, 머리 위치 지정, 이동성 추정 등) 및 상호작용 업무(예, 인간 대 인간, 인간 대 인프라, 등) 등이 이기종 업무로 구성될 수 있다. 이러한 몰입형 서비스들은 과도한 컴퓨팅과 큰 출력 크기로 인해 컴퓨팅 도메인과 네트워크(무선) 도메인 모두에 과도하고 동적인 부담을 준다. 따라서, 차세대 네트워크에서는 몰입형 컴퓨팅 도메인과 네트워크 도메인 사이에서의 부담을 조절할 필요가 있으나 아직까지 구체적은 방법이 제시되지 않은 실정이다.

선행기술문헌

특허문헌

[0006] (특허문헌 0001) 한국 공개특허 공보 10-2021-0149576호

발명의 내용

해결하려는 과제

[0007] 본 발명에서 이루고자 하는 기술적 과제는 응용 수준 성능 보장을 위한 무선 및 컴퓨팅 자원의 협력 스케줄링 방법을 제공하는 데 있다.

[0008] 본 발명에서 이루고자 하는 다른 기술적 과제는 응용 수준 성능 보장을 위한 무선 및 컴퓨팅 자원의 협력 스케줄링 장치를 제공하는 데 있다.

[0009] 본 발명에서 이루고자 하는 다른 기술적 과제는 응용 수준 성능 보장을 위한 무선 및 컴퓨팅 자원의 협력 스케줄링 방법을 컴퓨터에서 실행시키기 위한 프로그램을 기록한 컴퓨터-판독가능한 기록매체를 제공하는 데 있다.

[0010] 본 발명에서 이루고자 하는 기술적 과제들은 상기 기술적 과제로 제한되지 않으며, 언급하지 않은 또 다른 기술적 과제들은 아래의 기재로부터 본 발명이 속하는 기술분야에서 통상의 지식을 가진 자에게 명확하게 이해될 수 있을 것이다.

과제의 해결 수단

[0011] 상기의 기술적 과제를 달성하기 위한, 본 발명에 따른 응용 수준 성능 보장을 위한 무선 및 컴퓨팅 자원의 협력 스케줄링 방법은, 응용서비스의 업무 별 요구사항에 대한 제1 정보를 수신하는 단계; 제1 도메인 스케줄러로부

터 상기 응용서비스의 컴퓨팅과 관련된 제2 정보를 수신하는 단계; 제2 도메인 스케줄러로부터 사용자 별 채널 상태와 관련된 정보 및 네트워크 트래픽과 관련된 정보 중 적어도 어느 하나를 포함하는 제3 정보를 수신하는 단계; 상기 제1 정보, 상기 제2 정보 및 상기 제3 정보에 기초하여 상기 업무 별로 컴퓨팅 자원과 네트워크 자원에 대한 가치 또는 비용을 각각 산출하는 단계; 및 상기 산출된 가치 또는 비용에 기초하여 상기 응용서비스의 전체 지연시간을 컴퓨팅 지연시간과 네트워크 지연시간에 분배하여 할당하는 단계를 포함할 수 있다.

[0012] 상기 방법은, 상기 할당된 컴퓨팅 지연시간을 상기 제1 도메인 스케줄러로 전송하는 단계; 및 상기 할당된 네트워크 지연시간에 해당하는 데이터 전송량 혹은 데이터 전송률을 상기 제 2 도메인 스케줄러에게 전송하는 단계를 더 포함할 수 있다.

[0013] 상기 제 1 정보는 상기 응용서비스의 업무 별 요구 지연시간 및 상기 업무 별 예상 출력데이터 크기 중 적어도 어느 하나를 포함할 수 있다.

[0014] 상기 컴퓨팅 자원의 가치 혹은 비용을 산출하는 단계는, 연결된 서버의 종류, 상기 업무 별 종류 및 시간에 따라 변하는 트래픽 양 중 적어도 어느 하나에 더 기초하여 산출하는 단계를 포함할 수 있다.

[0015] 상기 네트워크 자원의 가치 혹은 비용을 산출하는 단계는, 상기 사용자 별 채널 추정의 정확도, 상기 사용자의 채널 상황 및 시간에 따라 변하는 트래픽 양 중 적어도 어느 하나에 더 기초하여 산출하는 단계를 포함할 수 있다.

[0016] 상기 제2 정보는 컴퓨팅 자원 요구량, 메모리 사용량 및 응용서버의 트래픽 양 중 적어도 어느 하나를 포함할 수 있다. 상기 제3 정보는 예측되는 채널 정보를 더 포함할 수 있다.

[0017] 상기의 다른 기술적 과제를 달성하기 위한, 본 발명에 따른 응용 수준 성능 보장을 위한 무선 및 컴퓨팅 자원의 협력 스케줄링을 수행하는 장치는, 응용서비스의 업무 별 요구사항에 대한 제1 정보를 수신하고, 제1 도메인 스케줄러로부터 상기 응용서비스의 컴퓨팅과 관련된 제2 정보를 수신하며, 제2 도메인 스케줄러로부터 사용자 별 채널상태와 관련된 정보 및 네트워크 트래픽과 관련된 정보 중 적어도 어느 하나를 포함하는 제3 정보를 수신하는 통신부; 및 상기 제1 정보, 상기 제2 정보 및 상기 제3 정보에 기초하여 상기 업무 별로 컴퓨팅 자원과 네트워크 자원에 대한 가치 또는 비용을 각각 산출하고, 상기 산출된 가치 또는 비용에 기초하여 상기 응용서비스의 전체 지연시간을 컴퓨팅 지연시간과 네트워크 지연시간에 분배하여 할당하는 프로세서를 포함할 수 있다.

[0018] 상기 통신부는, 상기 할당된 컴퓨팅 지연시간을 상기 제1 도메인 스케줄러로 전송하고, 상기 할당된 네트워크 지연시간에 해당하는 데이터 전송량 혹은 데이터 전송률을 상기 제 2 도메인 스케줄러에게 전송할 수 있다.

[0019] 상기 프로세서는, 상기 컴퓨팅 자원의 가치 혹은 비용을 산출하는 경우, 연결된 서버의 종류, 상기 업무 별 종류 및 시간에 따라 변하는 트래픽 양 중 적어도 어느 하나에 더 기초하여 산출할 수 있다.

[0020] 상기 프로세서는, 상기 네트워크 자원의 가치 혹은 비용을 산출하는 경우, 상기 사용자 별 채널 추정의 정확도, 상기 사용자의 채널 상황 및 시간에 따라 변하는 트래픽 양 중 적어도 어느 하나에 더 기초하여 산출할 수 있다.

발명의 효과

[0021] 본 발명에 따른 협력 스케줄링 방법은 두 도메인 간을 적응적이며 협력적으로 스케줄링하여 차세대 네트워크에서 제공할 몰입형 서비스 등의 응용 서비스의 전체 지연시간을 만족하도록 함으로써 응용 수준에서의 성능 보장이 가능하게 한다.

[0022] 본 발명에서 얻은 수 있는 효과는 이상에서 언급한 효과들로 제한되지 않으며, 언급하지 않은 또 다른 효과들은 아래의 기재로부터 본 발명이 속하는 기술분야에서 통상의 지식을 가진 자에게 명확하게 이해될 수 있을 것이다.

도면의 간단한 설명

[0023] 본 발명에 관한 이해를 돕기 위해 상세한 설명의 일부로 포함되는, 첨부 도면은 본 발명에 대한 실시예를 제공하고, 상세한 설명과 함께 본 발명의 기술적 사상을 설명한다.

도 1은 5G NR의 전체적인 시스템 구조를 도시한 도면이다.

도 2는 Open RAN(O-RAN) 시스템에서의 논리적 아키텍처(Logical Architecture)를 도시한 도면이다.

도 3은 서비스/응용 별 레이턴시(latency)에 대해 설명하기 위한 도면이다.

도 4는 5G, 6G 등의 프로토콜 스택-사용자 평면을 예시한 도면이다.

도 5는 6G 종단간 네트워크 구조를 예시한 도면이다.

도 6은 응용-수준에서의 서비스 지연시간(latency)를 설명하기 위한 도면이다.

도 7은 차세대 네트워크에서의 요구사항 중 하나인 지연시간을 설명하기 위한 도면이다.

도 8은 차세대 네트워크에서 제공할 서비스들의 업무 부담을 예시한 도면이다.

도 9는 컴퓨팅 오프로딩을 위한 기준 RPC 업무흐름(workflow)를 도시한 도면이다.

도 10은 본 발명에 따른 응용 수준 성능 보장을 위한 네트워크(혹은 무선) 및 컴퓨팅 자원의 협력 스케줄링을 수행하기 위한 장치의 구성을 블록으로 나타낸 도면이다.

도 11은 본 발명에 따른 응용 수준 성능 보장을 위한 네트워크(혹은 무선) 및 컴퓨팅 자원의 협력 스케줄링 방법을 설명하기 위한 예시적 도면이다.

도 12는 본 발명에 따른 응용 수준 성능 보장을 위한 네트워크(혹은 무선) 및 컴퓨팅 자원의 협력 스케줄링을 위한 정보 및 결과를 교환하는 인터페이스를 설명하는 도면이다.

발명을 실시하기 위한 구체적인 내용

[0024] 이하, 본 발명에 따른 바람직한 실시 형태를 첨부된 도면을 참조하여 상세하게 설명한다. 첨부된 도면과 함께 이하에 개시될 상세한 설명은 본 발명의 예시적인 실시형태를 설명하고자 하는 것이며, 본 발명이 실시될 수 있는 유일한 실시형태를 나타내고자 하는 것이 아니다. 이하의 상세한 설명은 본 발명의 완전한 이해를 제공하기 위해서 구체적 세부사항을 포함한다. 그러나, 당업자는 본 발명이 이러한 구체적 세부사항 없이도 실시될 수 있음을 안다. 예를 들어, 이하의 상세한 설명은 설명의 편의를 위해 이동통신 시스템이 3GPP 5G NR, 차세대 통신 시스템인 6G, Open RAN (O-RAN) 시스템인 경우를 가정하여 구체적으로 설명하나, 3GPP 5G NR, 차세대 통신 시스템인 6G, Open RAN (O-RAN)의 특유한 사항을 제외하고는 다른 임의의 이동통신 시스템에도 적용 가능하다.

[0025] 몇몇 경우, 본 발명의 개념이 모호해지는 것을 피하기 위하여 공지의 구조 및 장치는 생략되거나, 각 구조 및 장치의 핵심기능을 중심으로 한 블록도 형식으로 도시될 수 있다. 또한, 본 명세서 전체에서 동일한 구성요소에 대해서는 동일한 도면 부호를 사용하여 설명한다.

[0026] 본 발명은 다양한 변경을 가할 수 있고 여러 가지 실시예를 가질 수 있는 바, 특정 실시예들을 도면에 예시하고 상세하게 설명하고자 한다. 그러나 이는 본 발명을 특정한 실시 형태에 대해 한정하려는 것이 아니며, 본 발명의 사상 및 기술 범위에 포함되는 모든 변경, 균등물 내지 대체물을 포함하는 것으로 이해되어야 한다.

