Copyright© 2024 SK Telecom. All Rights Reserved.

SK telecom

SK텔레콤 6G 백서

View on Future AI Telco Infrastructure



ICT Infra

Copyright© 2024 SK Telecom. All Rights Reserved.

SK텔레콤 6G 백서

View on Future AI Telco Infrastructure

ICT Infra

SK telecom

Contents

[Executive Summary] 03

1. 들어가는 말 04

2. 6G 구조 설계에 대한 고려사항 06

2.1 5G 비전에 대한 고찰 07
2.2 6G 프레임워크의 포괄적 측면, 유비쿼터스 인텔리전스 08
2.3 차세대 ICT 진화의 촉매, 인공지능 09

3. AI Telco 인프라에 대한 관점 10

3.1 AI 컴퍼니 비전 10
3.2 유무선 네트워크 진화: AI Telco 인프라 11

4. Telco Edge AI 13

4.1 왜 Telco Edge AI인가? 13
4.2 Telco Edge AI란 무엇인가? 14

5. Telco Edge AI 진화 방향성 18

5.1 생태계 조성 18
5.1.1 사업자/제조사 협력 18
5.1.2 표준화 및 학술 연구 활동 19
5.2 비즈니스 모델 발굴 20
5.3 주요 기술 22
5.3.1 무선접속망 23
5.3.2 코어망 24
5.3.3 전송망 25
5.3.4 디바이스 27
5.3.5 AI 오케스트레이션 28

6. 맺음말 32

References 33

Abbreviations 34

Change History 35

Executive Summary

[들어가는 말]

- 트래픽 수요 증가 → 수익 창출 → 인프라 투자의 사이클로 연결되는 이동통신사(Telco)의 비즈니스 모델은 성장세가 정체되어, Telco 인프라의 가치를 제고할 수 있는 AI 기반 Telco 인프라의 구조 혁신을 통한 새로운 고객 가치 창출 노력 필요

[6G 구조 설계에 대한 고찰]

- 5G는 모바일 광대역(eMBB), 초고신뢰 저지연(URLLC), 대규모기기 통신(mMTC) 등의 이전 세대보다 다양한 사용 시나리오 수용이 가능한 구조로 설계되었으나, 결과적으로 Connectivity 차원의 성능 진화에만 집중
- 6G 프레임워크 권고서는 6G가 갖춰야 할 포괄적 측면으로 유비쿼터스 인텔리전스의 중요성을 제시, 통신(Connectivity)과 AI의 균형적 고려가 중요

[AI Telco 인프라에 대한 관점]

- SK텔레콤은 AI 컴퍼니로 변모하기 위한 비전으로 AI 피라미드 전략을 공개,
  ▲ AI 인프라(1층) ▲ AIX(2층) ▲ AI 서비스(3층)에 대한 추진 방향을 제시
- AI 인프라 중 유무선 인프라는 ▲ 네트워크의 효율을 높이는 AI-Assisted Network
  ▲ 새로운 비즈니스 모델과 가치를 제공하는 AI-Enabling Infra [Telco Edge AI] 의 활용 측면을 제공

[Telco Edge AI]

- 미래 Telco 인프라는 Connectivity와 AI 컴퓨팅을 동시 제공하는 Telco Edge AI로 진화해, 대규모 AI DC와 On-Device AI 사이에서 차별화된 가치를 제공
- Telco Edge AI 인프라는 Telco가 보유한 상면과 시설을 활용하고 신규 장비를 도입해 구축하며, 구성 요소는 ▲ 상면(Co-location) ▲ AI 서버 ▲ AI-RAN이 있음

[Telco Edge AI 진화 방향성]

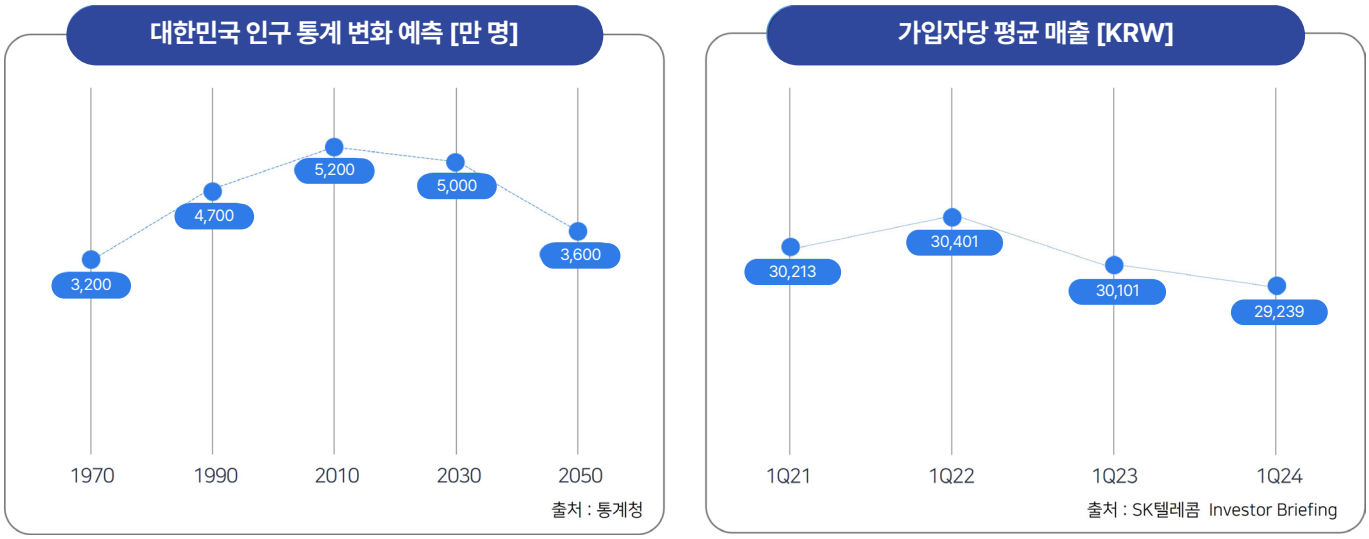
- 성공적 Telco Edge AI 진화를 위해, 제조사 · 사업자 · 고객 모두의 글로벌 생태계를 조성/확산하고, 유망 사업 모델을 발굴하는 것이 무엇보다 중요함
- SK텔레콤은 무선접속망, 코어망, 전송망, 디바이스, AI 오케스트레이션 등 기술 영역별 구현 방안을 구체화하고 이를 실행하기 위한 기술을 확보할 계획임

[맺음말]

- 긴 호흡이 필요한 Telco Edge AI로 성공적인 진화를 위해, 다양한 파트너 협력을 통해 의미 있는 성공 사례를 축적해 나감으로써 Telco 사업의 지속가능성을 높여가고자 함

1. 들어가는 말

이동통신 업계는 3G 상용화 이후 2000년대 후반 스마트폰의 등장과 4G 상용화를 거치며 본격적으로 개화했다. 모바일 인터넷 시대의 도래는 이동통신 산업이 한 단계 도약하는 계기가 되었지만, 한편으로는 통신 네트워크가 단순 데이터 전송 역할에만 머무는 덤파이프(Dumb Pipe)로 전락할 수 있다는 위기의식을 불러 일으키기도 했다. 예를 들어, 2003년 스카이프 출시, 2009년 출시된 왓츠앱과 같은 메시징 서비스의 출현, 2000년대 중반 유튜브와 넷플릭스 등의 동영상 스트리밍 서비스에 따른 이동통신 데이터 트래픽 폭증은 이동통신 생태계에 큰 도전 과제가 되고 있는 상황이다. 특히 국내의 경우, 인구 감소로 인한 신규 고객 확보는 더욱 어려워지는 반면 이동통신 비용 지출은 정체되고 있어, 기존의 세대교체(Generation Change) 형태의 이동통신망 진화 방식만을 고려할 경우 투자 여력이 부족해 이동통신 생태계 전반이 위축될 우려가 있다.



[그림 1] 대한민국 인구 통계 예측 및 SK텔레콤의 이동통신 가입자당 평균 매출

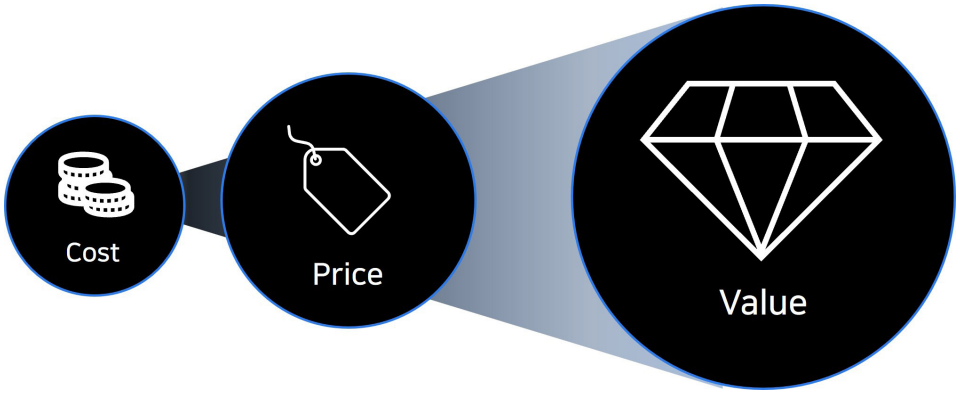
이를 타개하기 위한 노력으로 기존 B2C 중심의 이동통신 서비스 시장을 다양한 버티컬 영역으로 확장하기 위한 시도와 함께, 통신 이상의 가치(Beyond Connectivity)를 캐치프레이즈로 클라우드 서비스, 스마트 시티 솔루션[1], 사물인터넷 서비스, 디지털 광고 및 마케팅과 같은 통신을 활용한 인접 ICT 영역으로 진출하고 있다. 또한, 자율주행, 도심항공교통(UAM)과 같은 향후 상용화가 기대되는 첨단 영역의 신규 사업 개척을 위해서도 여러 파트너들과 다방면으로 협업하는 등 그 노력의 범위와 방향에 제한을 두지 않고 있다.

세상이 전례 없는 속도로 진화함에 따라, 더 진화되고 효율적인 통신 인프라에 대한 요구는 점점 증가하고 있다. 5G 는 이동통신의 광대역화에 중점을 두었던 4G에 비해 향상된 모바일 광대역(eMBB), 초고신뢰 저지연(URLLC), 대규모기기 통신(mMTC)을 목표로 이동통신이 더 다양한 사용 시나리오들을 수용할 수 있는 토양을 제공했으나, 여전히 이동통신 네트워크 자체의 속도나 용량 등 기술적인 성능을 진화하는데 중점을 두고 있다는 점에서는 3G, 4G와 같은 이전 세대 네트워크에서 지향했던 목표에서 크게 벗어나지 못한 것도 사실이다 [1].

[1] 외래어는 국립국어원 외래어표기법을 따르되 일부 관용적 표기를 절충했다. 영어 Solution의 정식 표기는 설루션이나 본 백서에서는 솔루션으로 표기한다.

한편, ICT 산업 전반은 이동통신 네트워크 인프라에 더 높은 수준의 통신 성능과 다양한 사용 시나리오를 요구하고 있으며, 이를 위해서는 더욱 진보된 통신 인프라가 중요해질 전망이다 [2]. 그리고 네트워크 인프라의 효율성 향상과 지속가능성 제고는 물론, 네트워크 인프라의 상면, 컴퓨팅 자원 등을 활용한 새로운 수익 창출 등 다양한 이유를 충족하기 위해 네트워크 인프라에 AI를 적용하고, 네트워크 인프라의 자원을 활용해 AI 서비스를 제공하기 위한 연구와 투자도 점차 활성화되는 추세다.

이와 같이, Telco는 인프라 진화와 AI를 통한 비즈니스 고객 가치 창출을 추구하는 방향으로 진화하고 있으며, 이를 위해서는 Telco 인프라에 AI 혁신 기술을 접목해 고객의 가치 및 신규 사업 모델 창출을 꾀해야 한다. 즉, 인프라 구축/운용 비용 및 지불 가격을 넘어서는 가치를 제공할 수 있어야 하고, 이를 가능케 하기 위해서는 고객이 느끼는 인프라의 가치를 훨씬 더 높이려는 노력이 매우 중요하다.



[그림 2] Telco 인프라 지향점(비용 < 가격 < 가치)

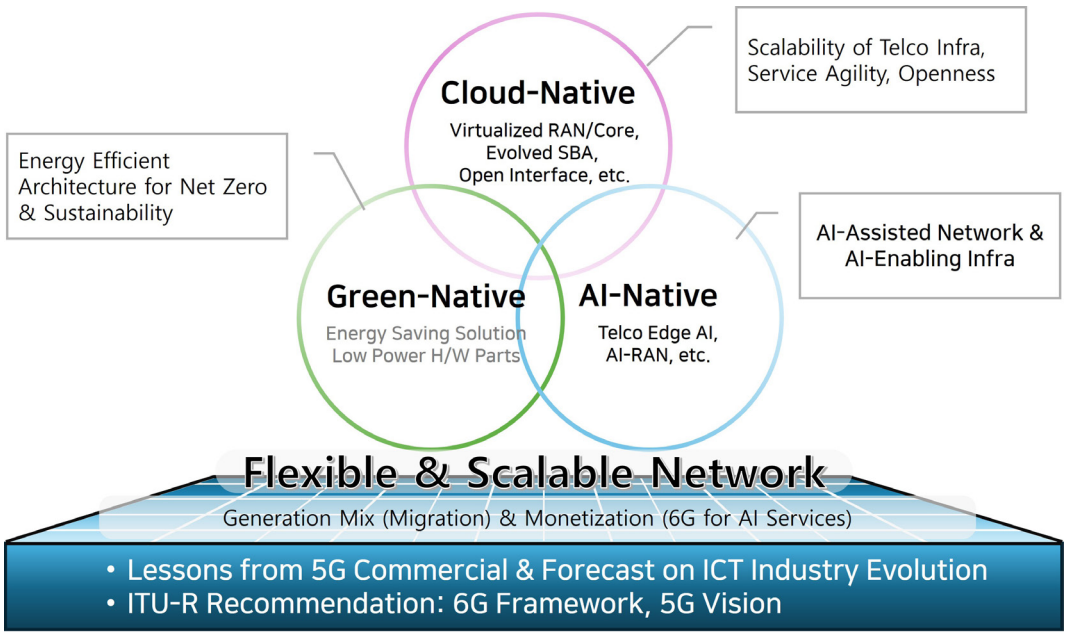
특히, 최근 ChatGPT가 주도하고 있는 생성형 AI가 촉발한 인공지능 서비스는 일상 생활의 소소한 개인 비서 앱 수준에서 벗어나 본격적인 생산성 향상 도구로서 비용 대비 높은 가치를 제공해 줄 것으로 기대되고 있어 사무실이나 연구실, 산업 현장 등에서 그 분야를 가리지 않고 도입되고 있다. 이에 따라, 기존의 하이퍼스케일러들이 AI 서비스에 필요한 학습과 추론에 특화된 고성능 칩셋을 채용한 AI DC와 같은 인프라를 구축하며 기존의 데이터 센터와의 차별화된 가치를 제공하기 위한 경쟁을 가속화하고 있는 반면, 이동통신 업계는 대규모로 시설되어 있는 네트워크 인프라를 통신 서비스(Connectivity) 용도만으로 활용하고 있는 실정이다.