[0027] 어떤 구성요소가 다른 구성요소에 "연결되어" 있다거나 "접속되어" 있다고 언급된 때에는, 그 다른 구성요소에 직접적으로 연결되어 있거나 또는 접속되어 있을 수도 있지만, 중간에 다른 구성요소가 존재할 수도 있다고 이해되어야 할 것이다. 반면에, 어떤 구성요소가 다른 구성요소에 "직접 연결되어" 있다거나 "직접 접속되어" 있다고 언급된 때에는, 중간에 다른 구성요소가 존재하지 않는 것으로 이해되어야 할 것이다.

[0028] 제1, 제2 등의 용어는 다양한 구성요소들을 설명하는데 사용될 수 있지만, 상기 구성요소들은 상기 용어들에 의해 한정되어서는 안 된다. 상기 용어들은 하나의 구성요소를 다른 구성요소로부터 구별하는 목적으로만 사용된다.

[0029] 본 명세서에서 사용한 용어는 단지 특정한 실시예를 설명하기 위해 사용된 것으로, 본 발명을 한정하려는 의도가 아니다. 단수의 표현은 문맥상 명백하게 다르게 뜻하지 않는 한, 복수의 표현을 포함한다. 본 명세서에서, "포함하다" 또는 "가지다" 등의 용어는 명세서상에 기재된 특징, 숫자, 단계, 동작, 구성요소, 부품 또는 이들을 조합한 것이 존재함을 지정하려는 것이지, 하나 또는 그 이상의 다른 특징들이나 숫자, 단계, 동작, 구성요소, 부품 또는 이들을 조합한 것들의 존재 또는 부가 가능성을 미리 배제하지 않는 것으로 이해되어야 한다.

[0030] 또한, 각 도면을 참조하여 설명하는 실시예의 구성 요소가 해당 실시예에만 제한적으로 적용되는 것은 아니며, 본 발명의 기술적 사상이 유지되는 범위 내에서 다른 실시예에 포함되도록 구현될 수 있으며, 또한 별도의 설명이 생략될지라도 복수의 실시예가 통합된 하나의 실시예로 다시 구현될 수도 있음은 당연하다.

[0031] 또한, 명세서에 기재된 "…부", "…유닛", "…모듈", "…기" 등의 용어는 적어도 하나의 기능이나 동작을 처리

하는 단위를 의미하며, 이는 하드웨어나 소프트웨어 또는 하드웨어 및 소프트웨어의 결합으로 구현될 수 있다.

[0032] 아울러, 이하의 설명에 있어서 단말은 기지국/네트워크로부터 하향링크(Downlink)를 통해 정보를 수신할 수 있으며, 단말은 또한 상향링크(Uplink)를 통해 정보를 기지국/네트워크로 전송할 수 있다. 단말이 전송 또는 수신하는 정보로는 데이터 및 다양한 제어 정보가 있으며, 단말이 전송 또는 수신하는 정보의 종류 용도에 따라 다양한 물리 채널이 존재한다.

[0033] 5G NR(new radio)은 usage scenario에 따라 eMBB(enhanced Mobile Broadband), mMTC(massive Machine Type Communications), URLLC(Ultra-Reliable and Low Latency Communications), V2X(vehicle-to-everything)을 정의한다. 그리고, 5G NR 규격(standard)는 NR 시스템과 LTE 시스템 사이의 공존(co-existence)에 따라 standalone(SA)와 non-standalone(NSA)으로 구분한다. 그리고, 5G NR은 다양한 서브캐리어 간격(subcarrier spacing)을 지원하며, 하향링크에서 CP-OFDM을, 상향링크에서 CP-OFDM 및 DFT-s-OFDM(SC-OFDM)을 지원한다.

[0034] 본 발명의 실시 예들은 무선 접속 시스템들인 5G NR, 6G, Open RAN 등의 개시된 표준 문서들에 의해 뒷받침될 수 있다. 즉, 본 발명의 실시 예들 중 본 발명의 기술적 사상을 명확히 드러내기 위해 설명하지 않은 단계들 또는 부분들은 상기 문서들에 의해 뒷받침될 수 있다. 또한, 본 문서에서 개시하고 있는 모든 용어들은 상기 표준 문서들에 의해 설명될 수 있다.

[0035] 본 발명은 6G 뿐만 아니라 현재 표준화가 진행중인 3GPP 5G, Open RAN 등에 적용될 수 있다. 먼저, 본 명세서에서 제안하는 방법이 적용될 수 5G NR, Open RAN 시스템에 대한 사항을 간략히 설명하고 본 발명에 대해 기술하기로 한다.

[0036] 도 1은 5G NR의 전체적인 시스템 구조를 도시한 도면이다.

[0037] 도 1을 참조하면, NG-RAN은 NG-RA 사용자 평면(새로운 AS sublayer/PDCP/RLC/MAC/PHY) 및 UE(User Equipment)에 대한 제어 평면(RRC) 프로토콜 종단을 제공하는 기지국(gNB)들로 구성된다. gNB는 Xn 인터페이스를 통해 상호 연결된다. gNB는 또한, NG 인터페이스를 통해 NGC로 연결된다. gNB는 N2 인터페이스를 통해 AMF (Access and Mobility Management Function)로, N3 인터페이스를 통해 UPF (User Plane Function)로 연결된다.

[0038] NG-C는 새로운 RAN과 NGC 사이의 NG2 레퍼런스 포인트(reference point)에 사용되는 제어 평면 인터페이스를 나타내고, NG-U는 새로운 RAN과 NGC 사이의 NG3 레퍼런스 포인트(reference point)에 사용되는 사용자 평면 인터페이스를 나타낸다. 비 독립형(Non-standalone) NR은 gNB가 LTE eNB를 EPC로 제어 플레인 연결을 위한 앵커로 요구하거나 또는 eLTE eNB를 NGC로 제어 플레인 연결을 위한 앵커로 요구하는 배치 구성을 가지고 있다. 비 독립형 E-UTRA: eLTE eNB가 NGC로 제어 플레인 연결을 위한 앵커로 gNB를 요구하는 배치 구성을 가지고 있다. 사용자 평면 게이트웨이는 NG-U 인터페이스의 종단점이다.

[0039] 5G 시스템에서 무선장치(Radio Unit, RU)는 디지털 프론트 엔드(DFE)와 PHY 계층의 일부, 그리고 디지털 빔포밍 기능을 처리하는 장치이다. 5G RU 설계는 본질적으로 지능적이어야 하지만 RU 설계의 주요 고려 사항은 크기, 무게 및 전력 소비이다. 분산 장치(Distributed Unit, DU)은 RU 가까이에 위치하며 RLC, MAC 및 PHY 계층의 일부를 실행하는 분산 처리 장치이다. 이 논리 노드는 기능 분할 옵션에 따라 eNB/gNB 기능의 하위 집합을 포함하고 그 작동은 중앙 장치(Centralized Unit, CU)에 의해 제어된다. DU는 COTS 서버의 현장에 배포되는 분산 장치 소프트웨어일 수 있고, DU 소프트웨어는 일반적으로 현장의 RU 근처에 배포되며 RLC, MAC 및 PHY 계층의 일부를 실행할 수 있다.

[0040] CU는 RRC 및 PDCP 계층을 실행하는 중앙 집중식 장치이다. gNB는 각각 CP 및 UP에 대한 Fs-C 및 Fs-U 인터페이스를 통해 CU에 연결된 CU와 1개의 DU로 구성됩니다. 여러 DU가 있는 CU는 여러 gNB를 지원한다. 분할 아키텍처를 통해 5G 네트워크는 중간 가용성 및 네트워크 설계에 따라 CU와 DU 간에 프로토콜 스택의 서로 다른 배포를 활용할 수 있다. 사용자 데이터 전송, 이동성 제어, RAN 공유(MORAN), 포지셔닝, 세션 관리 등과 같은 gNB 기능을 포함하는 논리 노드이다. 단, DU에만 할당되는 기능은 예외이다. CU는 midhaul 인터페이스를 통해 여러 DU의 작동을 제어한다. 중앙 장치인 CU는 주로 RRC, SDAP 및 PDCP 프로토콜 계층을 포함하며 주로 비실시간 RRC, PDCP 프로토콜 스택 기능을 담당한다. CU는 코어 네트워크 UPF 싱킹 및 엣지 컴퓨팅의 통합 배포를 지원하기 위해 클라우드에 배포할 수 있다. CU와 DU는 F1 인터페이스를 통해 연결됩니다. 하나의 CU에서 하나 이상의 DU를 관리할 수 있다.

[0041] 요컨대, DU는 데이터 링크 계층과 스케줄링 기능을 포함하는 실시간 계층 1(L1, 물리 계층)과 하위 계층 2(L2)를 담당하고, CU는 비실시간 상위 L2 및 L3(네트워크 계층) 기능을 담당할 수 있다.

[0042] 다음은, 본 발명이 적용될 수 있는 Open RAN(O-RAN) 시스템에서의 각 용어들에 대한 정의를 간략히 기재한다.

[0043] Near-RT RIC: O-RAN Near-Real-Time RAN 지능형 컨트롤러이다. E2 인터페이스를 통한 세분화된 데이터 수집 및 작업을 통해 RAN 요소 및 리소스의 거의 실시간 제어 및 최적화를 가능하게 하는 논리적 기능이다. 여기에는 모델 교육, 추론 및 업데이트를 포함한 AI/ML(인공 지능/기계 학습) 워크플로가 포함될 수 있다.

[0044] Non-RT RIC: O-RAN 비 실시간 RAN 지능형 컨트롤러이고, A1 인터페이스를 통해 전달되는 콘텐츠를 구동하는 SMO 내의 논리적 기능이다. Non-RT RIC 프레임워크와 기능이 아래에 정의된 Non-RT RIC 어플리케이션(rApp)으로 구성된다.

[0045] Non-RT RIC 어플리케이션(rApps): Non-RT RIC 프레임워크의 R1 인터페이스를 통해 노출된 기능을 활용하여 A1 인터페이스 구동 같은 RAN 운영과 관련된 부가 가치 서비스를 제공하는 모듈식 어플리케이션이다. O1/O2 인터페이스를 통해 후속적으로 적용될 수 있는 가치와 조치를 추천하고 다른 rApp의 사용에 대한 “"enrichment information "을 생성한다. Non-RT RIC 내의 rApp 기능은 RAN 요소와 리소스의 비실시간 제어 및 최적화와 Near-RT RIC의 어플리케이션/기능에 대한 정책 기반 지침을 가능하게 한다.

[0046] Non-RT RIC 프레임워크: Near-RT RIC에 대한 A1 인터페이스를 논리적으로 종료하고 R1 인터페이스를 통해 런타임 처리에 필요한 내부 SMO 서비스 세트를 rApp에 노출시키는 SMO 내부 기능이다.

[0047] Non-RT RIC 내의 Non-RT RIC 프레임워크 기능은 rApp에 필요한 모델 교육, 추론 및 업데이트를 포함한 AI/ML 워크플로를 제공한다.

[0048] NMS: 레거시 Open Fronthaul M-Plane 배포를 지원하기 위해 지정된 O-RU용 네트워크 관리 시스템이다.

[0049] O-Cloud: O-Cloud는 관련 O-RAN 기능(Near-RT RIC, O-CU-CP, O-CU-UP 및 O-DU 등)을 지원 소프트웨어 구성 요소(예: 운영 체제, 가상 머신 모니터, 컨테이너 런타임 등) 기능을 호스팅하기 위한 O-RAN 요구 사항을 충족하는 물리적 인프라 노드 모음으로 구성된 클라우드 컴퓨팅 플랫폼이다.

[0050] O-CU-CP: O-RAN Central Unit-Control Plane: PDCP 프로토콜의 RRC 및 제어 평면 부분을 호스팅하는 논리 노드이다.

[0051] O-CU-UP: O-RAN Central Unit-User Plane: PDCP 프로토콜 및 SDAP 프로토콜의 사용자 평면 부분을 호스팅하는 논리 노드이다.

[0052] O-DU: O-RAN Distributed Unit: 하위 계층 기능 분할을 기반으로 RLC/MAC/High-PHY 계층을 호스팅하는 논리 노드이다.

[0053] O-eNB: E2 인터페이스를 지원하는 eNB 또는 ng-eNB이다.

[0054] O-RU: O-RAN Radio Unit: 하위 계층 기능 분할을 기반으로 하는 Low-PHY 계층 및 RF 처리를 호스팅하는 논리 노드이다. 이는 3GPP의 "TRP" 또는 "RRH"와 유사하지만 Low-PHY 계층(FFT/iFFT, PRACH 추출)을 포함한다는 점에서 더 구체적이다.

[0055] 도 2는 Open RAN(O-RAN) 시스템에서의 논리적 아키텍처(Logical Architecture)를 도시한 도면이다.

[0056] 도 2를 참조하여 설명하면, O-RAN의 논리적 아키텍처 내에서 무선 측에는 Near-RT RIC, O-CU-CP, O-CU-UP, O-DU 및 O-RU 기능이 포함된다. E2 인터페이스는 O-eNB를 Near-RT RIC에 연결한다. 도 2에 도시되어 있지는 않지만 O-eNB는 O-DU와 O-RU 기능을 Open Fronthaul 인터페이스를 통해 지원한다.

[0057] 관리 측에는 Non-RT-RIC 기능을 포함하는 SMO 프레임워크가 포함됩니다. 반면에, O-클라우드는 O-RAN을 충족하는 물리적 인프라 노드 모음으로 구성된 클라우드 컴퓨팅 플랫폼이다. 관련 O-RAN 기능(Near-RT RIC, O-CU-CP, O-CU-UP 및 O-DU 등), 지원 소프트웨어 구성 요소(예: 운영 체제, 가상 머신 모니터, Container Runtime 등) 및 적절한 관리 및 오케스트레이션 기능. O-RU의 가상화는 향후 더 연구될 것이다. 도 2에서와 같이 O-RU는 O-DU와 SMO에 대한 Open Fronthaul M-Plane 인터페이스를 종료시킨다.

[0058] 도 3은 응용서비스 별 레이턴시(latency)에 대해 설명하기 위한 도면이다.

[0059] 도 3을 참조하면, 5G 시스템 등에서 다양한 응용서비스를 제공한다. 5G 시스템에서도 지연(latency 혹은 delay)이 매우 중요하게 여겨지는 응용서비스가 확대되었다. 도 3에 점선으로 표시한 영역에서 재난/재해 알림(Disaster alert), 실시간 게이밍(Real time gaming), 자율주행(Autonomous Driving), 증강 현실, 가상 현실,

촉각 인터넷 등의 응용서비스가 특히 지연(latency)가 매우 중요하기 때문에 원하는 시간 내에 응용 데이터 유닛(Application Data Unit, ADU)이 전달되어야 한다.

[0060] 도 4는 5G, 6G 등의 프로토콜 스택-사용자 평면을 예시한 도면이다.

[0061] 도 4를 참조하면, URLLC의 목표를 달성하기 위해 하위계층들(lower layers)에서의 초저지연(ultra-low latency)을 요구하고 있다. 즉, URLLC 서비스 제공을 위해 IP/SDAP/PDCP/RLC/MAC 계층들과 같은 하위계층들(layer 2/3)에서(즉, 패킷-수준 저지연에서) ultra-low latency를 요구하고 있다. 이와 같이, 5G 시스템 등에서 지연(latency)이 중요한 서비스는 ADU가 원하는 시간 내에 전달되어야 함이 요구되지만, 아직까지 응용-수준(application-level)에서는 이를 해결할 수 없었다.

[0062] 도 5는 6G 종단간 네트워크 구조를 예시한 도면이다.

[0063] 6G 종단간 응용수준에서의 성능 보장을 위해서는 효율적인 무선자원할당 기법이 요구된다. 차세대 네트워크에서는 인공지능을 활용하는 서비스들은 많은 컴퓨팅을 활용하기 위해 엣지 서버의 도움이 요구된다.

[0064] 또한, 도 5의 하측에 도시된 바와 같이, 송신측 종단의 전자 장치가 전송하고자 하는 ADU 크기는 실시간으로 가변할 수 있고 네트워크 혼잡 등의 네트워크 상태도 가변 할 수 있기 때문에, 송신측 종단의 전자 장치의 응용계층에서 네트워크 전송 API 함수를 호출한 시점부터 수신측 종단의 전자 장치의 응용계층에 타겟 응용시간지연 값 내로 전송 완료하는 것이 중요하다. 사용자 체감 성능의 일종인 QoE(Quality of Experience) 보장을 위해 각 ADU는 요구되는 ADU 전송 완료 시간 내에 전송 완료되어야 한다.

[0065] 도 6은 응용-수준에서의 서비스 지연시간(latency)를 설명하기 위한 도면이고, 도 7은 차세대 네트워크에서의 요구사항 중 하나인 지연시간을 설명하기 위한 도면이다.

[0066] 도 6은 차세대 네트워크에서 지원할 몰입형 서비스를 제공하기 위한 서비스 지연시간을 예시하고 있다. 사용자가 인지하는 지연시간(예, 서비스 지연시간)에는 컴퓨팅(계산) 지연시간과 네트워크 지연시간(예, front-haul(프런트홀) 지연시간, 무선 인터페이스(Air-interface) 지연시간 및 무선 지연시간이 모두 포함된다. 모바일 장치의 제한된 용량으로 인해 3D 렌더링이나 신경망 기반 추론과 같은 과중한 업무(task)들을 지원하려면 컴퓨팅 오프로딩(offloading) 프로세스가 필요하다. 네트워크 지연시간에는 프론트-홀 지연시간, 무선-인터페이스 지연시간, 무선 지연시간이 포함되는데, 프론트-홀 지연시간과 무선-인터페이스 지연시간은 매우 작은 값이어서 네트워크 지연시간은 무선 지연시간인 것으로 가정한다.

[0067] 도 7을 참조하면, 컴퓨팅 지연시간은 요청 발행부터 응용 서버(예, 엣지, fog, 클라우드 서버)에서 출력 생성까지의 지연을 말한다. 무선 (전송) 지연시간은 두 끝점(즉, 서버와 사용자 장치) 사이의 통신 서비스(사용자 평면)에 대한 통신을 위한 종단간 지연시간을 말한다. 도메인 별(즉, 컴퓨팅 도메인 및 네트워크 도메인) 요구 사항은 서비스 단위당 사전에 정의된 컴퓨팅 지연시간 및 무선 지연시간 예산(budget)에 의해 정적인 특성이 있다. 본 발명에서는 주로 컴퓨팅 지연시간과 무선 지연시간(혹은 네트워크 지연시간)에 대해 포커싱하여 설명하도록 한다. 또한, 컴퓨팅 지연시간은 서버에서의 컴퓨팅 지연시간만 고려하는 것으로 가정한다.

[0068] 도 8은 차세대 네트워크에서 제공할 서비스들의 업무 부담을 예시한 도면이다.

[0069] 도 8에 도시한 바와 같이, 컴퓨팅 부담은 업무 타입(예를 들어, 신경망 크기, 압축률 등)에 따라 결정된다. 3D 렌더링과 같은 업무는 상대적으로 컴퓨팅 도메인에서의 컴퓨팅 부담이 높고 네트워크 도메인에서의 네트워크 부담도 중간 이상으로 높은 업무임을 알 수 있다. 반면, 이미지의 segmentation과 estimation 과 같은 업무는 각각 컴퓨팅 부담은 중간 이상이나, 네트워크 부담은 각각 중간~낮은 수준 사이이거나 낮은 수준이다. 출력 크기 및 채널 품질에 의해 네트워크 부담이 결정될 수 있다. 따라서, 이질적인 업무에 대한 몰입형 서비스를 효율적인 비용으로 지원하기 위해서는 동적이고 상호적인 연합(혹은 협력) 자원 스케줄링이 필요하다.

[0070] 도 9는 컴퓨팅 오프로딩을 위한 기준 RPC 업무흐름(workflow)를 도시한 도면이다.

[0071] 컴퓨팅 작업이 많은 몰입형 서비스를 지원하려면 모바일 엣지 컴퓨팅(MEC) 노드로의 컴퓨팅 오프로딩이 필수적이다. 3D 렌더링, 위치 추정 또는 이동성 예측은 가상 세계와 실제 세계를 완벽하게 통합해야 한다. 서비스 지연시간에는 데이터를 생성하기 위한 컴퓨팅 지연시간과 출력을 사용자 장치에 전달하기 위한 네트워크 지연시간이 모두 포함된다. 몰입형 서비스는 도메인 별로 매우 다양한 부담을 요구한다. 다양한 환경과 상황은 서비스 중에도 동적인 부담을 준다. 따라서, 비용적인 효율성을 달성하려면 서비스의 업무(task)에 따른 budget 조정이 필요하다.

[0072] 이러한 비용적인 효율성을 달성하기 위해서는 도메인 간(예, 네트워크 도메인 및 컴퓨팅 도메인) 간의 정보 교환과 협력 스케줄링 및 도메인 별 스케줄링 방법이 요구된다. 이하, 도메인 간의 정보 교환과 협력 스케줄링 및 도메인 별 스케줄링 방법에 대해 설명한다.

[0073] 도메인 간의 정보 교환

[0074] 요구되는 컴퓨팅 지연시간, 메모리(량), 출력데이터 크기는 동적으로 변하기 때문에 응용 서버에서 이용할 수 있는 업무 별 동적성에 대한 정보가 교환될 필요가 있다. 무선자원(즉, 무선자원블록(RB) 또는 무선자원블록그룹(RBG))당 전송할 수 있는 데이터 비트는 채널 품질에 따라 달라진다. 따라서, 기지국(예, radio unit, distributed unit, and central unit in O-RAN)에서 이용할 수 있는 채널 동적성에 대한 정보가 교환될 필요가 있다.

[0075] 협력 스케줄링 및 도메인 별 스케줄링

[0076] 컴퓨팅 도메인과 네트워크 도메인 사이의 정밀하고 상호작용하는 budgeting이 필요해서 응용서비스의 각 업무에 대해 상대적 비용 혹은 가치를 결정할 필요가 있다. 스케줄링된 budget을 담보하기 위하여 작업 스케줄링 및 무선자원 스케줄링이 요구된다.