본 백서에서는 미래 이동통신 네트워크 인프라의 중요한 측면으로서 6G의 전체 네트워크 인프라가 AI-Native 특성을 기반으로 [1], 현재 한정된 용도만으로 활용되고 있는 네트워크 인프라의 구조를 혁신해 고객 접점에 위치한 서비스 제공의 Edge에서 AI DC와 비교해 차별화된 가치를 제공함으로써 새로운 수익 창출과 통신 인프라의 가치를 재고(Re-rating) 할 수 있는 Telco Edge AI 인프라로의 진화 방향을 제시하고자 한다.

2. 6G 구조 설계에 대한 고려사항

국제전기통신연합 전파통신부문(이하 ITU-R)에서는 향후 6G 개발을 위한 프레임워크와 전반적인 목표를 제시한 권고서인 ITU-R M.2160 문서를 '23년 11월 승인하였다 [3]. 이 권고서는 공식적인 6G 표준화 활동의 첫 번째 이정표가 되는데, 이 권고서에서 제시한 6G 시대가 수용해야 할 사용 시나리오(Usage Scenarios)와 성능 지표(Capabilities)에 따라 6G 네트워크의 기술 성능 요구사항이 설정되며, 3GPP와 같은 표준 개발 기구에서 실질적인 6G 규격을 개발하는 레퍼런스가 되기 때문이다.

SK텔레콤은 현재의 이동통신 네트워크 인프라를 6G로 진화시키는 데 있어 중요한 6G 아키텍처 설계의 중요 요소들을 검토 중으로, 아래 그림과 같이 5G 비전 및 6G 프레임워크 권고서와의 연계성, 현재 운용 중인 5G로부터의 교훈과 미래 네트워크 발전 방향에 대한 전망 등을 전반적으로 고찰해 6G 네트워크 인프라의 구조 설계 시 반영되어야 할 중요한 요소로 Cloud-Native, Green-Native, AI-Native와 이들 모든 요소들의 기반이 되는 Flexible & Scalable Network로의 진화를 검토하며 관련 연구를 진행하고 있다 [1].



[그림 3] Overall 6G Architecture 설계 시 주요 고려 요소

첫째, Cloud-Native는 기존의 물리적인 통신망 환경에서 벗어나 클라우드 기술을 기반으로 통신 서비스를 제공함으로써, 빠르게 변화하는 요구사항에 유연하게 대응하고, 높은 가용성과 확장성 및 운용 비용 절감을 달성하는 데 필수적으로 고려되어야 할 차세대 네트워크 인프라 구조 설계 시 설계 원칙이다.

둘째, Green-Native는 전 세계적 화두인 기후변화에 대응하기 위한 탄소 배출량 절감에 동참함은 물론, 이동통신 사업의 지속 가능성 확보를 위해 필수적으로 고려되어야 할 미래 네트워크 인프라 구조에 반영되어야 할 설계 원칙이다. 이에 대한 무선 접속망, 코어망 인프라에서 고려되어야 할 기술 요인들은 2023년 6G 백서에 기술되어 있다 [1].

셋째, AI-Native는 무선접속망/코어망/전송망 등 네트워크 인프라 전 영역에 걸쳐 인공지능을 융합함으로써 네트워크의 성능, 효율성, 안정성을 확보하는 데에 고려되어야 할 차세대 네트워크 인프라 구조 설계 시 설계 원칙이다 [1], [4]. 이는 단순히 AI 기술을 부분적으로 도입/적용하는 것을 넘어, 네트워크 인프라 전 영역에 걸쳐 AI를 기반으로 고객 경험 개선 및 운용상의 안정성과 효율성을 극대화시킴과 동시에 인프라 자산을 활용해 AI 서비스를 제공함으로써 수익을 창출하는 것까지 포함될 수 있다.

한편, 위 3가지 요소들의 기반이 되는 6G 구조 설계 원칙으로 유연하고 확장성이 있는 네트워크 구조(Flexible & Scalable)를 주목하고 있다. 이에 대한 키워드 중 하나인 Generation Mix는 현 이동통신 시스템에서 향후 6G로의 진화 과정에서 전체 네트워크 인프라를 Migration해 나갈 방안을 고려하는 것으로서, 차세대 이동통신의 기술 성능 향상에만 치중해 이전 세대와의 연계성에 대한 고려가 미흡했던 Clean Slate 구조를 지향한 과거 이동통신 세대 진화 방식인 Generation Change에 대한 교훈으로부터 SK텔레콤이 중요하게 연구하는 주제이다. Generation Mix는 기존 네트워크 시스템과의 공존을 통한 Smooth Migration은 물론, 차세대 이동통신 기술인 6G 서비스에 대한 수요 기반으로 인프라를 구축하는 Demand-based Deployment를 가능하게 함으로써 비용 효율적 투자를 통한 이동통신 사업의 지속가능성을 높이기 위한 핵심 고려 요소이다.

본 백서에서는 6G 네트워크 아키텍처의 중요 요소 중 AI-Native로 진화하기 위해 필수적인 요소로 당사에서 주목하고 있는 Telco Edge AI에 대한 우리의 관점을 제시한다.

2.1 5G 비전에 대한 고찰

2015년 ITU-R에서 발간한 5G 비전 권고서는 2020년 이후에 운용될 이동통신 시스템 개발을 위한 프레임워크와 목표를 제시하였는데, 대표적으로 eMBB, URLLC, mMTC의 3가지 사용 시나리오와 주요 성능 지표들이 정의되어 있다 [5]. 하지만 이러한 주요 성능지표들은 전통적인 이동통신 자체의 성능 목표를 정의하는 데 머물러 있음을 알 수 있다.

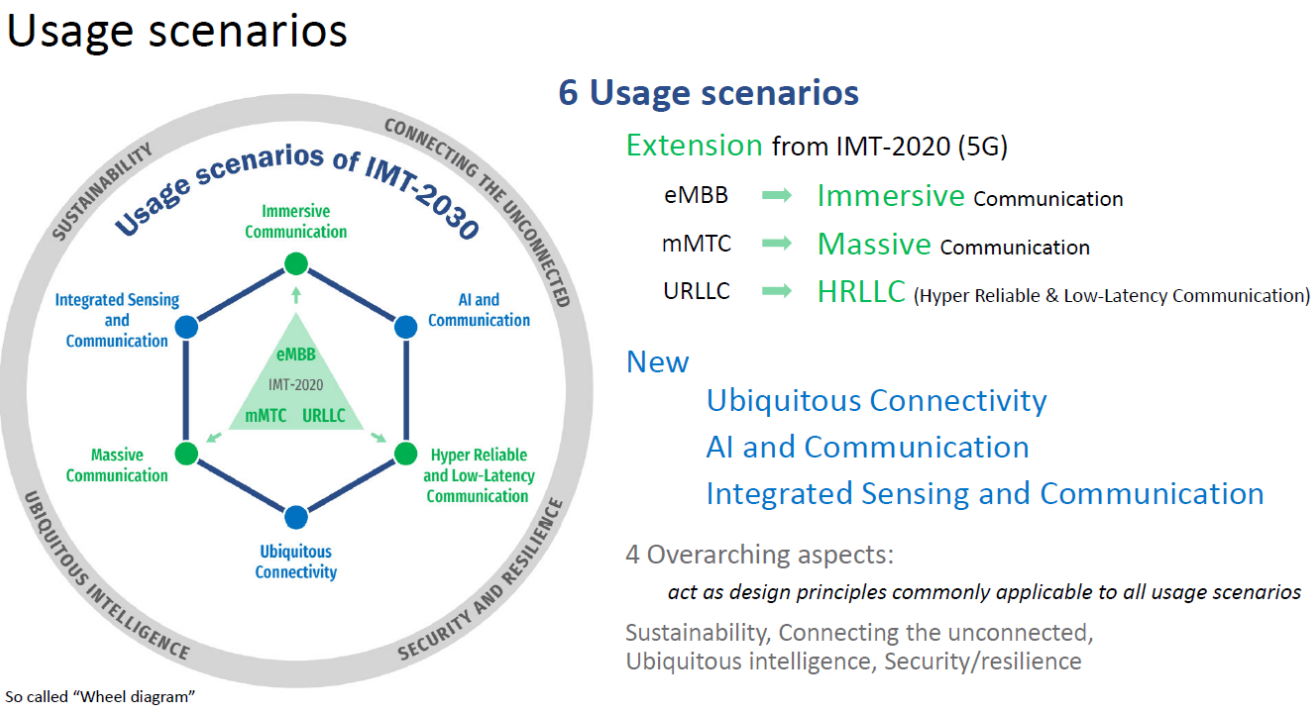
또 다른 측면에서, 2015년 5G 비전을 공표할 당시에 전망했던 2020년 이후의 미래상은 예측보다 더욱 빠른 속도로 변화하고 있는데, 2016년 알파고와 이세돌이 벌인 세기의 바둑 대국은 딥러닝과 강화학습을 기반으로 한 인공지능 활용의 가능성을 확인한 계기였다면, 2023년 거대 언어 모델을 기반으로 하는 생성형 인공지능 ChatGPT의 등장은 ICT 생태계 패러다임의 근간을 인공지능이 좌지우지할 정도로 혁신적인 변화를 이끌고 있다 [6]. 인공지능은 5G 비전을 수립할 당시에 모든 사람이 품었던 예상을 훨씬 뛰어넘는 속도로 우리에게 이미 다가와 있다.

5G 도입을 앞두고 이동통신 업계에서 이동통신 자체의 성능을 높이는 데에만 주력한 것은 어쩌면 이러한 빠른 변화를 예상치 못한 필연적 결과일 수 있다. 하지만, 이미 우리 모두의 앞에 다가와 있는 인공지능의 시대에 맞는 새로운 이동통신 네트워크 인프라의 발전 방향에 대해 진지하게 고민하기에 지금도 늦지 않았다는 자세로 준비해야 할 것이다.

이에 현재의 5G 네트워크 인프라가 AI-Native 구조로 진화하는 데 갖고 있는 한계를 극복하고자 6G 프레임워크 권고서에서 제시한 방향성을 바로 뒤이어 간략히 짚어 보고, 이를 통해 향후 6G 네트워크 인프라가 갖추어야 할 구조의 중요한 특성으로 인공지능과 관련한 측면을 살펴보고자 한다.

2.2 6G 프레임워크의 포괄적 측면, 유비쿼터스 인텔리전스

우선 ITU-R에서 지난해 11월 승인한 6G 프레임워크 권고서 중 5G 비전과의 비교를 직관적으로 확인할 수 있는 사용 시나리오를 살펴보도록 한다 [3].



[그림 4] IMT-2030의 사용 시나리오 및 포괄적 측면

* 출처: ITU 홈페이지: IMT towards 2030 and beyond (itu.int)

ITU-R에서 제시한 6G의 사용 시나리오는 5G로부터 확장된 3개 시나리오를 포함하면서, 새롭게 3개 시나리오가 추가되었는데, 이 중에서 AI와 커뮤니케이션 시나리오를 눈여겨볼 수 있다. 더 주목할 부분은 이 모든 6개 사용 시나리오에 공통적으로 적용될 설계 원칙으로서 4개의 포괄적 측면(Overarching aspects)을 제시했는데, 그중에서 유비쿼터스 인텔리전스를 통해 6G 네트워크가 어떻게 5G와 차별화되어야 하는지 설명하고 있다.

권고서에서 설명하는 유비쿼터스 인텔리전스를 간략히 소개해 보면 다음과 같다.

인공지능(AI)과 특히 머신러닝(ML) 기술의 지속적인 발전과 빠른 확산으로 인해, 통신 시스템의 모든 부분에 인텔리전스가 존재할 것으로 예상된다. 이는 스마트 시티와 커뮤니티 구축을 지원할 것이다. 미래의 연결된 기기들은 보다 직관적이고 효율적인 상호작용을 위해 완전히 컨텍스트를 인식하게 될 가능성이 있다. AI/ML을 통한 네트워크의 자율 관리가 가능해지면, 어느 정도 인간의 개입 없이 자체 모니터링, 자체 조직화, 자체 최적화 및 자체 치유를 수행할 수 있을 것이다.

IMT-2030은 지능형 응용 프로그램을 위한 서비스를 제공할 수 있는 AI 지원 인프라로 기능할 수 있다. AI 지원 무선 인터페이스와 분산 컴퓨팅 및 인텔리전스를 통해 엔드 투 엔드 AI 적용 가능성과 통신 및 컴퓨팅의 융합이 가능해질 것이다. 이러한 시스템은 추론, 모델 훈련, 모델 배포, 그리고 네트워크와 기기에 분산된 컴퓨팅 기능을 갖추게 될 것이다.

요약하자면, AI를 활용한 네트워크 인프라의 지능화/효율화와 함께, AI 응용 서비스를 제공할 수 있는 네트워크 인프라라는 두 가지 큰 목적성을 갖는 것으로 설명하고 있다. 즉, 앞으로 본 백서에서 다루어질 AI-Assisted Network와 AI-Enabling Infra로의 진화가 6G 네트워크 구조의 핵심 요소 중 하나가 되어야 하는 이유가 될 것이다.