[0077] 이러한 요구 사항을 해결하기 위해 도메인 간 정보 교환을 위한 프레임워크 및 오케스트레이션 워크플로우(orchestration workflow)와 새로운 협력 스케줄링 방법이 필요한데, 이하에서 본 발명에 따른 구체적인 방안을 설명한다.

[0078] 도 10은 본 발명에 따른 응용 수준 성능 보장을 위한 네트워크(혹은 무선) 및 컴퓨팅 자원의 협력 스케줄링을 수행하기 위한 장치의 구성을 블록으로 나타낸 도면이고, 도 11은 본 발명에 따른 응용 수준 성능 보장을 위한 네트워크(혹은 무선) 및 컴퓨팅 자원의 협력 스케줄링 방법을 설명하기 위한 예시적 도면이다.

[0079] 도 10을 참조하면, 본 발명에 따른 응용 수준 성능 보장을 위한 네트워크(혹은 무선) 및 컴퓨팅 자원의 협력 스케줄링을 수행하기 위한 장치(이하, 설명의 편의를 위해 '협력 스케줄링 장치' 로 약칭한다)(100)는 프로세서(110), 통신부(120)(일 예, 무선통신부) 및 메모리부(130)를 포함할 수 있다.

[0080] 프로세서(110)는 본 발명에서 제안한 응용 수준 성능 보장을 위한 네트워크(혹은 무선) 및 컴퓨팅 자원의 협력 스케줄링 방법을 구현하도록 구성될 수 있다. 메모리부(130)는 프로세서(110)와 전기적으로 연결되어 프로세서(110)의 동작과 관련한 다양한 정보를 저장한다. 통신부(120)는 프로세서(110)와 전기적으로 연결되어 있고 무선 혹은 유선 신호를 송신 및/또는 수신한다.

[0081] 협력 스케줄링 장치는 협력 스케줄러, 적응형 협력 스케줄러 등 다양하게 호칭될 수 있다. 협력 스케줄링 장치(100)의 통신부(120)는 응용 서버 혹은 응용 provider(200) 등으로부터 응용서비스의 업무 별 요구사항에 대한 제1 정보를 수신할 수 있다. 통신부(120)는 제1 도메인 스케줄러에 해당하는 컴퓨팅 스케줄러(300)로부터 응용서비스의 컴퓨팅과 관련된 제2 정보를 수신할 수 있다. 통신부(120)는 제2 도메인 스케줄러에 해당하는 무선 스케줄러(400)로부터 사용자 별 채널상태와 관련된 정보 및 네트워크 트래픽과 관련된 정보 중 적어도 어느 하나를 포함하는 제3 정보를 수신할 수 있다.

[0082] 협력 스케줄링 장치(100)의 프로세서(110)는 응용서비스의 업무 별 요구사항에 대한 제1 정보, 응용서비스의 컴퓨팅과 관련된 제2 정보 및 사용자 별 채널상태와 관련된 정보 및 네트워크 트래픽과 관련된 정보 중 적어도 어느 하나를 포함하는 제3 정보에 기초하여 응용 서비스의 업무 별 컴퓨팅 자원과 네트워크 자원에 대한 가치 또는 비용을 각각 산출할 수 있다. 프로세서(110)는 산출된 가치 또는 비용에 기초하여 응용서비스의 전체 지연시간을 컴퓨팅 지연시간과 네트워크 지연시간에 분배하여 할당할 수 있다. 이와 같이, 프로세서(110)는 컴퓨팅 도메인과 네트워크 도메인 간의 부담을 조절할 수 있다.

[0083] 여기서, 응용 서버(100)가 전송하는 상기 제1 정보에는 응용서비스의 업무 별 요구 지연시간(예, task 1의 지연시간, task 2의 지연시간) 및 업무 별 예상 출력데이터 크기(예, task 1의 출력데이터 크기, task 2의 출력데이터 크기) 중 적어도 어느 하나가 더 포함될 수 있다.

[0084] 연결된 서버의 종류(예를 들어, 엣지 서버, fog, 클라우드 등)에 따라 자원의 비용이 달라지는데, 중앙 서버일수록 컴퓨팅 자원 및 메모리 자원이 풍부하기 때문에 비용이나 가치가 낮다고 할 수 있다. 또한, 시간에 따라 변하는 트래픽 양에 따라 자원의 비용이 달라지고, 업무의 종류에 따라 자원의 비용이 달라진다. 따라서, 프로세서(110)는 컴퓨팅 자원의 가치 혹은 비용을 산출하는 경우에 연결된 서버의 종류, 업무 별 종류 및 시간에 따

라 변하는 트래픽 양 중 적어도 어느 하나에 더 기초하여 산출할 수 있다.

[0085] 네트워크 자원(무선 자원)의 경우에도 사용자의 채널 상황에 따라 자원의 비용이 달라질 수 있고, 시간에 따라 변하는 트래픽 양에 따라 자원의 비용이 달라질 수 있으며, 채널추정의 정확도에 따라 자원의 비용이 달라질 수 있다. 따라서, 프로세서(110)는 사용자 별 채널 추정의 정확도, 사용자의 채널 상황 및 시간에 따라 변하는 트래픽 양 중 적어도 어느 하나에 더 기초하여 네트워크 자원의 가치 혹은 비용을 산출할 수 있다.

[0086] 프로세서(110)는 산출된 컴퓨팅 자원 및 네트워크 자원의 각각의 가치에 따라 전체 지연시간을 컴퓨팅 도메인과 네트워크 도메인에 분리하여 할당할 수 있다. 컴퓨팅 자원인 컴퓨팅 및 메모리 자원은 요구 컴퓨팅 지연시간으로 결정될 수 있다. 네트워크 자원은 데이터 전송률 혹은 데이터 전송량으로 결정될 수 있다. 네트워크(무선) 자원은 각 TTI(transmission time interval) 마다 요구되는 데이터 처리량에 따라 달라진다. 일 예로서, 기본적인 무선자원할당 시간 혹은 단위 무선자원할당 시간으로 프레임 구조에서의 뉴머롤로지(Numerology)에 따라 250 μs에서 1ms로 정해질 수 있다. 이러한 단위 무선자원할당 시간은 5G, 6G, O-RAN에 사용되는 프레임 구조의 뉴머롤로지에 의해 결정될 수 있다.

[0087] 도 11을 참조하면, 상술한 바와 같이 협력 스케줄링 장치(100)는 적응형 협력 스케줄러로 호칭할 수 있다. 협력 스케줄링 장치(100)는 End-to-End Orchestrator(10) 혹은 E2E 오케스트레이터(10)에 속할 수 있다. E2E 오케스트레이터(10)는 적응형 협력 스케줄러(100)(혹은 cooperative scheduler) 외에 admission 제어부, 슬라이스 관리부(slice management), 모빌리티 관리부(mobility management)부를 더 포함할 수 있다.

[0088] 응용 서버 혹은 응용 provider(200)는 응용서비스의 업무 별 요구사항에 대한 제1 정보를 협력 스케줄링 장치(100)로 전송할 수 있다. 제1 도메인 스케줄러에 해당하는 컴퓨팅 스케줄러(300)는 엣지 오케스트레이터(30)에 속하고, 엣지 오케스트레이터(30)는 컴퓨팅 스케줄러(300)외에 자원 관리부(resource management)를 포함할 수 있다. 컴퓨팅 스케줄러(300)는 응용서비스의 각 업무 별 서비스 지연시간을 보장하기 위한 업무의 컴퓨팅 자원(예, Flops), 메모리 사용량, 서버의 트래픽 정보 등을 획득한다. 컴퓨팅 스케줄러(300)는 컴퓨팅 자원 요구량, 메모리 사용량, 서버 트래픽 정보 등을 응용서비스의 컴퓨팅과 관련된 정보라고 결정할 수 있고, 컴퓨팅 스케줄러(300)는 응용서비스의 컴퓨팅과 관련된 제2 정보를 협력 스케줄링 장치(100)에 전송해 줄 수 있다.

[0089] 제2 도메인 스케줄러에 해당하는 무선 스케줄러(400)는 RAN 오케스트레이터(40)에 속할 수 있다. RAN 오케스트레이터(40)는 무선 스케줄러(400) 외에도 자원관리부(resource management), 채널 예측기(혹은 추정기)(channel predictor) 등을 포함할 수 있다. 무선 스케줄러(400)는 응용서비스의 각 업무 별 서비스 지연시간을 보장하기 위해 채널 정보, 네트워크 트래픽 정보를 획득할 수 있다. 무선 스케줄러(400)는 채널 정보, 네트워크 트래픽 정보, 예측 채널 정보를 네트워크와 관련된 정보로 결정할 수 있다. 무선 스케줄러(400)는 사용자 별 채널상태와 관련된 정보 및 네트워크 트래픽과 관련된 정보 중 적어도 어느 하나를 포함하는 제3 정보를 협력 스케줄링 장치(100)에 전송할 수 있다.

[0090] 협력 스케줄링 장치(100)의 프로세서(110)는 상기 제1 정보에 포함된 응용서비스 제공자 혹은 응용 서버(200)로부터 제공된 업무 별 요구 지연시간, 출력데이터 크기(생성된 결과의 크기), 제2 정보에 포함된 연결된 컴퓨팅 자원에 따른 업무의 컴퓨팅 및 예측 메모리 사용량, 제3 정보에 포함된 채널 정보(예, CQI(Channel Quality Indicator), 거리, 이동속도, LOS probability, 신호지표(예, SINR, RSRP, RSRQ))에 기초하여 응용서비스의 각 업무 별 컴퓨팅 지연시간과 네트워크 지연시간을 분배할 수 있다.

[0091] 이후, 협력 스케줄링 장치(100)는 컴퓨팅 스케줄러(300)에게 할당한 컴퓨팅 지연시간에 대한 정보를 전송해 줄 수 있고, 무선 스케줄러(400)에게 할당한 네트워크 지연시간에 해당하는 데이터 전송량 혹은 데이터 전송률 등을 전송해 줄 수 있다.

[0092] 도 12는 본 발명에 따른 응용 수준 성능 보장을 위한 네트워크(혹은 무선) 및 컴퓨팅 자원의 협력 스케줄링을 위한 정보 및 결과를 교환하는 인터페이스를 설명하는 도면이다.

[0093] 도 12를 참조하면, 응용 서버 혹은 응용 provider(200)는 응용서비스의 업무 별 요구사항에 대한 제1 정보를 X2 인터페이스를 통해 협력 스케줄링 장치(100)로 전송할 수 있다. 컴퓨팅 스케줄러(300)는 응용서비스의 컴퓨팅과 관련된 제2 정보를 F1-U 인터페이스 등을 통해 협력 스케줄링 장치(100)에 전송해 줄 수 있다. 한편, O-RAN 시스템에서는 O-RU(radio unit)가 E2 인터페이스를 통해 O-CU(central unit) 또는 O-DU(distributed unit)로 사용자 별 채널 정보 등의 상기 제3 정보를 전송할 수 있다. 무선 스케줄러(400)는 사용자 별 채널상태와 관련된 정보 및 네트워크 트래픽과 관련된 정보 중 적어도 어느 하나를 포함하는 제3 정보를 F1-U 인터페이스 혹은 프론트-홀(Front-haul)을 통해 협력 스케줄링 장치(100)에 전송할 수 있다. 한편, O-RAN 시스템에서는 O-

RU(radio unit)가 E2 인터페이스를 통해 O-CU(central unit) 또는 O-DU(distributed unit)로 사용자 별 채널 정보 등의 상기 제3 정보를 전송할 수 있다.