2.3 차세대 ICT 진화의 촉매, 인공지능

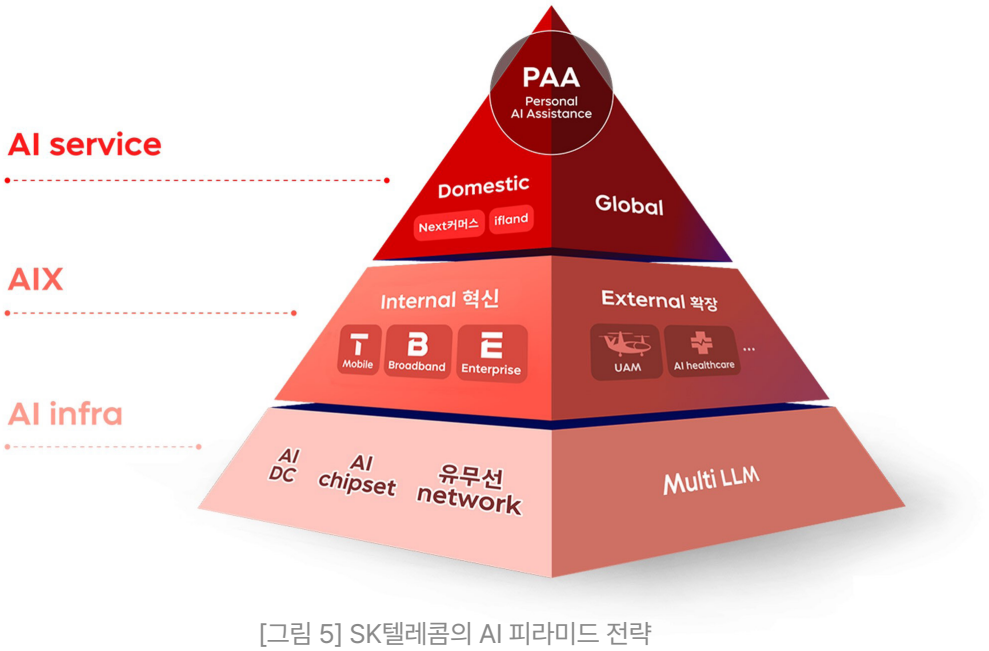
AI는 차세대 정보통신기술(ICT)에서 중요한 촉매 역할을 할 것으로 기대된다. AI-Native 네트워크는 AI가 통신 네트워크의 관리와 운영에 깊이 통합된 형태로, 자율적이고 지능적인 네트워크 관리를 가능하게 한다. 구체적으로 AI-Native 네트워크는 자율 네트워크 관리, 지능형 트래픽 관리, 예측 및 분석, 네트워크 보안 강화, 에너지 효율 관리 등의 기능을 포함할 수 있다. AI는 네트워크 상태를 실시간으로 모니터링하며, 이상 탐지와 자동 조정을 통해 네트워크 성능을 최적화한다. 또한 AI 알고리즘을 통해 데이터 트래픽을 효율적으로 분배해 혼잡을 최소화하고, 서비스 품질(QoS)을 보장한다. AI는 네트워크 트래픽 패턴을 예측하고 잠재적 문제를 사전에 파악해 대응할 수 있으며, 사이버 공격 탐지 및 대응 시스템을 강화해 실시간으로 위협을 식별하고 차단한다. 또한 AI는 네트워크 인프라의 에너지 소비를 최적화해 지속가능성을 높인다. AI의 이러한 다양한 쓰임새 때문에, 전 세계의 주요 ICT 기업들은 AI-Native 네트워크의 중요성을 인식하고 적극적으로 투자와 개발을 추진하고 있다.

시장조사업체 Value Market Research에 따르면 2030년까지 이동통신 시장에서 AI의 수요는 200억 달러에 이를 것으로 추산되고 있다 [7]. 2022년에 조사한 20억 달러 규모에서 10배 가량의 기하급수적인 성장이 예상된다. 세계 이동통신 네트워크 장비 시장은 2024년 현재 전체 480억 달러로 추산되며, 그 중에서도 무선 접속망을 위한 네트워크 장비 시장은 360억 달러 정도로 매우 큰 비중을 차지한다 [8]. 따라서 현재의 네트워크 장비가 AI-Native 네트워크로 교체된다면 이는 차세대 ICT 시장의 판도를 변화시킬 큰 파장을 불러일으킬 것이다.

결론적으로 AI는 차세대 통신 인프라의 핵심 요소로, 자율적인 네트워크 관리, 성능 최적화, 운영 비용 절감, 사용자 경험 개선 등 다양한 분야에서 중요한 역할을 담당한다. AI-Native 네트워크로의 진화는 통신 인프라의 효율성과 신뢰성을 극대화하며, 미래의 혁신적인 ICT 서비스와 애플리케이션을 가능하게 할 것이다.

3. AI Telco 인프라에 대한 관점

인공지능(AI)이 기업의 미래를 좌우하는 핵심 기술로 부상함에 따라, 이러한 변화에 대응해 글로벌 AI 기업으로 도약하기 위한 AI 컴퍼니 비전을 수립하였다. 본 장에서는 AI 컴퍼니 비전의 핵심인 AI 피라미드 전략과, 그 전략을 구성하는 초석인 AI 인프라로 진화하기 위한 SK텔레콤의 관점을 제시한다.



[그림 5] SK텔레콤의 AI 피라미드 전략

3.1 AI 컴퍼니 비전

2023년 9월, SK텔레콤은 'AI 피라미드 전략'을 통해 AI 기술을 고도화하고, 이를 통해 글로벌 AI 리더로 자리매김하기 위한 청사진을 공개했다. AI 피라미드 전략은 AI 기술을 고도화하고 AI 서비스를 통해 고객과의 관계를 더욱 밀접하게 만들며, AI 얼라이언스 중심의 협력을 통해 실행해 나갈 계획이다.

AI 인프라, AIX(AI Transformation), AI 서비스의 세 가지 주요 영역으로 구성되어 있는 AI 피라미드 전략의 각 영역은 SK텔레콤의 핵심 비즈니스와 긴밀히 연결되어 있으며, 이를 중심으로 산업과 생활 전 영역의 혁신을 이루는 것을 목표로 한다.

1) AI 인프라
AI 인프라는 SK텔레콤의 기술 역량이 집약된 부분으로, Telco LLM(Large Language Model), AI 유무선 네트워크, AI 데이터 센터 등을 포함한다. SK텔레콤은 글로벌 텔코 AI 연합(GTAA)을 통해 통신사 특화 AI 모델 개발과 협력을 추진하고 있다. 또한 AI 데이터 센터에 대한 투자와 협력을 통해 유무선 인프라에 AI를 적용함으로써 네트워크의 효율성과 안정성을 강화하고자 한다.

2) AIX (AI Transformation)
AIX는 SK텔레콤의 주요 사업 영역에 AI를 적용해 생산성과 고객 경험을 혁신하는 부분을 담당한다. AI 기술을 활

용해 모바일, 브로드밴드, 기업 솔루션 등의 비즈니스 운영을 효율화하고, 새로운 성장 동력을 발굴하고 있는데, AI 기반 고객센터의 혁신, 맞춤형 방송 서비스 'AI Btv'의 도입, 기업 맞춤형 AI 솔루션 제공 등을 통해 AIX를 실현하고 있다. 또한, 도심항공교통(UAM)과 AI 헬스케어 분야에서도 AI 기술을 활용한 혁신을 적극 추진 중이다.

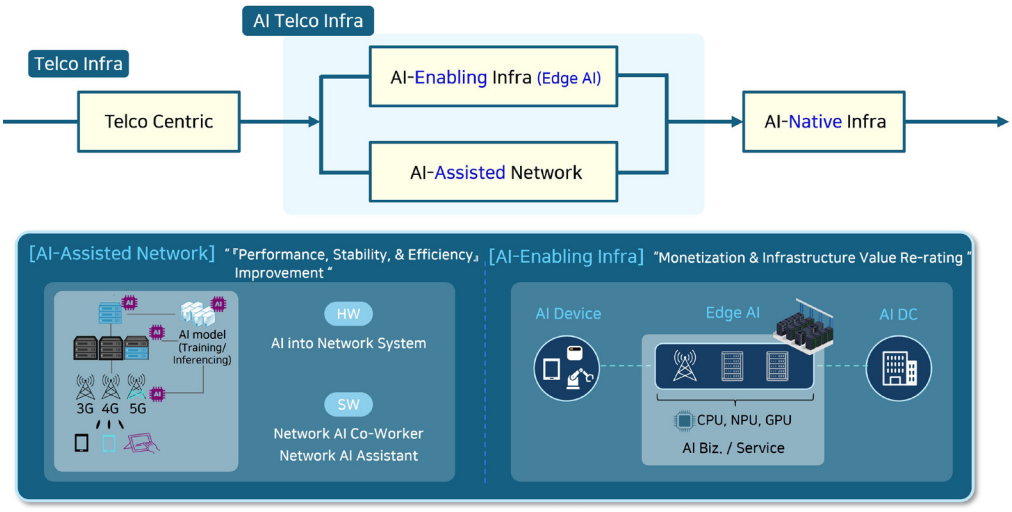
3) AI 서비스

AI 서비스는 SK텔레콤의 AI 기술을 바탕으로 글로벌 시장에서 경쟁력을 확보하기 위한 전략이다. SK텔레콤은 개인 AI 비서(PAA) 'A.(에이닷)'을 통해 사용자 경험을 혁신하고 있으며, 이 서비스를 글로벌 시장으로 확장해 AI 에이전트 분야에서의 입지를 넓혀 글로벌 AI 시장에서의 리더십을 확보하고자 한다.

SK텔레콤은 AI 피라미드 전략의 성공적 실행을 위해 'T.H.E. AI'라는 AI 거버넌스 원칙을 제정했다. 'T.H.E.'는 'by Telco, for Humanity, with Ethics'의 약자로, 이는 AI 기술이 고객과 이해관계자를 상호 연결하고(Telco), 다양성과 포용, 인류 복지 증진을 목표로 하며(Humanity), 투명하고 윤리적인 AI 의사결정 과정을 보장하는(Ethics) 원칙을 담고 있다. 이러한 AI 거버넌스 원칙은 SK텔레콤의 AI 사업 전반에 걸쳐 신뢰성과 안정성을 확보하기 위한 지침을 제공하며, SK텔레콤은 이를 통해 신뢰할 수 있고 투명하며 안전한 AI를 개발하고자 한다.

3.2 유무선 네트워크 진화: AI Telco 인프라

SK텔레콤은 과거의 전통적인 통신 서비스 전용 용도의 인프라를 벗어나 AI 유무선 인프라로의 진화를 추진 중이며, 최종적으로는 AI-Native 인프라로의 전환을 목표로 하고 있다. 이러한 인프라 진화는 기존과는 다른 방식의 접근이 필요하며, Telco 인프라와 AI 핵심 인프라의 변혁을 추구하고 있다. AI Telco 인프라는 크게 활용 측면에 따라 AI-Assisted Network와 AI-Enabling Infra로 구분할 수 있다.



[그림 6] AI Telco 인프라의 두 가지 활용 측면

1) AI-Assisted Network

AI-Assisted Network는 이동통신 네트워크의 품질 향상 및 운용상의 안정성과 효율성을 극대화하는 개념이다. 이는 네트워크 시스템에 시뮬레이션, Vision AI 등 AI 기반의 다양한 솔루션들을 도입함으로써 투자 및 자산 관리, 운용 현장 네트워크 내 이상 징후 감지 등에 적용하는 것을 의미한다. 또한 전국적으로 발생하는 현장 이슈에 대한 최적화를 통해 고객 체감 품질 강화 업무에 기여할 수 있다.

고객 체감 품질 저하가 일어나지 않는 강한 네트워크가 만들어지기 위해서, 고객이 체감할 수 있는 모든 세부 지표까지 AI를 통해 감시, 장애 및 과부하 등 어떠한 문제 발생 시에도 즉시 조치, 복구할 수 있어야 한다. 또한, In-Network Computing 기술을 도입해 대용량 데이터 트래픽 처리시 에도 효율적인 Deep 데이터 수집/분석이 가능해 AI 기반 이상 탐지(Anomaly Detection) 및 근본 원인 분석(Root Cause Analysis)을 효율적으로 활용할 수 있다.

2) AI-Enabling Infra(Telco Edge AI)

AI-Enabling Infra는 이동통신 사업자의 인프라 자산을 활용해 디바이스와 AI 데이터 센터를 연결하는 역할을 수행한다. 최근 On-Device AI 기술의 보급으로 디바이스 단에서의 추론 위주 AI 서비스가 활성화되는 동시에, 초거대 규모의 AI 데이터 센터에서의 대규모 데이터 기반 학습 모델 개발 또한 병행되는 양상을 보인다. 이 둘 사이에서 이동통신 사업자가 보유한 네트워크 인프라와 컴퓨팅 자원 등의 자산을 활용한 B2B/B2C 서비스를 제공함으로써 새로운 수익화와 인프라의 가치를 재창출하는 개념을 Telco Edge AI로 명명한다.

Telco Edge AI의 목표는 인프라 진화와 AI 비즈니스의 고객 가치 창출로 나눌 수 있다. 인프라 진화 측면에서는 구현 방안을 구체화하고, 주요 기술 및 차별화 기술을 발굴하며, 글로벌 생태계를 조성하고 이끌어가는 것을 포함한다. AI 비즈니스 고객 가치 창출 측면에서는 Edge AI 인프라 비전의 구체화를 통한 인프라 가치 증대, Telco 자산 활용 극대화, 그리고 새로운 비즈니스 모델의 발굴 및 구체화를 추진한다.

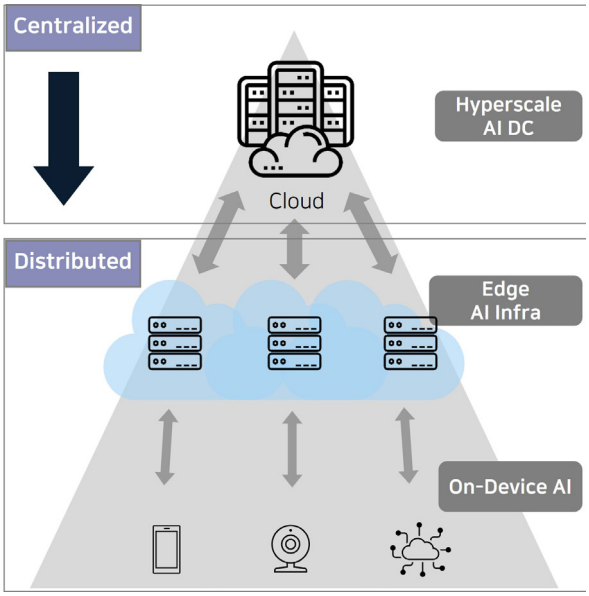
SK텔레콤은 Telco 인프라의 Edge AI 혁신을 통해 고객 가치 및 신규 비즈니스 창출을 위한 노력을 지속 중이며, 통신 네트워크의 시의적절한 진화 방향성을 주도적으로 제시하고자 한다.

4. Telco Edge AI

4.1 왜 Telco Edge AI인가?

AI 기술의 발전과 서비스에 대한 수요가 폭발적으로 증가하면서 고성능 AI 컴퓨팅에 대한 수요 또한 급증하고 있다. 더불어 향후 AI 서비스가 본격화될 경우 대규모의 컴퓨팅 파워뿐만 아니라 저지연 추론 및 학습에 대한 수요가 커질 것으로 전망되며, 이러한 수요와 요구 조건을 고려해 AI 서비스를 제공하기 위한 AI Telco 인프라 또한 진화되어야 한다. 전통적인 AI Telco 인프라는 클라우드 기반 대규모 AI 데이터 센터(Hyperscale AI DC)를 구축하는 중앙 집중 형태였으며, 이는 고성능 GPU 기반 대규모의 컴퓨팅 파워를 제공할 수 있어 초거대의 고성능 AI 모델을 학습하고 서비스하는 데 적합한 반면, 상면 및 전력 등 높은 유지비용과 중앙 집중 형태의 숙명인 서비스 속도의 지연, 그리고 보안 취약 상황 발생 시 정보 유출의 피해가 상대적으로 크다는 단점을 가지고 있다.