[0094] 컴퓨팅 스케줄러(300)은 협력 스케줄링 장치(100)로부터 전달받은 할당된 컴퓨팅 지연시간내에 각 업무를 마칠 수 있게 CPU, GPU(Graphics Programming Unit), TPU(Tensor Processing Unit) 등의 자원과 cache와 같은 메모리를 할당할 수 있다. 이때, 컴퓨팅 스케줄러(300)는 일반적인 컴퓨팅 자원 스케줄러와 다르게 무선 자원의 채널 상태를 고려하여 우선순위를 가지도록 구성할 수 있다. 예를 들어, 컴퓨팅 스케줄러(300)는 채널 상태가 가장 좋은 특정 응용서비스의 업무가 가장 높은 우선순위를 갖도록 구성할 수 있다. 컴퓨팅 스케줄러(300)는 각 업무별로 결정된 시간이 짧은 업무일수록 우선순위를 높게 가지도록 구성할 수 있고, 할당된 컴퓨팅 지연시간을 지킬 수 있으면 업무의 대기가 가능하다.

[0095] 무선 스케줄러(400)는 협력 스케줄링 장치(100)로부터 전달받은 데이터 전송률 혹은 데이터 전송량의 정보에 기초하여 각 TTI 마다 데이터 처리량을 할당할 수 있다. 무선 스케줄러(400)는 예측 및 추정된 채널에 따라 각 TTI 마다 데이터 처리량에 따른 필요 무선자원 양을 결정할 수 있다. 무선 스케줄러(400)는 채널 추정의 오류로 인한 데이터 처리량 부족은 다음 TTI에서 추가적으로 할당하도록 구성될 수 있다. 무선 스케줄러(400)의 무선자원할당은 컴퓨팅이 끝나 결과가 나온 후부터 요구 전제 지연시간 내에서만 가능하다.

[0096] 이상에서 설명한 바와 같이 협력 스케줄링 장치(100)가 두 도메인 간을 적응적이며 협력적으로 스케줄링하여 차세대 네트워크에서 제공할 몰입형 서비스 등의 응용 서비스의 전체 지연시간을 만족하도록 함으로써 응용 수준에서의 성능 보장이 가능하게 한다.

[0097] 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, FPGA(field programmable gate 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 어플리케이션을 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.

[0098] 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, 가상장치(virtual equipment), 컴퓨터 저장 매체 또는 장치, 또는 전송되는 신호 파(signal wave)에 영구적으로, 또는 일시적으로 구체화(embody)될 수 있다. 소프트웨어는 네트워크로 연결된 컴퓨팅장치 상에 분산되어서, 분산된 방법으로 저장되거나 실행될 수도 있다. 소프트웨어 및 데이터는 하나 이상의 컴퓨터 판독 가능 기록 매체에 저장될 수 있다.

[0099] 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), CDROM, 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 코드 뿐만 아니라 인터프리터 등을 사용해서 컴퓨터에 의해서 실행될 수 있는 고급 언어 코드를 포함한다. 상기된 하드웨어 장치는 실시예의 동작을 수행하기 위해 하나 이상의 소프트웨어 모듈로서 작동하도록 구성될 수 있으며, 그 역도 마찬가지이다.

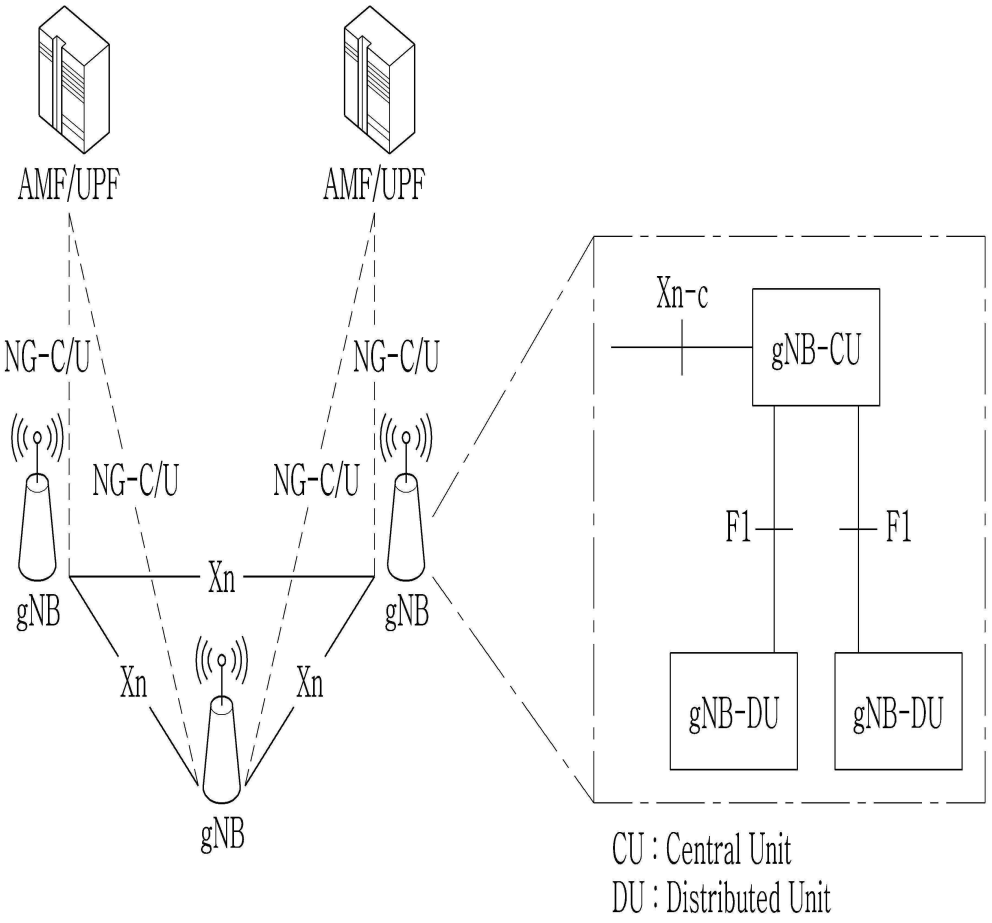
[0100] 이상에서 설명된 실시예들은 본 발명의 구성요소들과 특징들이 소정 형태로 결합된 것들이다. 각 구성요소 또는 특징은 별도의 명시적 언급이 없는 한 선택적인 것으로 고려되어야 한다. 각 구성요소 또는 특징은 다른 구성요소나 특징과 결합되지 않은 형태로 실시될 수 있다. 또한, 일부 구성요소들 및/또는 특징들을 결합하여 본 발명의 실시예를 구성하는 것도 가능하다. 본 발명의 실시예들에서 설명되는 동작들의 순서는 변경될 수 있다. 어느 실시예의 일부 구성이나 특징은 다른 실시예에 포함될 수 있고, 또는 다른 실시예의 대응하는 구성 또는 특징과 교체될 수 있다. 특허청구범위에서 명시적인 인용 관계가 있지 않은 청구항들을 결합하여 실시예를 구성하거나 출원 후의 보정에 의해 새로운 청구항으로 포함시킬 수 있음은 자명하다.

[0101] 본 발명에서 프로세서(110)는 하드웨어(hardware) 또는 펌웨어(firmware), 소프트웨어, 또는 이들의 결합에 의해 구현될 수 있다. 하드웨어를 이용하여 본 발명의 실시예를 구현하는 경우에는, 본 발명을 수행하도록 구성된 ASICs(application specific integrated circuits) 또는 DSPs(digital signal processors), DSPDs(digital signal processing devices), PLDs(programmable logic devices), FPGAs(field programmable gate arrays) 등이 프로세서에(110)에 구비될 수 있다.

[0102] 본 발명은 본 발명의 필수적 특징을 벗어나지 않는 범위에서 다른 특정한 형태로 구체화될 수 있음은 당업자에게 자명하다. 따라서, 상기의 상세한 설명은 모든 면에서 제한적으로 해석되어서는 아니되고 예시적인 것으로 고려되어야 한다. 본 발명의 범위는 첨부된 청구항의 합리적 해석에 의해 결정되어야 하고, 본 발명의 등가적 범위 내에서의 모든 변경은 본 발명의 범위에 포함된다.