최근 모바일 영역에서는 On-Device AI 기술 및 이를 활용한 AI 서비스가 화두가 되고 있다. 그 배경으로는 모바일 Application Processor (AP)의 AI 추론 성능이 지속적으로 개선되고 있고, 또한 AI 모델 경량화 기술이 발전해 특정 응용 분야에서 거대 모델과 견줄 만한 성능을 갖는 소형 모델이 등장하였다는 점을 들 수 있다. 이에 모바일 단말의 제한적 환경에 특화된 AI 서비스를 자체적으로 제공할 수 있게 된 것이다. 향후에도 On-Device AI는 스마트폰 및 신규 AI 디바이스 활용 시 고객 경험 향상을 위한 주요 도구로써 진화할 것으로 예상된다.



[그림 7] AI 인프라 개념

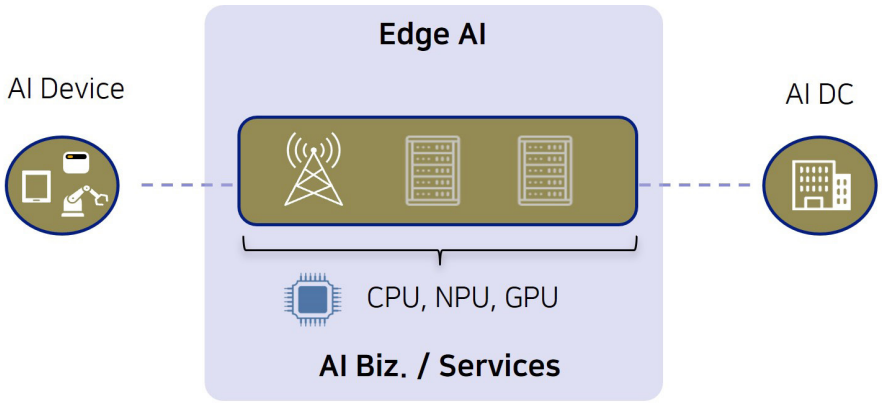
SK텔레콤은 미래 Telco 인프라가 대규모 AI 데이터 센터와 On-Device AI 사이에서 각각의 장점을 살려 상호 보완적이며 차별화된 AI 서비스를 제공할 것으로 전망한다. 기존의 Telco 인프라가 Connectivity 제공과 고도화를 위해 발전해 왔다면, 미래 Telco 인프라는 Connectivity 제공과 동시에 AI 컴퓨팅 파워를 제공함으로써 중앙 집중화된 대규모 AI 컴퓨팅을 사용자에 더 가깝게 분산 배치해 상대적으로 저지연 및 높은 보안성의 AI 서비스를 제공하는 Telco Edge AI[2] 인프라로 진화하는 것이다.

[2] Edge AI는 기존의 중앙집중형 AI 데이터 센터와 구분되어, 데이터에 근접한 네트워크 끝단(edge)에서 AI 연산이 수행되는 형태를 지칭하고, On-Device AI도 포함하는 개념이지만, 본 백서에서의 Edge AI는 대규모 AI 데이터 센터 / On-Device AI와 구분되어, 통신 인프라에 위치하면서 AI 서비스를 제공하는 형태를 의미한다.

구분	On-Device AI	Telco Edge AI 인프라	Hyperscale AI DC
위치	단말	통신 인프라	중앙 집중
구성	Low-End GPU/NPU/CPU	xPU[3] (GPU, NPU, CPU 등)	High-End GPU
규모	단말 내장	소규모 서버 랙 통신 장비 내장	대규모 서버 랙
활용 용도	소규모 AI 서비스 (일부 추론)	적정 규모 AI 서비스 + 통신 서비스	대규모 AI 서비스
Low Latency, Security, Agility	높음 (단말내 직접 처리)	중간	낮음 (다수 노드 경유)

[표 1] On-Device AI, Telco Edge AI, Hyperscale AI DC의 차이점

4.2 Telco Edge AI란 무엇인가?

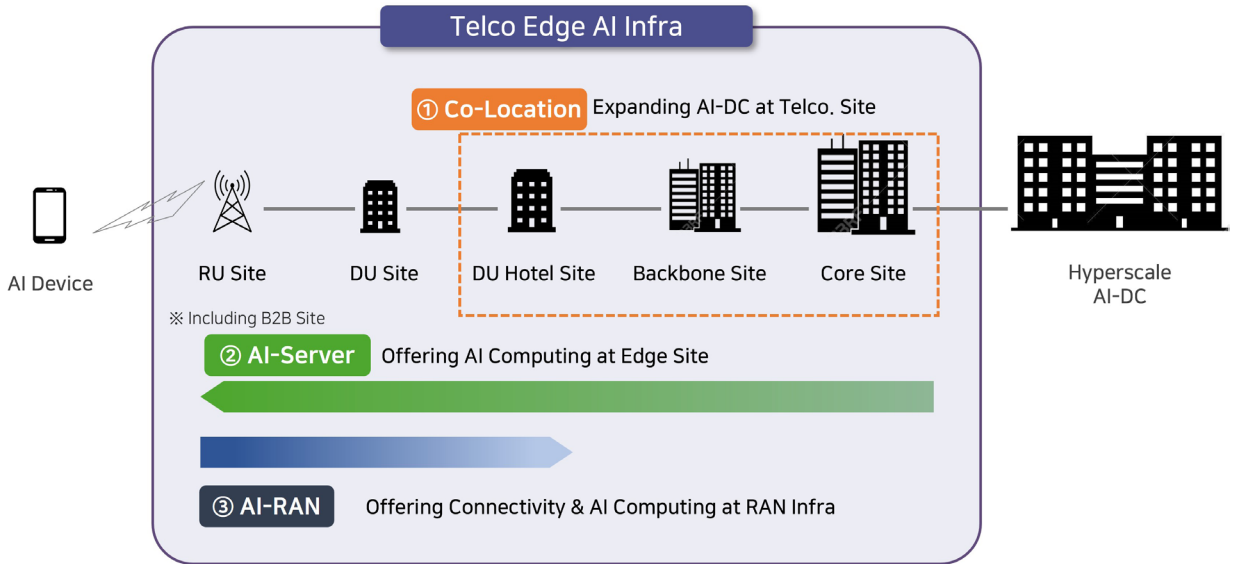


[그림 8] On-Device AI와 AI DC 사이의 Edge AI

[3] xPU: 각기 다른 처리 요구에 최적화하기 위한 목적으로 특화된 명령 세트/아키텍처 기반해 설계된 전용 처리 장치(예: CPU/GPU/NPU)

Telco Edge AI 인프라는 이동통신 사업자가 가진 인프라의 상면 및 시설, 장비를 활용해 기존에 제공하던 통신 서비스뿐 아니라 AI 서비스에 요구되는 AI 컴퓨팅 파워를 함께 제공해, Connectivity 중심의 전통적 Telco 인프라에 새로운 가치를 더한 인프라이다.

사업자는 통신 서비스를 제공하기 위한 시설을 전국적으로 구축해 운용하고 있으며 이를 한 번에 Telco Edge AI 인프라로 교체하는 것은 사실상 어렵다. 따라서 SK텔레콤은 Telco Edge AI 인프라를 사업과 서비스 특징에 따라 다음의 세 가지 영역으로 구분한다.



[그림 9] Telco Edge AI의 세 가지 영역

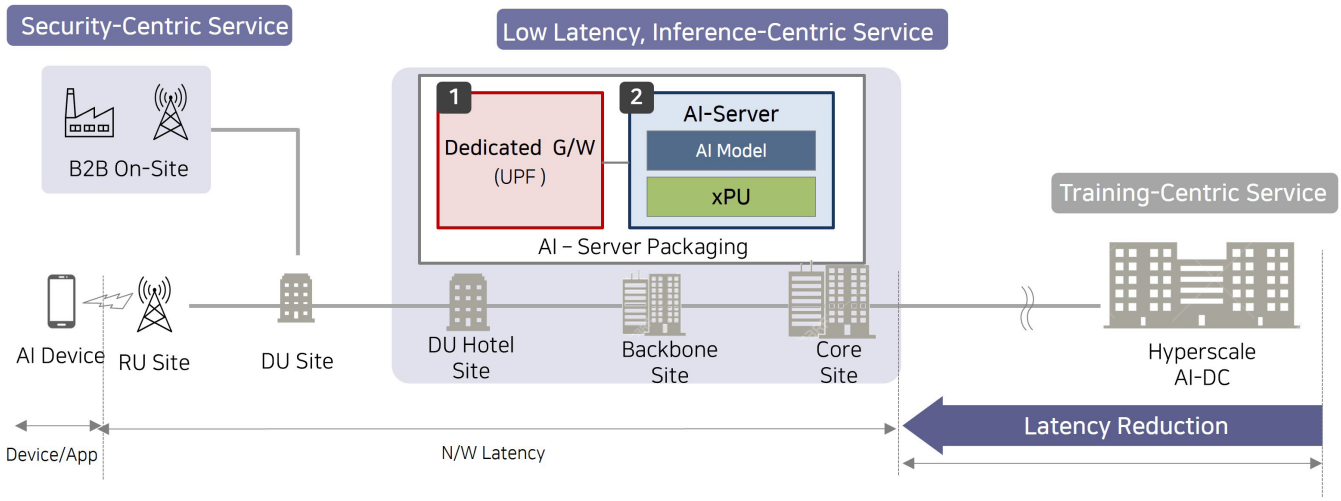
1) 상면(Co-location)

첫 번째 영역은 "상면"으로, 이동통신 사업자가 가지고 있는 통신 국사의 유휴 공간을 AI 컴퓨팅을 제공하기 위한 상면 인프라로 바꾸고 비용 효율적으로 AI서비스를 제공하는 것이다. AI 컴퓨팅 파워를 제공하는 xPU 기반 서버들은 전통적인 통신 장비들보다 발열과 전력 소모량이 높아 통신 국사에 유휴 상면이 있더라도 바로 도입하기 어렵다. 따라서, 과도한 구축 소요 시간/비용이 요구되는 하이퍼스케일 AI DC에 대한 대안으로, 기존 통신 국사 시설에 고성능, 고효율의 냉방 시설과 전력 공급 시설을 확충해 미리 준비함으로써 AI 서비스를 위한 장비를 유연하고 빠르게 도입할 수 있다.

2) AI 서버

두 번째 영역은 "AI 서버"로, 첫 번째 영역인 상면에 준비된 냉방과 전력 시설을 바탕으로 AI 서버를 보다 전진 배치해 통신 성능 향상과 Edge에서 AI 컴퓨팅 파워를 제공하는 것이다. AI 서버는 직접 Connectivity를 제공하지 않지만 네트워크 자동화, 최적화, 에너지 절감 등 통신서비스의 성능을 향상시키기 위한 AI Workload와 서드파티 또는 자사가 개발한 (통신 외) 서비스 목적의 AI Workload 제공을 위한 장비로 DU 집중 국사 이상의 시설에 도입될 수 있다. 이렇게 도입된 AI 서버는 Edge에서 AI 서비스를 제공해 대규모 AI 데이터 센터보다 상대적으로 낮은 지연과 높은 보안성을 고객에게 제공할 수 있다.

AI 서버는 대규모 AI 데이터 센터 수준의 AI 컴퓨팅 파워를 제공하는 것이 목적이 아니며 주로 저지연의 추론 기반 AI 서비스 제공을 목적으로 한다. 따라서 첫째, 제공하고자 하는 서비스의 AI Workload에 최적화된 xPU를 선정하는 것과 둘째, AI 서버와 단말 간 저지연 서비스 제공을 위한 UPF 구축에 대한 고려가 필요하다.



[그림 10] AI 서버

3) AI-RAN

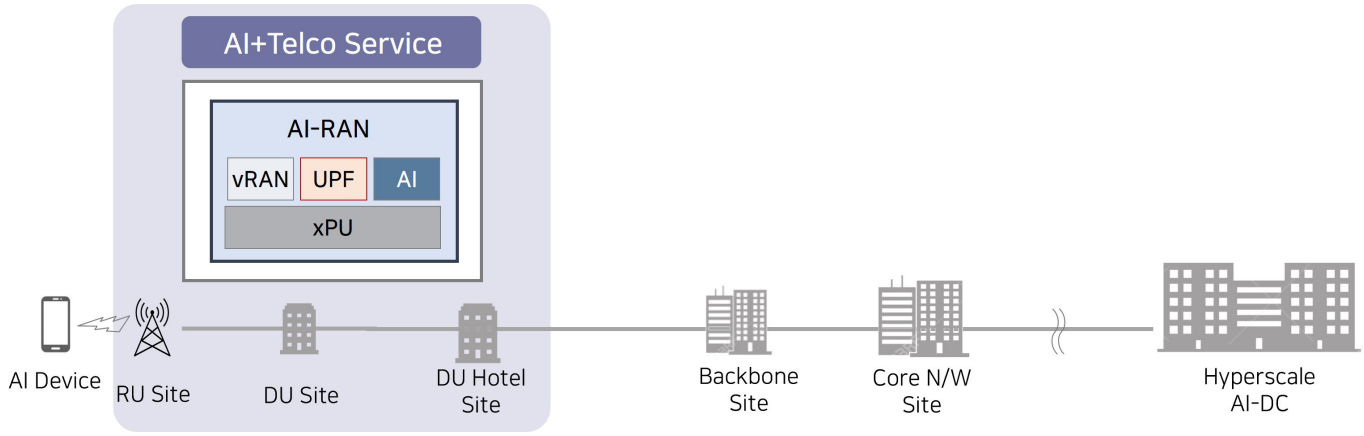
세 번째 영역은 "AI-RAN"으로, xPU 기반의 서버를 활용해 기존 Connectivity 위주의 통신 서비스를 제공하는 RAN 기능과 AI 서비스를 위한 AI 컴퓨팅 연산을 동시에 수행하는 영역을 의미한다. AI-RAN은 6G에서 논의되는 AI-Native를 구현하기 위한 핵심 기술 중 하나로, AI와 통신 서비스를 단일 장비에서 동시에 제공한다는 점이 두 번째 구성 요소인 AI 서버와의 차이라고 할 수 있다.

AI-RAN에서는 AI 컴퓨팅 파워를 크게 두 가지 분야에서 활용할 수 있다. 기존 RAN의 주요 Workload인 물리계층 신호처리나 복잡한 연산이 필요한 무선 자원 할당 기능 등을 AI 모델을 활용해 간략화하고 RAN 성능을 향상시키는 분야와, Connectivity를 위한 RAN 서비스 제공 후 남는 컴퓨팅 자원을 활용해 AI 서비스를 동시에 제공하는 분야이다.

AI-RAN은 기존 기지국 장비를 대체해 도입될 수 있다. 따라서, 도입 대상이 되는 사업자 시설의 범위가 넓고 장비의 절대적인 수량도 많으며, 기술적인 측면에서는 안정적인 통신 서비스를 기본적으로 제공해야 한다는 점에서 고려해야 할 사항이 많아 중 · 장기적 관점에서 검토가 필요하다.

다만 범용 하드웨어 기반의 RAN 장비는 아직 전통적인 통신 장비에 비해 셀 용량, 소모 전력 및 안정성 측면에서 개선되어야 할 부분이 많은 상태이다. 또한 AI 분야의 기술은 너무 빠르게 변화하고 있으며 새로운 AI 칩과 이를 활용한 플랫폼들이 계속해서 시장에 나오고 있다. 이에 비해 구형 제품에 대한 기술 지원은 상대적으로 짧은 편으로, 사업자의 시설에서 장기간에 걸쳐 안정적인 서비스를 제공하는 것이 필수 조건인 전통적인 통신 장비들과 다른 특징을 보인다. xPU에 기반한 AI-RAN 장비가

충분한 기술 유지 보수에 대한 보장이 되지 않는다면, AI 서비스뿐 아니라 안정적인 통신 서비스 제공에도 우려가 있어 장비의 수명주기 관리(LCM) 관점의 고려가 필요하다.



[그림 11] AI-RAN

Telco Edge AI 인프라에서 공통적으로 고려되어야 할 기술은 사업자가 제공하던 통신 서비스를 안정적으로 제공함과 동시에 AI 컴퓨팅 파워를 제공하는 것이다. 이를 위해서는 Telco Edge AI 인프라에서 제공하고자 하는 AI 애플리케이션 및 통신 서비스 요구사항을 고려한 xPU를 도입하는 것이 중요하다.

고성능 xPU의 경우 충분한 AI 컴퓨팅 파워를 제공하지만 과도한 전력 소모와 발열이 발생할 수 있다. 이에 비해 On-Device AI를 목적으로 한 AI 칩들의 경우, 낮은 전력 소모를 지향하지만 제공하는 AI 컴퓨팅 파워가 제한적이다. 최근 CPU, GPU, NPU를 포함한 AI 컴퓨팅 파워를 제공하기 위한 칩 기술이 경쟁적으로 발전하고 있다. 이들을 고려해 최적화된 xPU 기반의 Telco Edge AI 인프라를 구축할 수 있다면 효율성과 성능을 극대화해, 혁신적인 ICT 서비스와 애플리케이션을 제공하는 기반이 될 것으로 기대된다.

또한, AI-RAN도 AI 서버와 마찬가지로 AI 컴퓨팅 리소스의 Edge 전진 배치로 인한 저지연 및 높은 보안성을 제공해 줄 수 있다. 이 외에도 AI 서버와 AI-RAN 같은 Telco Edge AI에서는 AI 서비스가 네트워크 슬라이싱, QoS 관리 기술 등 네트워크 고유 기능과의 결합을 통해 혼잡 상황에서의 효율적인 트래픽 제어 등 하이퍼스케일러 대비 Telco만이 제공할 수 있는 추가적인 Value를 고객에게 제공할 수 있을 것으로 기대된다.

5. Telco Edge AI 진화 방향성

Telco Edge AI로의 진화가 성공적으로 달성되기 위해서는, 구현 방안을 구체화하고 이를 실행하기 위한 차별화 기술을 확보함과 동시에, 특화된 사업 모델을 지속적으로 발굴하며 제조사 · 사업자 · 고객이 서로 윈-윈할 수 있는 글로벌 생태계를 조성/확산하는 것이 매우 중요하다고 할 수 있겠다.

또한, 앞선 3.2절에서 살펴본 AI 인프라 활용의 또 다른 측면인 AI-Assisted Network 역시 이동통신 사업자가 현재 운용 중이거나 향후 구축할 네트워크 인프라의 성능과 효율을 높이기 위해 간과되어서는 안 되며, 향후 Telco Edge AI로의 진화 이전 단계에서 반드시 거쳐야 할 인프라 구조 진화의 단계라고 할 수 있다.

따라서, 본 장에서는 Telco Edge AI 생태계 조성/확산을 위한 활동들에 대해서 먼저 살펴보고, Telco Edge AI 특화 사업 모델의 예시들을 제시하며, 마지막으로 차세대 인프라 구조 진화를 위해 요구되는 주요 기술들에 대해 AI 인프라의 두 가지 활용 측면을 포함해 서술하고자 한다.

5.1 생태계 조성

5.1.1 사업자/제조사 협력

한정된 용도만으로 활용되고 있는 기존의 통신 네트워크 인프라에 Telco Edge AI 기술을 적용하고 차별화된 가치를 제공하는 인프라로의 구조 혁신을 위해서는 관련 기반 생태계 조성이 필수적이다.

이를 위해서는 새로운 비즈니스 모델과 수익 창출 방안을 모색 중인 이동통신 사업자들 간의 협력이 필수적이다. SK텔레콤은 이를 위해 주요 글로벌 이동통신사들과 미래 이동통신 네트워크 인프라 진화를 위한 비전 공유 및 연합을 통해 글로벌 Telco Edge AI 생태계를 구축하고자 한다. SK텔레콤은 지난 2024년 2월, 글로벌 통신사 간 협력을 통해 AI 기술 혁신과 적용을 가



[그림 12] MWC24 GTAA 창립총회

속화하기 위해 도이치텔레콤, 이앤그룹, 싱텔그룹, 소프트뱅크와 함께 MWC24에서 GTAA(Global Telco AI Alliance)를 창립하고, 텔코 LLM(Telco Large Language Model)을 공동 개발하는 등 '글로벌 AI 컴퍼니'로의 전환을 위한 전략을 발표한 바 있다. 이를 기반으로 협력 범위를 Telco Edge AI로 확장해 글로벌 사업자 Alliance를 통해 각 사업자가 보유한 기술과 노하우를 공유하고, Telco Edge AI 기반 서비스와 솔루션을 확보하기 위한 사업자 간의 협력과 연대를 추진할 계획이다.

또한, SK텔레콤은 주요 장비 및 칩셋 제조사와의 파트너십을 통한 글로벌 생태계 구축에도 힘쓸 예정이다. 과거 가상화 기술이 코어망과 무선접속망에 도입되기 시작하며 기존의 통신 특화 제조사 위주의 생태계에 일부 IT 업체들이 진입하면서 Telco 생태계에 변화가 일고 있다. 통신과 AI의 융합은 2030년경 도래할 것으로 예상되는 6G의 핵심 요소로 부각되고 있으며, 주요 AI 서버 및 칩셋 제조사들도 이와 비슷한 전략으로 통신 시장 진입을 시도하고 있다. 특히, GPU 중심으로 형성된 AI 컴퓨팅 인프라의 주도권을 선점한 업체에 대항해 여러 빅테크 기업들이 연합 세력을 구축하는 등 경쟁이 심화되고 있다.

특히, Telco Edge AI의 성공을 위해서는 기존 IT산업계에서 사용되는 시스템이나 솔루션과는 차별화된 Telco Grade의 Edge 인프라에 최적화된 사양이 필수적이다. 부연하자면, AI DC에서 대규모의 AI 학습과 추론에 사용되는 초고사양의 xPU와 서버는 Telco Edge AI에서는 총소유비용(TCO) 측면에서 비효율적일 뿐 아니라, 오버 스펙일 수 있다. 따라서, Telco와 단말, 칩셋, 서버, 네트워크/스토리지, 플랫폼, 컨텐츠 등 여러 분야의 플레이어들 간 긴밀한 소통과 협력을 통해 Telco Edge AI의 생태계 외연이 확장될 수 있도록 공동의 노력이 반드시 필요하다.

SK텔레콤은 통신과 IT 기술뿐만 아니라 AI 기술까지 아우르는 생태계 조성과 함께 당사 망 환경에 최적화된 Telco Edge AI 인프라 설계, 구축 및 관리, 그리고 통신 장비의 특성을 고려한 SLA(신뢰성, 수명주기관리, 기술지원) 및 TCO 확보 등을 위해 주요 장비 및 칩셋 제조사와의 협력을 진행하고 있다. 또한 정부와의 협력을 통한 생태계 조성/확산 노력의 일환으로써, 당사는 대한민국 과학기술정보통신부에서 발의한 6G 핵심기술개발사업(2021년~2025년, 총 1,917억원) 및 차세대 네트워크(6G) 산업 기술개발사업(2024년~2028년, 총 4,407억원) 국책과제를 참여하고 있다.

5.1.2 표준화 및 학술 연구 활동



[그림 13] 3GPP Stage 1 Workshop on IMT-2030 Use Cases, O-RAN Alliance 표준회의 발표

Telco Edge AI 인프라가 다양한 제조사의 하드웨어와 소프트웨어에서 원활하게 작동하고, Telco Edge AI를 구성하는 다양한 장치와 시스템 간의 상호운용성을 보장하는 등 생태계 확장 및 파편화 방지를 통한 규모의 경제를 실현하기 위한 기술 표준화 작업도 필수적이다. SK텔레콤은 2024년 5월 초에 열린 "3GPP Stage 1 Workshop on IMT 2030 Use Cases"에 패널리스트로 참여해 Telco Edge AI 인프라의 중요성 및 표준화 필요성을 제언했으며, 2024년 5월 말/6월에 열린 3GPP SA1#106 표준회의 [9] / O-RAN 얼라이언스(O-RAN Alliance) 표준회의 [10]에서 차세대 인프라 구조 진화를 위한 핵심 기술 중 하나인 Telco Edge AI 인프라의 표준화 반영을 제안했다. 앞으로도 SK텔레콤은 3GPP, O-RAN Alliance, ETSI 등 사실 표준화 단체에 관련 표준화 아이템을 제안하고 표준화 작업을 진행함으로써 혁신적인 서비스 및 새로운 가치를 제공할 수 있는 Telco Edge AI 인프라로의 진화를 추진해 나갈 계획이다.

또한 지적재산권 확보의 일환으로, Telco Edge AI 인프라 구축을 위한 장비 노드, 노드 내 구성 요소 등을 정의하고, 기존 장비와의 연동을 위한 링크/인터페이스 및 Edge 인프라에서의 AI 추론/학습 지원을 지원하기 위해 필요로 하는 교환 메세지를 정의함으로써, 유무선 인프라의 Telco Edge AI 인프라 진화를 위한 기초적인 구조를 제안했다. 또한 Edge 노드가 통신과 AI 서비스를 동시에 지원할 시, 복수의 Edge 노드 간 자원관리/최적화 구조 및 인터페이스, 교환 메시지 등을 정의해 인프라의 최적 자원 활용을 제안했다.

이러한 연구 활동 결과들을 바탕으로, 2024년 6월 한국통신학회 학술대회에서는 AI 네이티브 무선접속망 연구 개요 및 방향성과 6G 시뮬레이터 기반의 최적의 망 설계를 발표했고, 2024년 8월 IEEE Communications Magazine 논문 게재, 2024년 10월 International Conference on ICT Convergence(ICTC) 논문 발표 등 지속적으로 학술 연구 분야에도 기여할 계획이다 [11], [12], [13], [14].

또한 다양한 산학 협력 활동을 통한 연구개발 결과물을 통해 Telco Edge AI 생태계 확장에 대한 노력을 기울이고 있다. 일례로, AI 활용 사용 사례(Use Case)를 포함한 AI-RAN S/W 개발 및 AI Workload 부가 시의 xPU서버 기반 Telco Edge AI 인프라의 효용성 정량 평가를 진행할 예정이다. 또 다른 산학협력의 예로, IMT-2030 Framework 권고서 [3] 내 사용 시나리오(Usage Scenario) 중 하나인 ISAC 기술의 송수신 모뎀을 FPGA로 설계/구현하고 xPU 기반의 서버와 연결해 AI 학습/추론을 수행함으로써, 센싱 및 통신 서비스를 제공하는 부분을 시험/검증할 계획이다.

5.2 비즈니스 모델 발굴

Telco Edge AI 인프라 기반의 비즈니스 모델은 앞에서 언급되었듯이, 통신 인프라의 가치 재평가 및 고객 가치 창출 가능한 인프라로의 진화를 위해서 필수적으로 발굴되어야 하는 요소이다. 본 장에서는 Telco Edge AI 인프라의 활용 가능성이 예상되는 비즈니스 모델들을 예시로 제시한다.



[그림 14] Telco Edge AI 비즈니스 모델의 예시

1) 자율주행

현재까지 차량 제조사는 차량에 탑재 가능한 On-Device 기반 자율주행 플랫폼 개발에 집중하고 있는 상황이다. 자율주행을 위해서는 낮은 지연속도와 높은 보안성을 포함한 V2X 통신이 필수인데 AI 서버를 활용한 대규모 상면 및 서버 제공, AI-RAN을 활용한 실시간 추론 전용의 컴퓨팅 자원을 제공하는 비즈니스 모델이 고려될 수 있다.

2) 물류/유통

물류/유통의 경우, 비교적 단순하고 정형화된 업무 특성으로 인해 AI 기술 도입이 용이하다고 볼 수 있다. 재고 관리/창고 설비에 대한 자동화, 물류 운송 경로의 최적화 등의 영역에 Telco Edge AI 인프라를 활용함으로써 자동화 및 실시간 업무 처리에 대한 사업적 니즈를 충족시킬 수 있겠다.

3) 제조

제조업의 경우, 디지털 변환(DX)과 자동화/무인화 추세가 가속화되고 있는데, 공장 내 격리된 공간에서의 물류 운반, 설비 로봇, 공정 등에 빅데이터 분석/학습을 통한 효율화를 꾀할 수 있고 이러한 영역에 Telco Edge AI 인프라를 활용함으로써 신규 비즈니스 모델을 발굴할 수 있다.

4) 보안

보안 영역의 경우, 영상 분석, 객체 탐지, 지능화 모니터링 등의 영역에서 AI를 활용함으로써 효율성 증대 및 저지연을 통한 실시간 대응이 가능해질 수 있고, 이를 지원하기 위해 Telco Edge AI 인프라가 활용될 수 있겠다.

5) AI Device

최근 단말 기기 자체에서 인공지능 연산을 수행하는 On-Device AI 기술의 대두로 인해, 고객의 AI Device에 대한 기대감이 고조되고 있다. 다만 사용 사례나 현재 단말의 AI 성능 한계로 인해 AI Device에 대한 확산이 제한되고 있는 상황이다. 따

라서, Telco Edge AI 인프라를 활용한 AI Device의 AI 성능 개선이나 지연시간 감소 혹은 AI 지원을 위한 최소의 기능만 탑재된 저가 Device에 대해 Telco Edge AI 인프라 활용 AI 서비스 제공 등의 비즈니스 모델이 고려될 수 있다.