도면

도면1



도면2

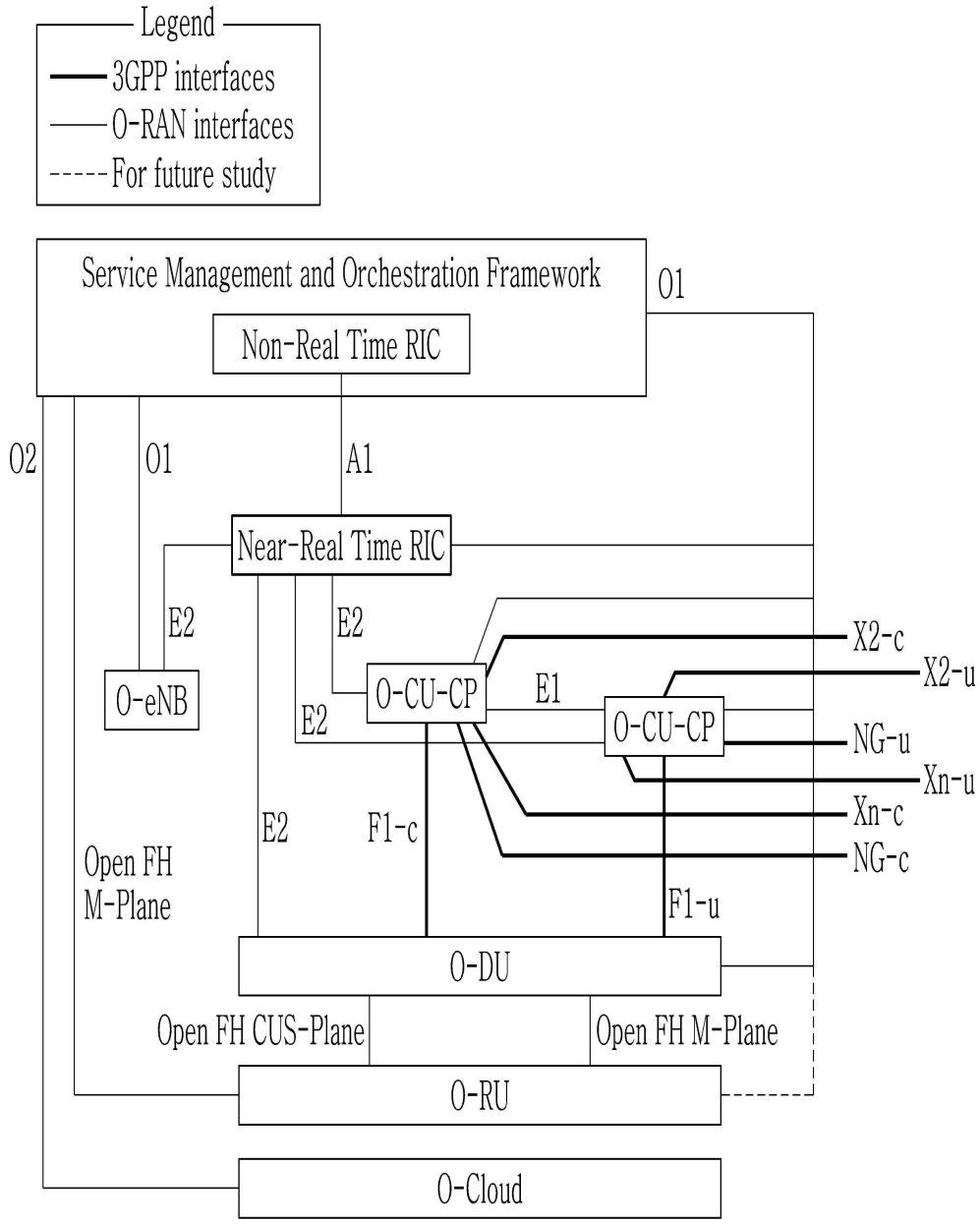
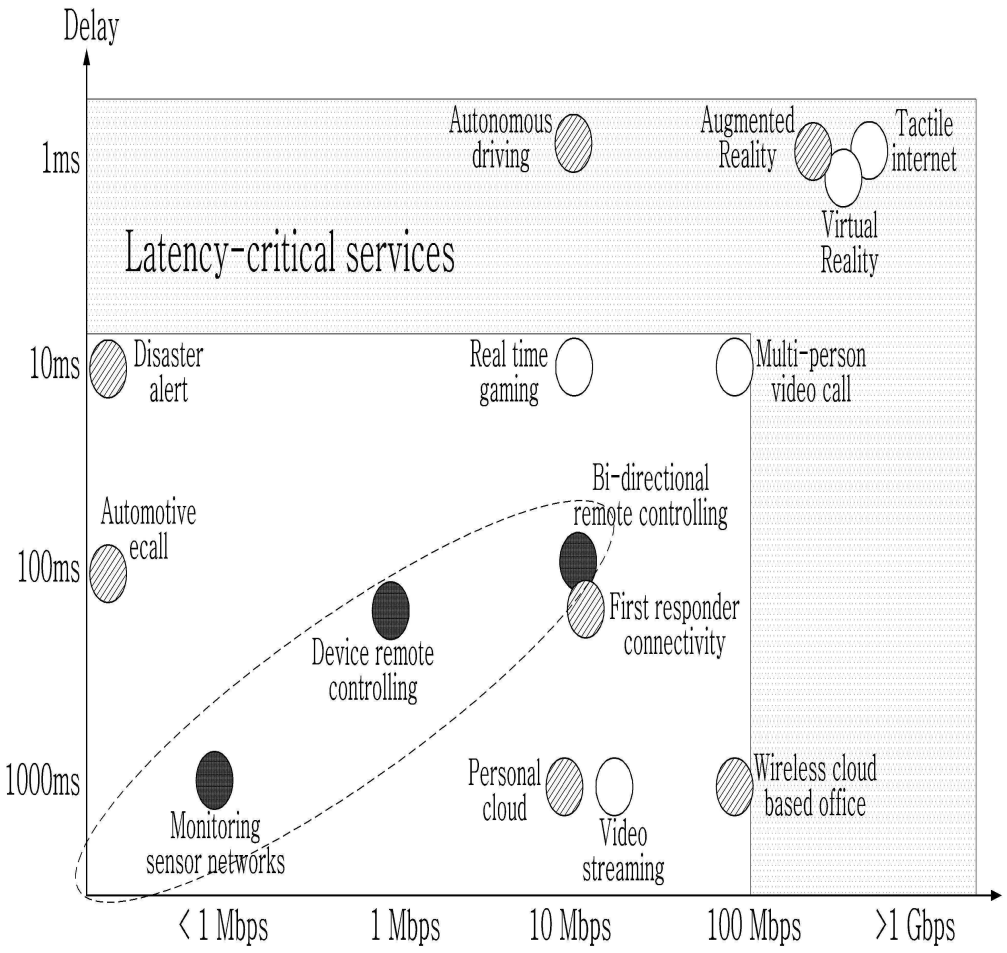
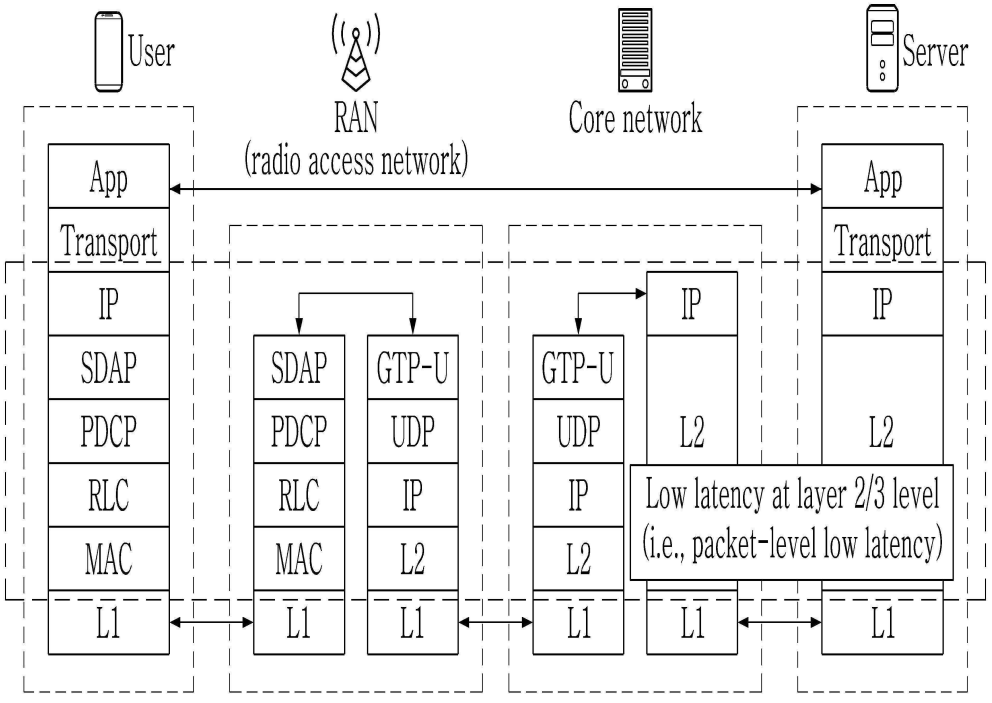


Figure 4.1-2: Logical Architecture of O-RAN

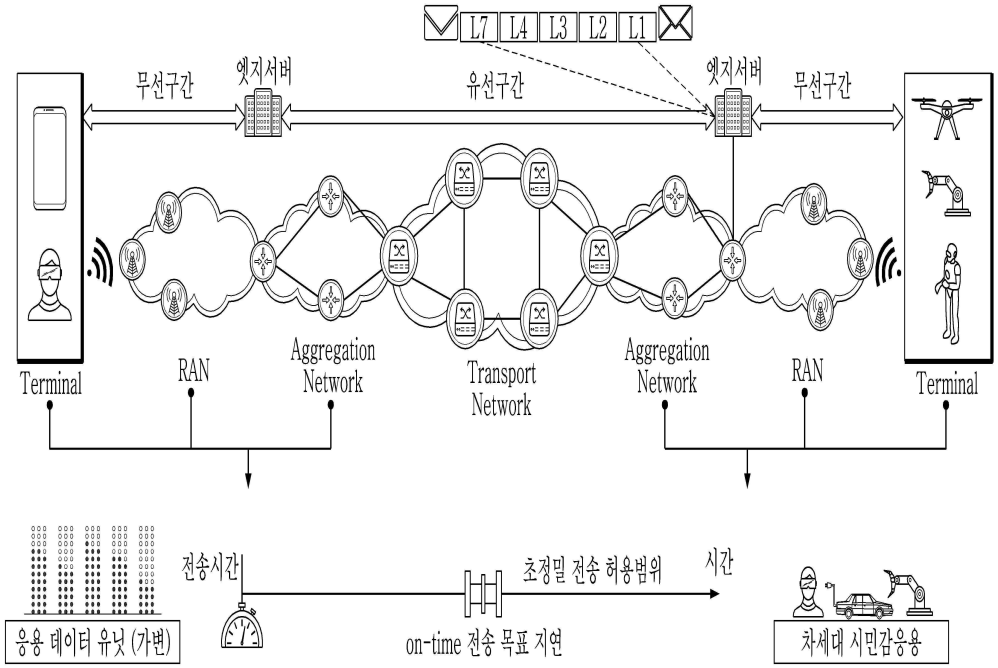
도면3


Delay
1ms
10ms
100ms
1000ms
Latency-critical services
Autonomous driving
Augmented Reality
Tactile internet
Virtual Reality
Disaster alert
Real time gaming
Multi-person video call
Automotive ecall
Bi-directional remote controlling
First responder connectivity
Device remote controlling
Monitoring sensor networks
Personal cloud
Video streaming
Wireless cloud based office
< 1 Mbps
1 Mbps
10 Mbps
100 Mbps
>1 Gbps