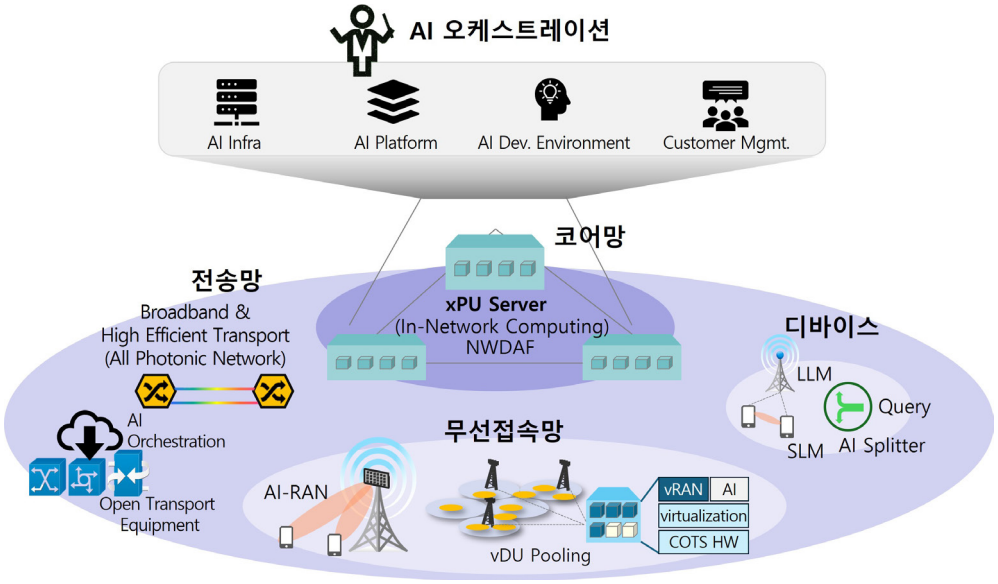
6) 의료

의료 분야에 있어서도 AI 기반의 방대한 의료 데이터 분석 및 질병 진단/치료/예방 등, 다양한 영역에서의 AI의 쓰임새가 기대되고 있다. 특히, 개인 의료 데이터에 대한 프라이버시 관점에서 높은 보안성을 요구할 수 있기 때문에, Telco Edge AI 인프라를 활용한 의료 관련 비즈니스 모델을 발굴하는 고민이 필요하다.

상기 비즈니스 모델 외에도 Telco Edge AI 인프라의 활용 가능성이 예상되는 신규 유망 비즈니스 모델을 발굴하는 노력이 중요하고, 이를 기반으로 통신 인프라의 AI Telco 인프라로의 진화를 촉진시킬 수 있을 것이다. SK텔레콤은 Telco Edge AI로의 성공적인 진화를 위한 가장 중요한 요인인 유망 비즈니스 모델 발굴을 가속화하기 위해 여러 가지 Use Case들에 대한 PoC를 진행해 유의미한 결과 확보 가능성을 검증할 예정이다.

5.3 주요 기술

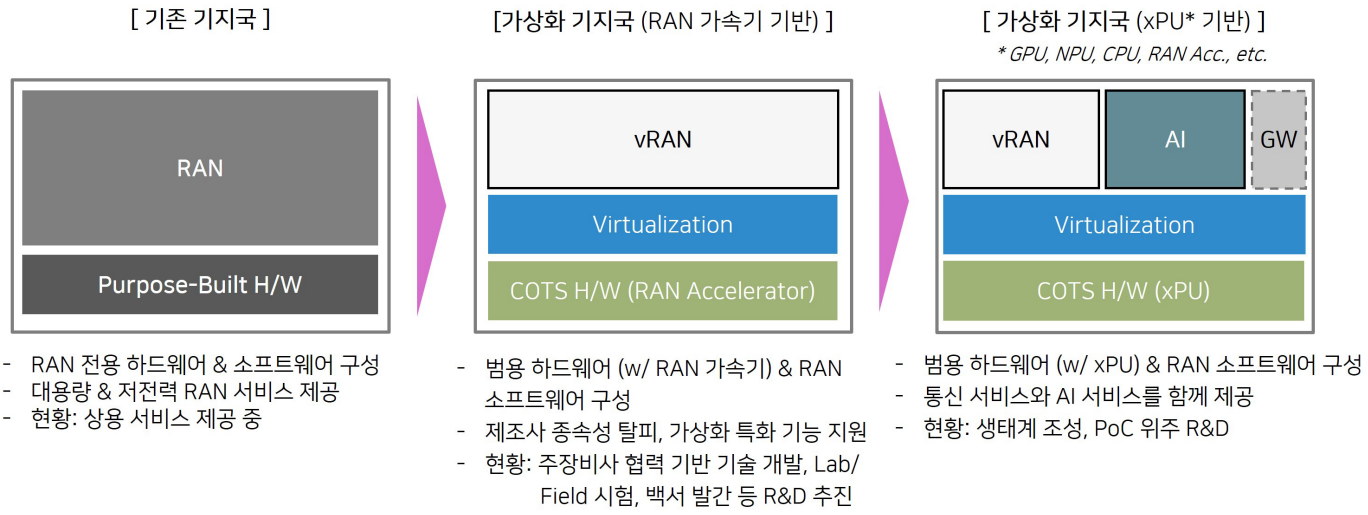
본 장에서는 Telco Edge AI로의 진화를 위해 필요한 기술 중 SK텔레콤에서 주목하고 있거나 R&D 진행 중인 기술을 소개한다. 소개된 기술들은 AI 컴퓨팅 자원의 효율적 배분/구축/관리 기술, Edge-Client/Edge-Edge 간 초저지연 통신을 위한 기술과 같이 Telco Edge AI에 필요한 요구사항을 충족하기 위한 기술들이다. 이러한 기술들 중 일부는 동시에 AI-Assisted Network를 위해서도 활용이 가능한데, 이 경우 AI-Assisted Network에서 활용되는 기대효과를 기술하였다. 네트워크 인프라의 전체 영역에서 조망해 본 Telco Edge AI 관련 주요 기술은 아래와 같다.



[그림 15] Telco Edge AI 관련 영역별 주요 기술

5.3.1 무선접속망

기존의 기지국은 제조사의 RAN 소프트웨어 운용에 최적화된 전용 하드웨어로 설계 및 구성되어 있어, RAN과 AI 서비스를 동시에 제공해야 하는 AI-RAN을 실현하기에는 어려움이 있다. 이에 반해 가상화 기지국은 범용 서버 하드웨어에 RAN과 AI Workload를 처리하기 위한 xPU를 장착하고, 하드웨어와 소프트웨어 사이의 가상화 Layer를 통해 기지국 제조사의 RAN 소프트웨어 가속과 서드파티 제조사의 AI 소프트웨어 가속에 xPU 자원을 공유해 사용할 수 있게 되므로, 하나의 Telco 인프라에서 RAN과 AI 서비스를 동시에 제공하고자 하는 AI-RAN 실현의 핵심 기술이다.



[그림 16] AI-RAN 실현 위한 기지국 구조 진화 방향성

한편, 현재의 가상화 기지국은 기존의 전용 하드웨어 기반 기지국 대비 Cell 용량, 소모전력, TCO 측면에서 일부 한계점이 존재한다. SK텔레콤은 CPU, RAN 전용 가속기, GPU 등 다양한 xPU 기반의 가상화 기지국을 개발, 연구실과 필드 환경에서의 시험을 통해 성능, 용량, 소모 전력, TCO 수준을 확인하고, 개선하기 위한 R&D를 수행하고 있다.

2022년 2월에는 주장비 제조사와의 5G 가상화 기지국에 대한 연구 결과를 MWC 2022를 통해 발표하였으며, 2023년 1월에는 상용 필드 환경에서 5G 가상화 기지국에 대한 실증에 성공하였다. 이후 RAN 전용 가속기 기술이 다변화 및 진화함에 따라 2023년 11월, 주장비 제조사와 협력을 통해 연구실 환경에서 기존 가상화 기지국 대비 신규 RAN 전용 가속기 기반 가상화 기지국의 Cell 용량 및 소모전력 측면의 개선을 확인하였다.

SK텔레콤은 이러한 가상화 기지국에 대한 연구 성과를 바탕으로 2024년 2월, NTT DOCOMO와 오픈랜 가상화 기지국 기술 백서를 공동 발간해 가상화 기지국에 대한 핵심 고려사항과 진화 방향을 제시하였다 [15]. 향후 RAN 전용 가속기뿐만 아니라 GPU 등 다양한 xPU 기반 가상화 기지국 개발 및 검증을 추진해 AI-RAN으로의 구조 진화 기반을 확보하기 위한 중장기 가상화 기지국 R&D를 지속해 나갈 예정이다.

또한 SK텔레콤은 AI-RAN 장비 기반으로 기지국의 여러 기능에 AI 기술을 적용해 통신 서비스 관점에서 성능 및 품질을 향상시키기 위한 R&D를 활발하게 수행하고 있다. 대표적인 사례로, 글로벌 사업자 및 제조사와 협력을 통해 기지국과 단말 간 무선 인터페이스에 AI 기술을 적용하는 연구를 진행하고 있다. 통상적인 이동통신 시스템은 송수신기간 통신을 위해 참조 신호(RS, Reference Signal)를 사용하는데, AI 기반 변복조 송수신 기술을 통해 참조 신호 부하를 최소화함으로써 주파수 이용 효율을 개선하였다. 해당 기술에 대한 PoC를 통해 실험실 환경에서 전송속도 향상 가능성을 확인하였으며, MWC 2024 에서 세부 연구개발 결과를 제시하였다. SK텔레콤은 이 외에도 다양한 Use Case에 대해 실제로 망을 운용하는 사업자 관점에서 효용성을 가장 염두에 두고 R&D를 수행하고 있다.

5.3.2 코어망

코어망은 단말기에서 발생하는 모든 음성과 데이터 트래픽이 인터넷 망으로 접속하기 위해서 필수적으로 거쳐야 하는 관문이다. 기능적으로는 단말 접속, 인증/등록, IP 주소 할당, 데이터 세션 관리, Handover, QoS 및 과금 처리를 수행하며, 각 역할별 다양한 NF(Network Function) 처리 장비들로 구현/구축 및 상호 연동된다.

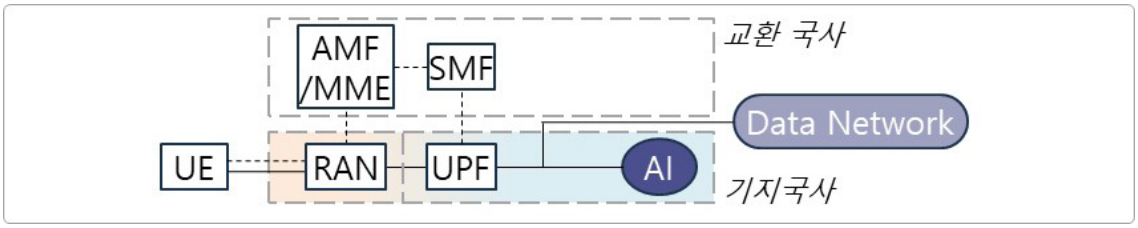
코어망은 대용량 & 저지연 처리, 안정성 및 고객 품질이 중요하다. Cloud-native로 진화할 때 SW 블록이 분화하면서 지연이 늘어나는 단점이 있다. 코어망은 서버 성능 향상에 따라 대용량화되고 있으며 높은 안정성을 요구하므로, 고장의 징후, 망의 품질 저하를 사전에 예측 감지해 자동으로 혹은 사전에 조치하는 것이 중요하다.

xPU 장착한 서버의 도움과 SW 기술을 통해서 시스템의 성능을 높이는 것은 중요하다. SK텔레콤은 Global 개발사와의 협업을 통해서 xPU를 활용한 대용량화 & 저지연 6G 핵심 기술을 개발했으며 더욱 발전시키고 있다. 그간의 연구 성과로 2024년 1월 통신 지연을 최대 70% 감소할 수 있는 '인라인 서비스 메시' 기술 백서를 발간하였고[16] 국제 이동통신 표준화 협력 기구(3GPP)에 반영을 추진하고 있다 [17].

In-Network Computing 기술을 통해서 대용량 데이터 트래픽 처리 시에도 효율적인 시그널, 패킷 데이터 품질을 분석할 수 있는 기술과 향후 AI 기반으로 이상 탐지(Anomaly Detection), 근본 원인 분석(Root Cause Analysis)을 할 수 있는 기술을 연구하고 있으며 논문 발표 및 연구 활동을 지속하고 있다 [18].

NWDAF(Network Data Analytics Function) 는 NF에서 발생하는 데이터를 분석해서, 장비 기능 최적화, 고객 품질, 성능 향상, 운용 자동화를 Telco는 인프라 진화와 AI를 통한 고객 가치 창출을 추구하는 방향으로 진화하고 있으며, 지원하는 표준 장비이다. SK텔레콤은 최상의 고객 품질을 제공하기 위한 네트워크 혼잡 예측 개발에 성공하였으며, 비정상 활동 단말 탐지 등 다양한 Use Case로 확장 연구를 지속하고 있다.

Telco Edge AI를 위해 xPU 장치에서 동작하며 UPF를 고객과 가까운 기지국사에 분산 전진 배치하기 위한 기술을 연구하고 있다. 전진 배치 UPF는 고객의 통화 및 데이터 서비스 트래픽과 달리, AI 서비스 트래픽을 AI 컴퓨팅 파워로 연결해, 고객과 가장 가까운 곳에서 추론 서비스를 제공할 수 있도록 한다.



[그림 17] 전진 분산 배치 UPF의 트래픽 처리 개념

5.3.3 전송망

생성형 AI 학습/서비스에 적합한 분산형 AI 데이터 센터가 도입되면서 유선 네트워크 구조가 변화하고 있다. AI 학습은 주로 데이터 센터의 GPU 서버를 통해 이루어지며, 이를 위해 광대역 및 무손실 전송 기술이 필수적이다.

AI 서비스는 GPU 의 높은 가격과 전력 소모 증가로 인해, 고객의 단말 또는 온프레미스(On-Premise)에서 AI 추론을 수행하는 구조가 적합하다. 이에 따라 분산형 AI 데이터 센터를 지원하는 유선 네트워크는 Edge AI DC 간 Inter-Connect 의 역할과 Edge AI DC 내에서 초광대역과 무손실 저지연 및 저전력 연결 서비스를 필요로 한다.

5.3.3.1 광대역·고효율 전송 기술

분산형 데이터 센터를 추구하는 Edge AI 특성상 효율적으로 대용량 데이터를 학습시키기 위해 Edge AI DC 간 광대역 · 고효율의 연결성을 확보해야 한다. 고효율을 높이기 위해 Edge AI DC 내 또는 Edge AI DC간 외부 광전 변환 없이 전력 효율을 높이는 방향으로 진화가 필요하다. 이에 대한 연구/개발을 하고 있는 주요 단체 중 하나는 IOWN 글로벌 포럼이며, E2E 광전송을 통해 고대역폭화와 초저지연화를 이룩하는 차세대 통신 네트워크를 APN(All Photonic Network)이라고 한다.

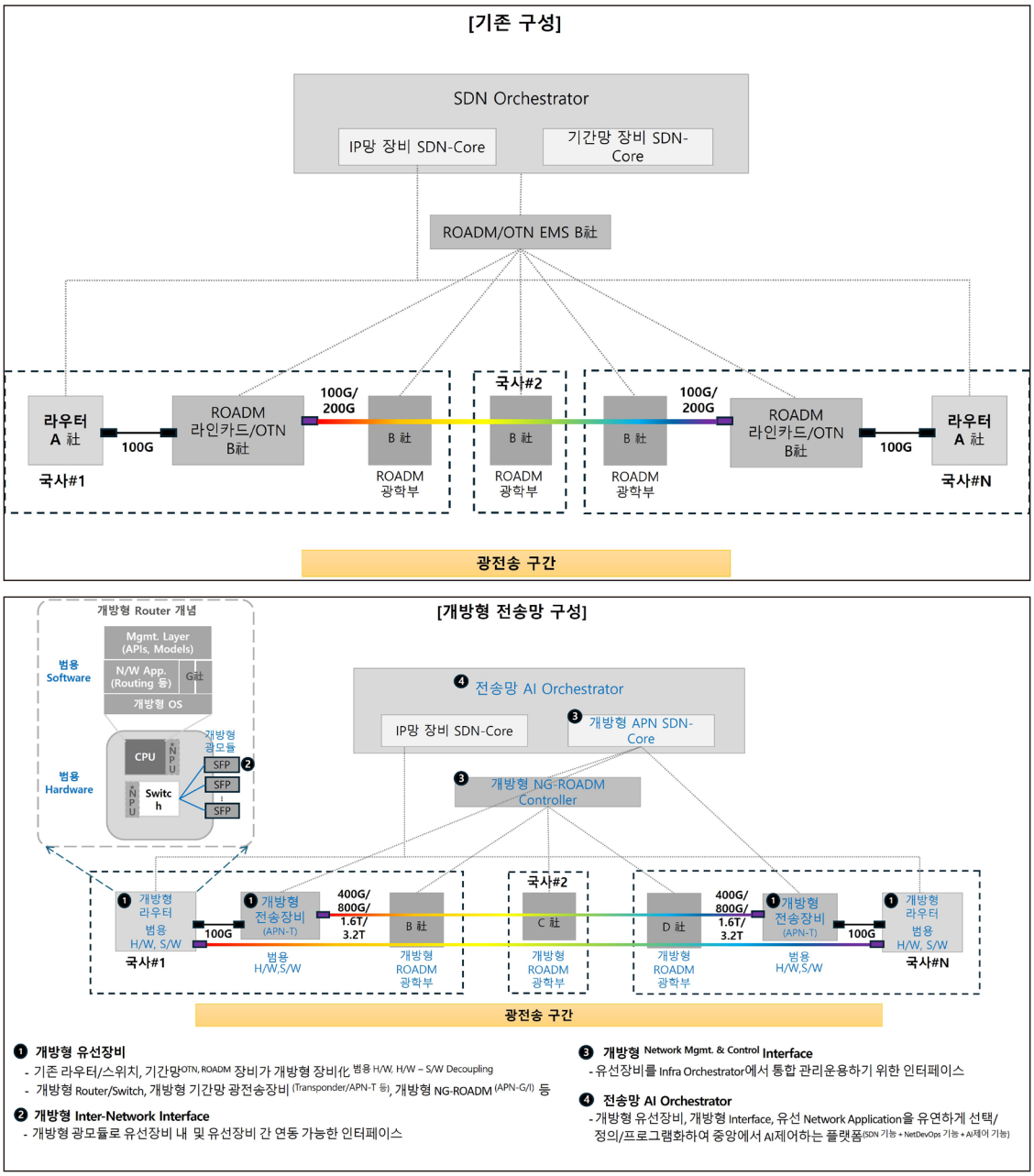
APN은 기존에 통합해서 사용하던 광전송 시스템을 용도(Transponder, Multiplexor, Wavelength Selective Switch 등)에 따라 분리시켜 각각 APN-T, APN-G, APN-I라는 이름으로 분류되며, 개방형 장비 생태계를 구축하기 위해 노력하고 있다.

개방형 전송 장비는 하드웨어와 소프트웨어를 분리해 범용 스위치/라우터, 광전송장비 하드웨어에 개방형 Network 운영체제 소프트웨어를 탑재해 사용이 가능하다. 이러한 개방형 유선장비를 활용하면 지역별 스위칭 용량에 따라 일원화된 유니트를 적용할 수 있으며, 하드웨어 교체 또는 증설에 유연하게 대응할 수 있다. 또한, 기능에 따라 소프트웨어를 On-Demand 기반으로 설치해 AI Telco 인프라에 최적화된 기능만을 효율적으로 사용할 수 있다.

Edge AI DC 내에는 CPO(Co-Package Optics) 기술이 유력한 기술 진화 후보로 활발하게 연구개발 중이며, Edge AI DC 간에는 400Gpbs 이상 500km 전송이 가능한 개방형 광모듈 기술 또한 중요하다.

Edge AI DC 내 및 DC 간에는 All Photonic Network를 달성하기 위한 광 스위칭 기술도 후보 기술 중 하나이다. 고속 스위칭이 장점인 패킷 스위칭을 완전히 대체하기는 어렵지만, 패킷 스위칭과 광 스위칭을 조합한 네트워크 구조를 통해 확장성이 높고 저지연 및 무손실 저전력을 만족하는 네트워크를 구현할 수 있다.

5.3.3.2 전송망 AI 오케스트레이션



[그림 18] AI Telco 인프라로의 전환을 위한 개방형 전송망 구성

Telco Edge AI 인프라의 네트워킹을 위해 위에서 언급한 개방형 전송 장비, 광 모듈, 광 스위칭 등을 모두 하나의 네트워크로 관리하는 전송망 AI Orchestrator 기술이 필요하다. 기존 벤더의 장비 수용뿐만 아니라 개방형 전송 장비와 광 모듈을 직접 관리하고, Telemetry 데이터 수집 및 AI 제어를 이용해 자동화 수준을 높이는 연구개발을 진행 중이다.

이러한 활동을 통해 궁극적으로 전송망에서의 TCO 절감, 유선 장비 완성도 확보, 운영 안정성 증대, AI 고도화를 추진 중이다.

당사는 개방형 유선 장비 R&D를 통해 AI Telco 인프라에 필요한 유선 네트워크 기술을 확보하고, AI 서비스에 필수적인 고대역폭, 초저지연 기술을 구현하기 위한 활동을 수행하고 있다.

AI Telco 인프라에서 전송망은 Edge AI 내 연결을 담당하는 개방형 IP 스위치, Edge AI 간 연결을 담당하는 개방형 라우터/개방형 기간망 광전송 장비로 구성된다.

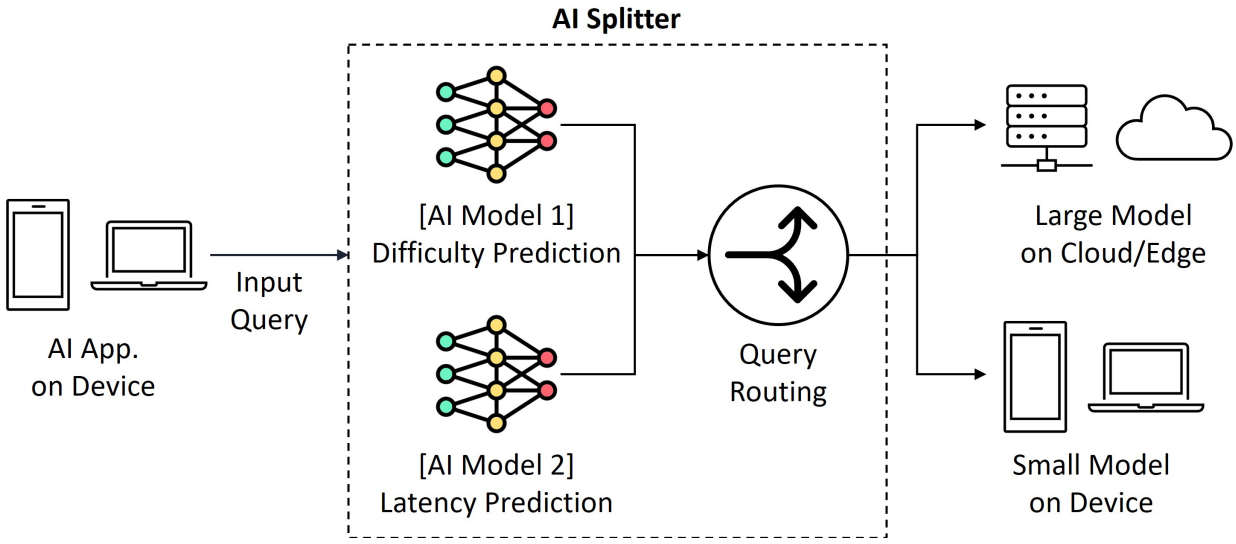
기존 전송 장비로 이루어진 망과의 가장 큰 차이점은 광 Coherent 기술 집적화를 통한 고대역폭화, 광전 변환의 최소화를 통한 초저지연화, 유선 장비의 개방화, 그리고 AI와 자동화를 포함한 망 관리 기술의 진화이다. 당사는 이에 대응하는 개방형 전송 장비, 개방형 장거리 광 모듈, 개방형 Optical Path 스위칭 기술, 전송망 AI 오케스트레이터 관련 기술 개발과 실증을 진행해오고 있다.

5.3.4 디바이스

AI가 중심으로 부각되는 6G에서도 디바이스는 고객과의 접점을 형성하는 매우 중요한 역할을 수행할 것으로 예상된다. 이에 SK 텔레콤에서는 Telco Edge AI 환경에서 통신과 AI의 효과적인 융합을 위한 디바이스 측면의 R&D를 적극적으로 추진하고 있다.

5.3.4.1 단말 기반 Hybrid AI 최적화

AI 기술 발전의 흐름 속에서 디바이스의 역할 강화를 통한 AI 서비스 품질 및 효율성 향상 방안을 연구하고 있다. 최근 모바일 AP의 지속적인 AI 컴퓨팅 성능 향상으로 디바이스에서 구동 가능한 소형 언어 모델이 등장하고 있다. 구글의 제미나이 나노, 마이크로소프트의 파이3 미니, 메타의 모바일LLM이 대표적인 예이다. 이들은 클라우드에서 구동되는 LLM 대비 Q&A와 같은 특정 시나리오에서 유사 수준의 성능을 보임과 동시에 인프라 구축/운용 측면에서 비용 효율적인 특징을 갖는다. 따라서 점차 현실화되고 있는 하이브리드 AI, 즉 디바이스-Telco Edge AI-클라우드 공존 환경에서는 사용자의 요청을 어느 노드에서 처리하는 것이 가장 유리한 지 효과적으로 판단해야 하는 문제가 발생한다. 따라서 SK텔레콤에서는 Telco Edge AI 및 On-Device AI를 활용해 사용자의 요청에 대한 다양한 특성을 분석/예측한 후 이를 최적의 노드에 할당하는 AI Splitter 기술을 개발하고 있으며, 향후 PoC를 통해 그 효과성을 입증할 예정이다.



[그림 19] AI Splitter 동작 개념도

5.3.4.2 Telco Edge AI 활용 신규 디바이스

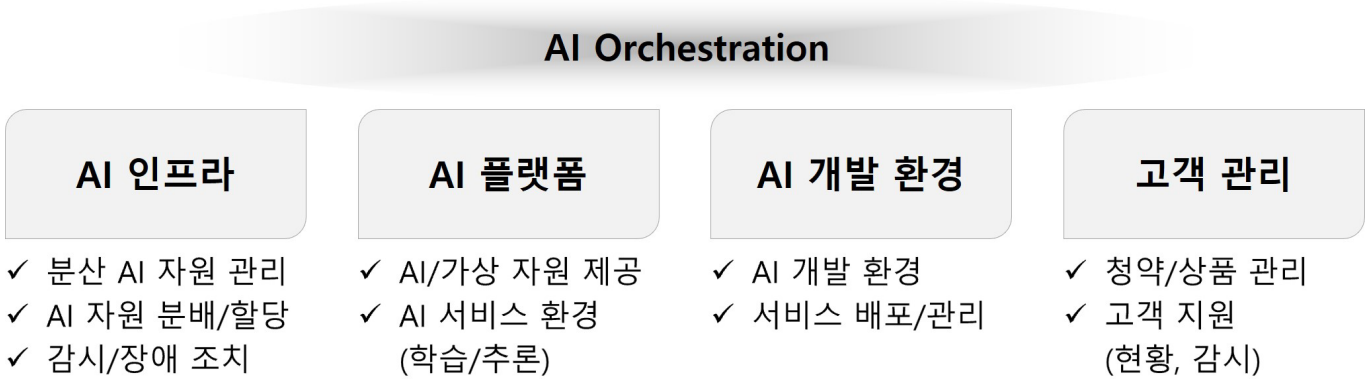
6G와 Telco Edge AI 기반 신규 비즈니스 모델 및 Use Case를 발굴/지원하기 위한 디바이스에 대한 R&D 활동을 수행하고 있다. 최근 AI 기술을 효과적으로 활용하기 위한 다양한 폼 팩터를 가진 신규 디바이스가 출현하고 있다. 일반적인 스마트폰의 동작 방식에서 벗어나 음성 및 제스처를 기반으로 개인 비서 서비스를 제공하는 Rabbit의 R1이나 AI Pendant, 그리고 AR/VR에 AI를 접목해 고객에게 새로운 경험을 제공할 수 있는 메타의 스마트 글라스 및 퀘스트3가 대표적인 예라고 할 수 있다. 이러한 새로운 형태의 디바이스는 현재 프로세서 및 메모리 제약으로 인해 서버 기반의 AI 서비스를 채택하고 있다. 하지만, 서버에만 활용하는 AI 서비스는 고객의 체감 품질 및 정보 보호 관점에서 한계가 존재한다. 따라서 당사는 저지연과 높은 보안성 확보를 위해서 향후 도입될 6G 기술과 함께 Telco Edge AI 및 On-Device AI를 복합적으로 활용할 수 있는 AI 특화 디바이스를 발굴할 예정이다. 또한 이와 관련 AI 서비스도 함께 개발해 AI 시대의 새로운 사용자 경험을 제공하기 위한 노력을 지속할 예정이다.