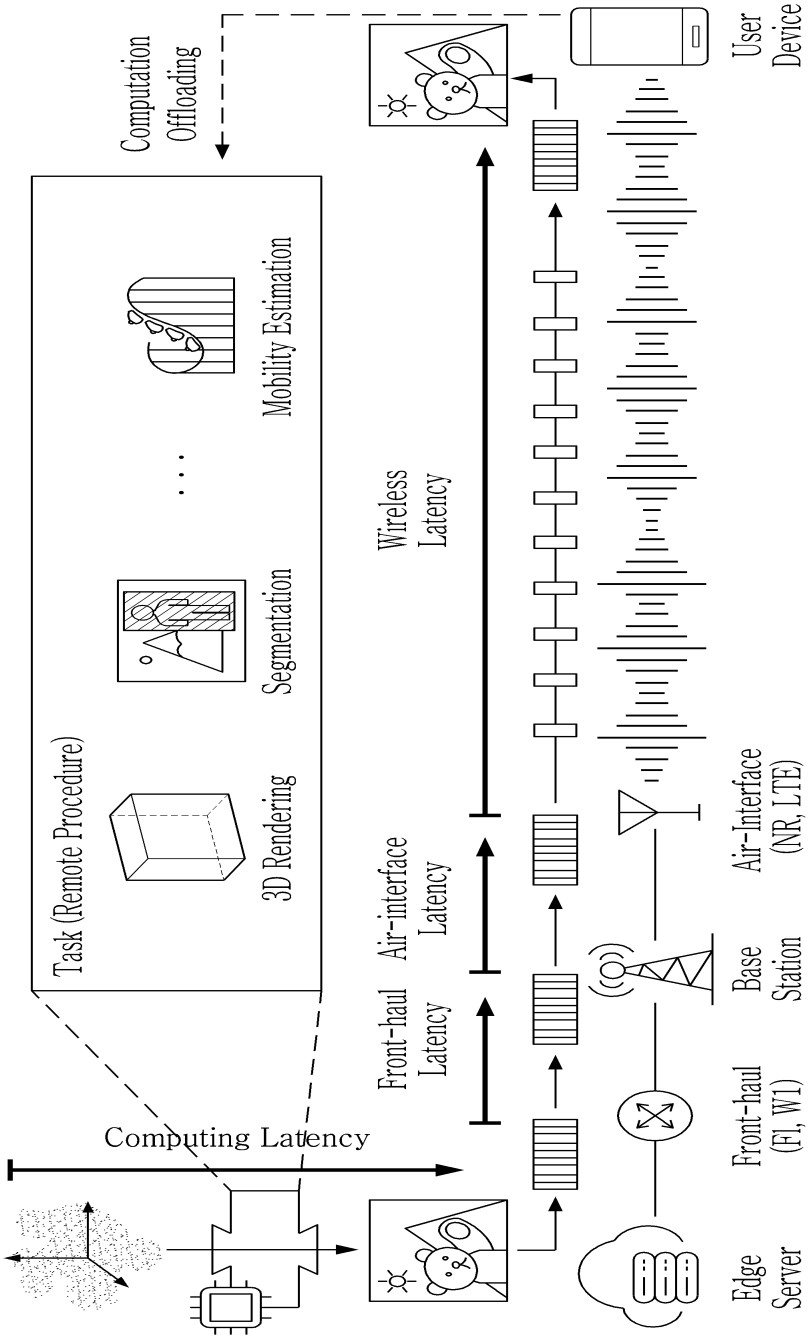
도면4


User
RAN
(radio access network)
Core network
Server
App
Transport
IP
SDAP
PDCP
RLC
MAC
L1
SDAP
PDCP
RLC
MAC
GTP-U
UDP
IP
L2
L1
GTP-U
UDP
IP
L2
L1
IP
L2
L1
App
Transport
IP
L2
L1
Low latency at layer 2/3 level
(i.e., packet-level low latency)

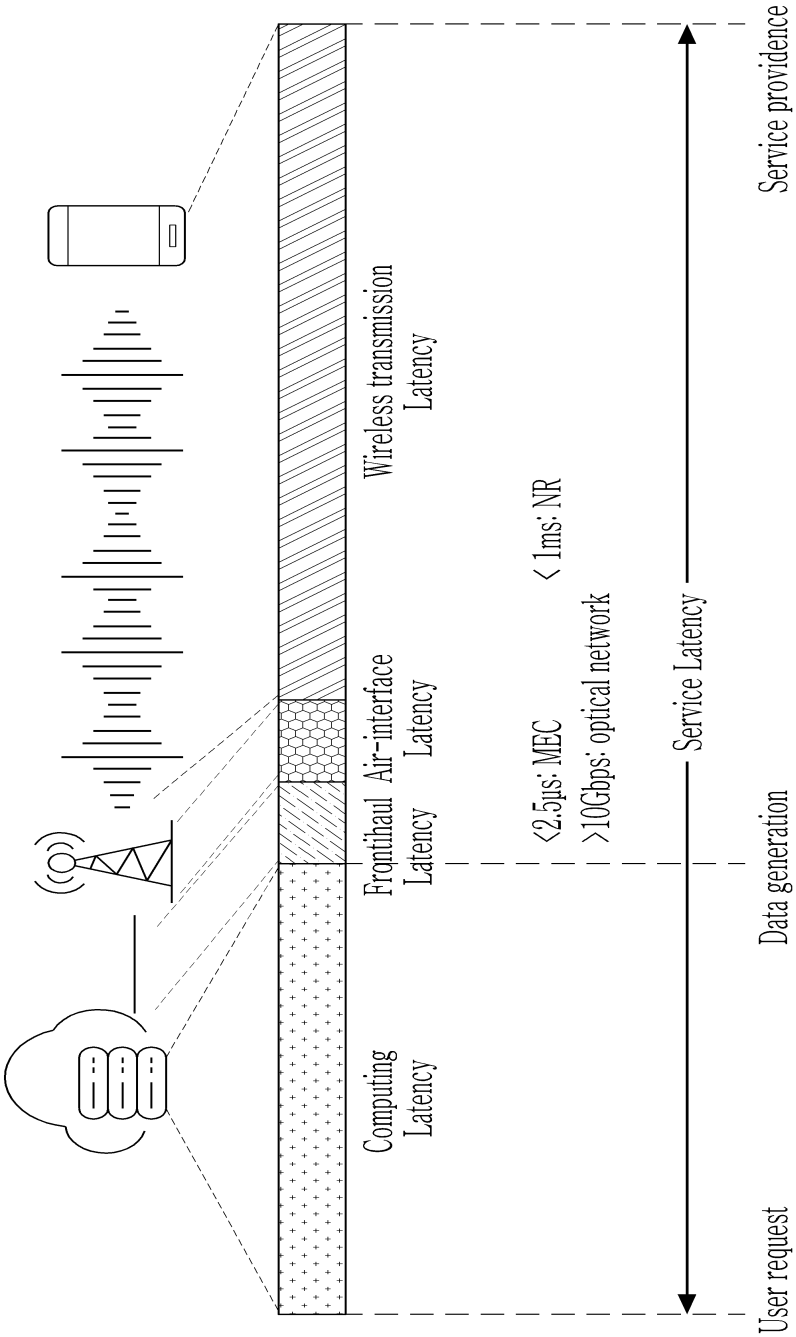
도면5


L7
L4
L3
L2
L1
무선구간
엣지서버
유선구간
엣지서버
무선구간
Terminal
RAN
Aggregation Network
Transport Network
Aggregation Network
RAN
Terminal
전송시간
초정밀 전송 허용범위
시간
응용 데이터 유닛 (가변)
on-time 전송 목표 지연
차세대 시민감응용

도면6


Computation Offloading
User Device
Task (Remote Procedure)
3D Rendering
Segmentation
Mobility Estimation
Wireless Latency
Air-interface Latency
Front-haul Latency
Computing Latency
Edge Server
Front-haul (FI, W1)
Base Station
Air-Interface (NR, LTE)

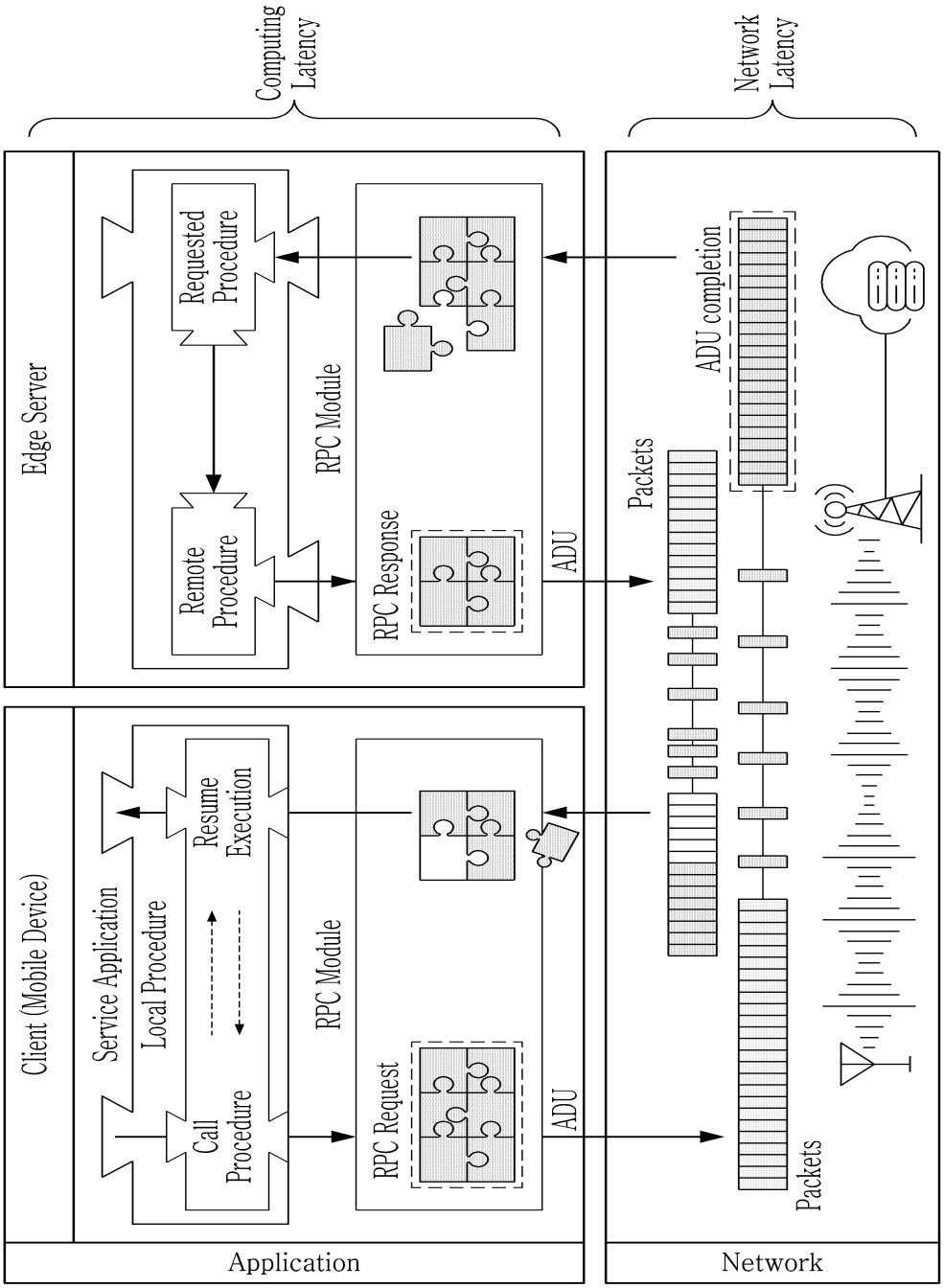
도면7


Service providence
Wireless transmission Latency
Fronthaul Latency
Air-interface Latency
Computing Latency
<2.5μs: MEC
>10Gbps: optical network
< 1ms: NR
Service Latency
Data generation
User request

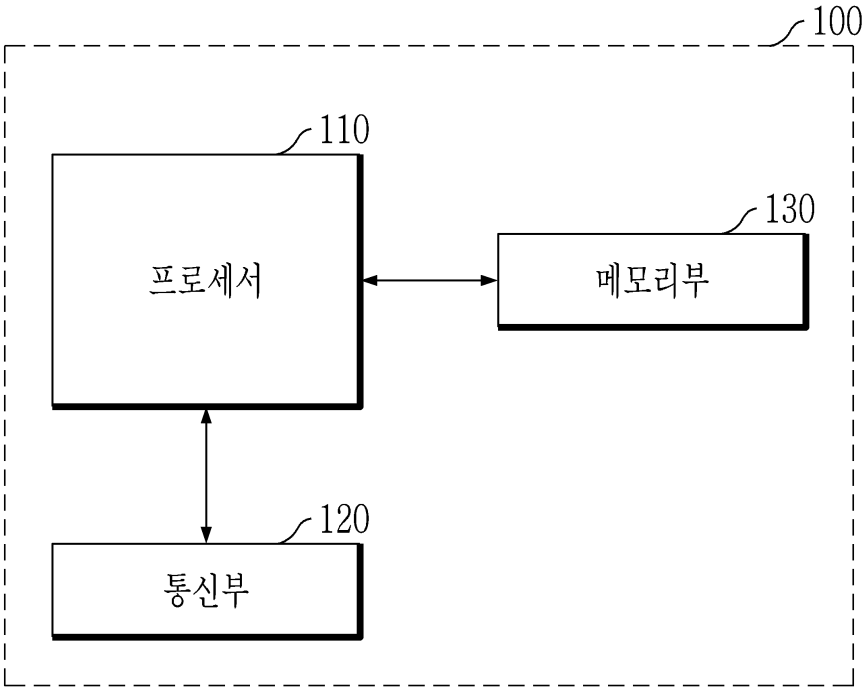
도면8

: Determined by the task type (e.g., neural network size, compression ratio, etc.)
Computational burden
Network burden
: Determined by output size and channel quality
3D Rendering
high
middle ~ high
Segmentation
middle ~ high
low ~ middle
Estimation
middle ~ high
low

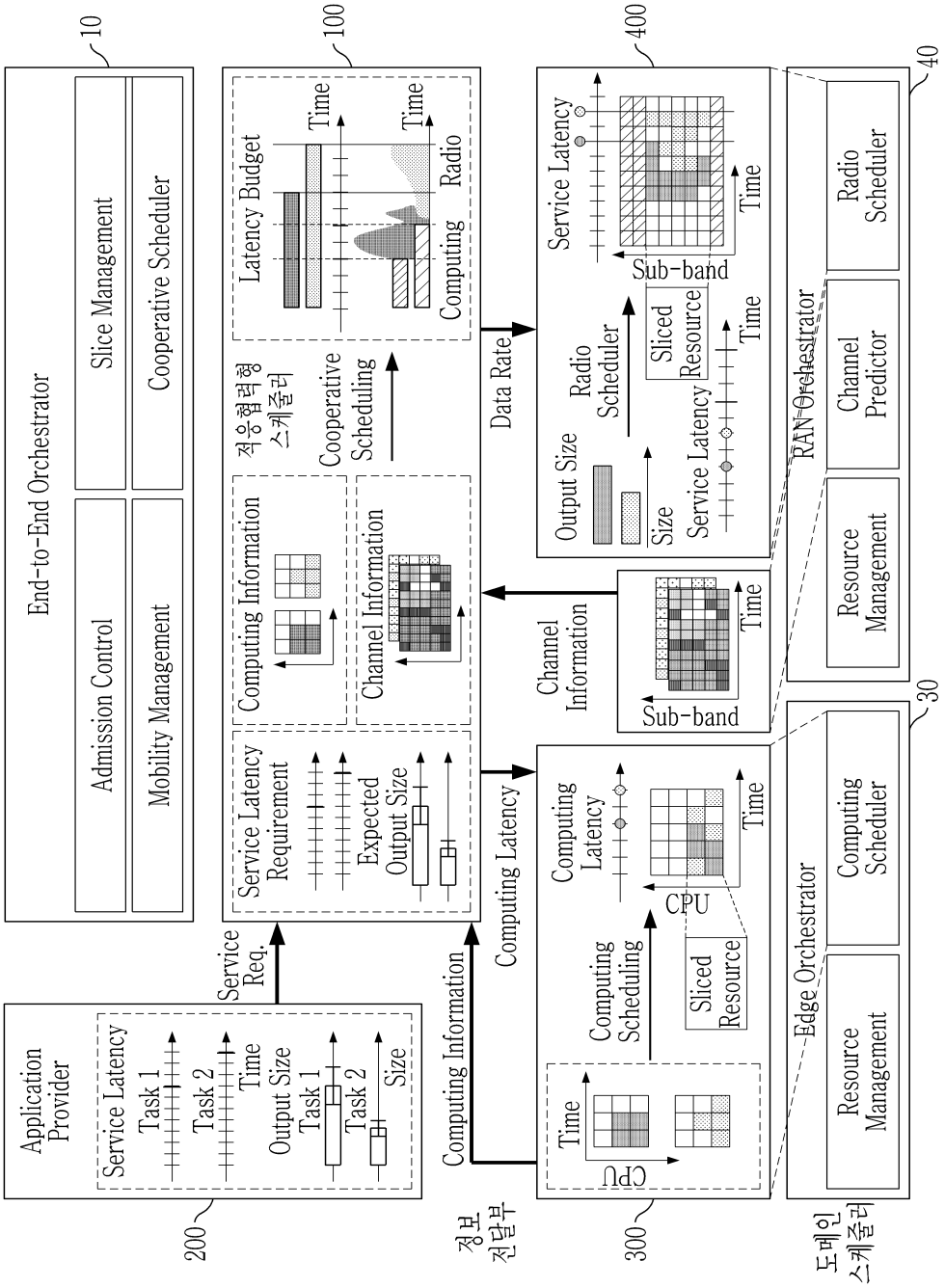
도면9


Computing Latency
Network Latency
Edge Server
Requested Procedure
Remote Procedure
RPC Module
RPC Response
ADU
Packets
ADU completion
Client (Mobile Device)
Service Application
Local Procedure
Call Procedure
Resume Execution
RPC Module
RPC Request
ADU
Packets
Application
Network

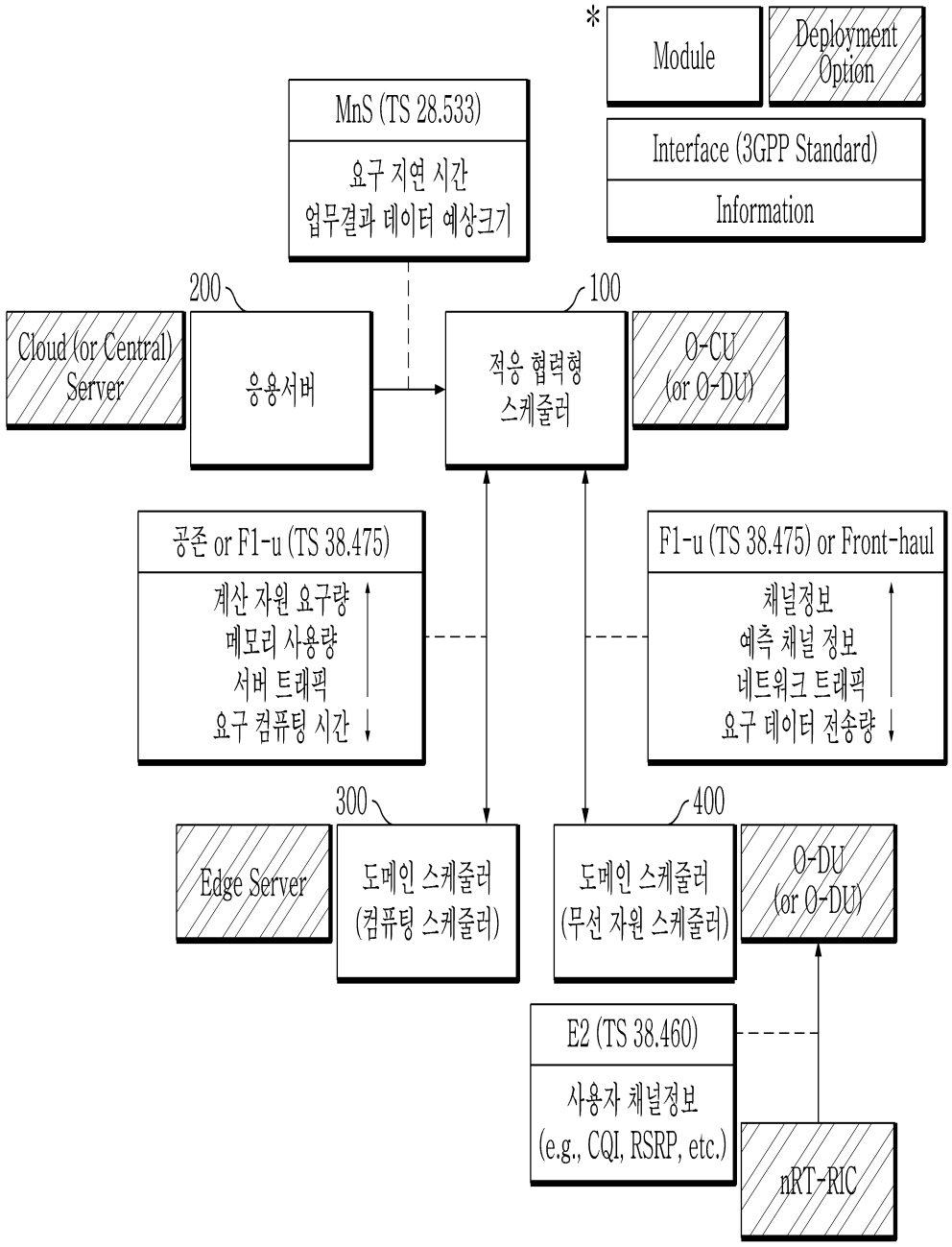
도면10



도면11


End-to-End Orchestrator
Admission Control
Mobility Management
Slice Management
Cooperative Scheduler
10
Application Provider
Service Latency
Task 1
Task 2
Time
Output Size
Task 1
Task 2
Size
200
Service Req.
Service Latency Requirement
Expected Output Size
Computing Information
Channel Information
적응협력형 스케줄러
Cooperative Scheduling
Latency Budget
Time
Computing
Time
Radio
100
Data Rate
Computing Latency
Channel Information
정보 전달부
Computing Information
Computing Latency
Time
CPU
Computing Scheduling
Sliced Resource
300
Edge Orchestrator
Resource Management
Computing Scheduler
30
도메인 스케줄러
Service Latency
Time
Sub-band
Output Size
Size
Radio Scheduler
Sliced Resource
Service Latency
Time
Time
Sub-band
400
RAN Orchestrator
Resource Management
Channel Predictor
Radio Scheduler
40

도면12


*
Module
Deployment Option
Interface (3GPP Standard)
Information
MnS (TS 28.533)
요구 지연 시간
업무결과 데이터 예상크기
200
100
Cloud (or Central) Server
응용서버
적응 협력형 스케줄러
O-CU (or O-DU)
공존 or F1-u (TS 38.475)
계산 자원 요구량
메모리 사용량
서버 트래픽
요구 컴퓨팅 시간
F1-u (TS 38.475) or Front-haul
채널정보
예측 채널 정보
네트워크 트래픽
요구 데이터 전송량
300
400
Edge Server
도메인 스케줄러 (컴퓨팅 스케줄러)
도메인 스케줄러 (무선 자원 스케줄러)
O-DU (or O-DU)
E2 (TS 38.460)
사용자 채널정보 (e.g., CQI, RSRP, etc.)
nRT-RIC