5.3.5 AI 오케스트레이션

Telco Edge AI 에서 사용될 AI 컴퓨팅 자원을 End-to-End로 관리하기 위한 R&D를 진행해 오고 있다. 고객에게 효율적이고 빠른 AI 개발 환경을 제공하고 AI 서비스를 제공하기 위해 AI 오케스트레이션이 필요하다. AI 인프라와 플랫폼을 고객에게 제공하고, AI 서비스 개발 환경에 도움을 주며, 고객이 필요한 정보를 드릴 수 있도록 관리하는 영역에서 기술 개발을 진행 중이다.

AI 오케스트레이션은 AI 서비스를 제공하기 위해 필요한 AI 컴퓨팅 자원을 제공한다. Telco Edge AI 환경에서는 AI 컴퓨팅

자원이 소규모 국사, 교환 국사에 분산되어 있다. 따라서, 분산된 AI 컴퓨팅 자원을 효과적으로 관리하고, 고객에게 제공해 사용자 경험을 향상시킬 필요가 있다. AI 오케스트레이션은 분산된 AI 컴퓨팅 자원을 통합 관리하며, 고객 요구에 맞추어 자원을 할당하고 배치할 수 있도록 지원한다. 아울러, 고객이 AI 서비스를 개발하고 상용으로 배포하기 위한 환경을 제공한다. AI 오케스트레이션은 AI 인프라 관리, 플랫폼 관리, 개발 환경 관리, 고객 관리 영역에서 중요한 역할을 수행한다.



[그림 20] AI 오케스트레이션의 주요 역할

AI 서비스의 학습과 추론에 사용되는 AI 컴퓨팅 자원은 초저지연의 광대역 네트워크, 저장 공간을 지원해야 한다. AI 컴퓨팅 자원은 실시간 할당과 연결이 필요하며, 서비스가 다양하게 발전하면서, 복잡한 AI 컴퓨팅 자원을 유연하고 비용 효율적으로 제공해야 한다. 분산된 AI 컴퓨팅 자원의 상태를 관리하고, 자원을 최적으로 사용하도록 관리해야 한다. 또한, 고객의 AI 서비스 품질을 높이기 위해 다양한 AI 컴퓨팅 자원을 통합 관리하고 AI 서비스 Workload의 분산 및 스케줄링이 필요하다.

SK텔레콤은 가상화 환경을 맞이해 SDI(Software Defined Infrastructure) 기술을 개발, 상용망에 적용하고 있다. SDI는 가상화 컴퓨팅 자원을 설치하고 관리하는 솔루션이다. 또한, AI 컴퓨팅 자원관리, AI 플랫폼, AI 개발 환경, 고객관리(요금 최적화) 솔루션을 보유하고 있다. 해외 솔루션과 협업 및 솔루션 간의 통합을 통한 AI 오케스트레이션을 제공하기 위한 연구를 진행하고 있다. SK텔레콤이 주목하고 있는 기술 분야를 살펴본다.

5.3.5.1 AI 인프라 관리

AI 컴퓨팅 자원을 제공하는 서비스 제공자로서 AI 컴퓨팅 자원을 구축하고 지속 관리해야 한다. AI 컴퓨팅 자원으로서, 서버/GPU의 모델과 성능, 네트워크, 저장 공간, 가상 환경을 분산된 지역에 원격으로 자동화해 구축해야 한다. 또한, 정기적인 업그레이드를 통해서 최신 상태를 유지해 고객에게 제공할 수 있어야 한다.

고객에게 무중단 서비스를 제공하기 위해 온도, 하드웨어 및 소프트웨어 고장 감시, 알람을 제공해야 한다. 고장을 자동 복구하고, 자원을 자동 업그레이드하기 위해서는 고객에게 제공 중인 AI 컴퓨팅 자원은 무중단 이전하거나 사용 중단까지 대기하는 기술이 필요하다.

AI 서비스는 높은 처리 능력을 요구하므로, GPU, 네트워크, 저장공간과 같은 AI 컴퓨팅 자원을 효율적으로 관리하지 않으면 서비스 품질 저하나 중단이 발생할 수 있다. AI 서비스 품질을 위해 GPU를 포함한 AI 컴퓨팅 자원을 최적으로 할당하고, 병목이 발생하지 않도록 AI 컴퓨팅 자원을 관리해야 한다. 즉, AI Workload를 여유 GPU로 자동 재배치해 고성능의 AI 서비스를 안정적으로 제공하는 자원 최적 관리가 중요하다.

5.3.5.2 AI 플랫폼 관리

AI 오케스트레이션은 물리적인 AI 컴퓨팅 자원을 가상 자원으로 변환해 효율적으로 활용할 수 있게 한다. 고객의 요청에 따라 VM 환경, 컨테이너 환경의 생성, 배포를 지원해야 한다. VM 가상화로 독립된 작업 공간을 제공할 수 있고, 컨테이너 형태의 가상화 환경을 제공한다. 쿠버네티스와 같은 컨테이너 관리 플랫폼을 통해 컨테이너의 배포, 증설, 운영 자동화를 지원한다.
AI 플랫폼 관리는 AI 컴퓨팅 자원의 유연성과 효율성을 극대화하기 위해 필요하다. AI 서비스의 다양한 요구사항을 효과적으로 충족하기 위해 컨테이너 가상화 기술이 필수적이다. AI플랫폼 관리 기술을 통해, 다양한 AI Workload를 효과적으로 운영하고, 자동화된 관리와 안정적이고 신뢰성 높은 서비스를 제공한다.

AI 컴퓨팅 자원은 AI 서비스의 학습 및 추론 작업에서 핵심적인 역할을 하기 때문에, AI 컴퓨팅 자원을 가상화 플랫폼에 유연하게 연결해 고객의 서비스 환경으로 제공하는 것은 중요하다. AI 컴퓨팅 자원을 최대한 활용해 AI Workload 처리 시간을 단축하고 AI 서비스 성능을 최대화할 수 있도록 AI 컴퓨팅 자원을 최적 배분하고 관리하는 AI 컴퓨팅 자원 스케줄링은 중요하다.

특히, 대규모의 AI 서비스 학습/추론을 위해 다수의 GPU를 묶어서 사용하는 클러스터링 기술이 필요하다. NVIDIA는 학습을 위한 데이터를 고대역폭으로 처리하는 메모리 기술로 GDR(GPU Direct RDMA), 네트워크 기술로 NVLink, Infiniband를, 저장 장치 기술로 GDS(GPU Direct Storage)를 제공하고 있다. 그 외 네트워크 기술에는 UALink(Ultra Accelerator Link), 이더넷에서 사용하는 RoCE(RDMA over Converged Ethernet) 기술이 발전하고 있다. 하지만, 동일 건물 내에서는 사용이 가능하지만, 수십 km 이상 떨어진 곳의 AI 컴퓨팅 자원은 지연 증가와 속도 감소라는 물리적인 한계로 사용이 어렵기 때문에 이를 극복하기 위한 방안을 다각적으로 검토하고 있다.

나아가 GPU 클러스터 자체를 가상화하고 동적으로 분할하는 기술이 필요하다. AI 서비스에서 GPU 할당량에 부합하게 Infiniband 혹은 RoCEv2 자원을 할당하지 않으면 네트워크가 병목이 되어 GPU 사용률이 낮아진다. AI 플랫폼은 GPU 서버끼리 메세지를 주고받는 패턴과 Rail-Optimized 구조의 물리 네트워크 형상을 고려해 고객의 자원 요청량에 맞게 동적으로 GPU 클러스터 전용 가상 네트워크를 구축하고 격리한다.

분산된 Edge에서 AI 컴퓨팅 자원의 클러스터링은 소프트웨어적인 병렬, 분산처리를 통해 극복해야 한다. 학습 데이터와 모델의 크기에 따라서 데이터 병렬처리, 모델 병렬처리를 사용한다. Keras, Tensorflow, PyTorch, MxNet을 모두 지원하는 잘 알려진 분산 학습 프레임워크인 오픈소스 Horovod는 데이터와 모델 병렬 처리를 모두 단일 프레임워크 기반의 분산 처리 대비 좀 더 빠르고, 좀 더 쉽게 사용할 수 있게 지원한다. GPU 클러스터링은 AI 컴퓨팅 자원, 데이터와 모델의 최적화를 통해서 학습 성능을 높여야 하며, AI 컴퓨팅 자원이 분산된 Edge 환경에서 AI 오케스트레이션은 GPU 클러스터링을 위한 소프트웨어를 통합하는 기술이 중요하다.

5.3.5.3 AI 개발 환경 관리

AI 오케스트레이션은 AI 서비스 개발을 위한 개발과 검증 환경을 제공한다. AI 모델 학습에 필요한 데이터를 효율적으로 처리할 수 있도록 데이터 처리 및 관리 도구를 제공한다. 데이터 처리 및 모델 배포 도구는 AI 모델 기반이 되는 데이터 처리와 모델 배포를 자동화해 개발 효율성을 향상한다.

통합 개발 환경은 AI 개발자들이 신속하게 AI 모델을 개발하고 검증할 수 있도록 한다. AI 개발 과정의 복잡성을 낮추고 효율성을 높이기 위해 AI 개발 도구와 프레임워크 지원이 필요하다. AI 모델의 학습, 배포 기능을 자동화해 반복적인 작업을 최소화하고 모델 성능을 지속적으로 개선하도록 MLOps 환경을 지원한다. AI 개발 환경은 AI 개발 주기를 단축하고 AI 서비스 품질을 지속적으로 향상할 수 있도록 한다.

5.3.5.4 고객 관리

AI 컴퓨팅 자원의 이용을 위해서는 상품 관리, 구매 지원이 필요하다. AI 오케스트레이션은 고객이 AI 서비스에 접근해 이용하고, AI 컴퓨팅 자원을 모니터링하며 서비스를 관리할 수 있는 포털 기능을 제공한다. 고객 포털에는 다양한 AI 컴퓨팅 자원, 개발 환경, AI 서비스 상품을 구성해 고객이 선택하고 사용할 수 있어야 한다.

고객 포털을 통해 고객은 구매할 수 있는 AI 컴퓨팅 자원을 확인하고, 구매한 AI 컴퓨팅 자원을 관리하고, 감시할 수 있어야 한다. 고객 스스로 AI 컴퓨팅 자원 사용을 예상하고 평가할 수 있고, 고객에게 최적의 AI 컴퓨팅 자원 사용을 제안하는 것도 중요한다. 또한, 기술 지원, 문제 해결, 유지보수, 커뮤니티 등 고객 지원 서비스를 제공한다.

고객은 AI 컴퓨팅 자원을 효과적으로 활용할 수 있으며, 기술 지원을 받아 비즈니스 목표를 달성할 수 있도록 지원해야 한다. 지속적인 기술 지원과 유지보수를 통해 안정적이고 신뢰성 높은 AI 서비스를 제공할 수 있도록 지원해야 한다. AI 컴퓨팅 자원을 쉽게 관리하고 모니터링 할 수 있으며, 계약 조건을 유연하게 조정해, 서비스를 비용 효율적으로 이용할 수 있어야 한다.

6. 맺음말

SK텔레콤은 Telco 네트워크 인프라를 Telco Edge AI 인프라로 변모시켜 나가고자 하는 비전을 갖고 있으며, 이를 통해 고객이 체감할 수 있는 새로운 가치와 비즈니스 모델을 만들어 Telco 사업의 지속가능성을 높일 수 있을 것으로 기대한다.

SK텔레콤은 CDMA 세계 최초 상용화를 달성한 이후 수많은 도전의 역사를 일구어 온 혁신에 대한 노력을 인정받아 2024년 IEEE 마일스톤에 등재되는 영예를 안았다.



[그림 21] 'CDMA 상용화' IEEE 마일스톤 국내 최초 등재

1996년 CDMA 세계최초 상용화 이후 긴 호흡으로 도전과 혁신을 지속해 온 SK텔레콤의 역사를 돌아볼 때, Telco 네트워크 인프라의 혁신과 변화는 단기간에 완수할 수 있는 과제는 아닐 것이다. 하지만, ICT 분야에서 수많은 파트너들과의 협력을 통해 작지만 의미 있는 Quick-Win 사례들을 축적하고, 체계적인 마일스톤을 설정해 중장기적 목표를 달성해 나아가고자 하는 지금 이 시점이 이동통신 분야에서 또 다른 획을 긋는 새로운 출발점이 될 것으로 기대한다.

References

[1] SK Telecom, "5G Lessons Learned 및 6G 핵심 요구사항, 6G 네트워크 기술 진화 방향, 6G 스펙트럼," SK Telecom 6G 백서, Aug. 2023. [Online]. Available: https://bit.ly/3RmZOMk

[2] Deloitte Insights Report, "새로운 고차원 가치 창출하는 통신산업의 진화", Apr. 2024.

[3] ITU-R M.2160-0, "Framework and overall objectives of the future development of IMT for 2030 and beyond", Nov. 2023

[4] K. B. Letaief, W. Chen, Y. Shi, J. Zhang and Y. -J. A. Zhang, "The Roadmap to 6G: AI Empowered Wireless Networks," IEEE Communications Magazine, vol. 57, no. 8, pp. 84-90, Aug. 2019

[5] ITU-R M.2083-0, "IMT Vision - Framework and overall objectives of the future development of IMT for 2020 and beyond", Sep. 2015

[6] McKinsey & Company, "The economic potential of generative AI: The next productivity frontier", Jun. 2023

[7] Value Market Research, "Global AI In Telecommunication Market Research Report –Industry Analysis, Size, Share, Growth, Trends and Forecast 2023 to 2030," Nov. 2023

[8] "통신장비산업의 가치사슬별 경쟁력 진단과 정책 방향," 월간 KIET 산업경제, vol. 285, pp. 68-80, June 2022

[9] SK Telecom, S1-241246, "SK Telecom's View on Future Telco. Infrastructure", 3GPP SA1#106, May 2024

[10] SK Telecom, "The path to AI Telco Infra Evolution", 6th O-RAN 6G Workshop: The Road to 6G Standards: Views from Industry and Academia, Jun. 2024

[11] 이경필 외, "AI 네이티브 무선접속망 연구 개요 및 방향성", 한국통신학회 하계학술대회, Jun. 2024

[12] 오지혜 외, "6G 시뮬레이터 기반의 최적의 망 설계", 한국통신학회 하계학술대회, Jun. 2024

[13] M. Na, et al, "Operator's Perspective on 6G: 6G Services, Vision, and Spectrum", IEEE Communications Magazine, vol. 62, no. 8, pp. 178-184, Aug. 2024

[14] H. Lee, et al, "The Evolution Path of Telco Infrastructure: From Open RAN to 6G AI-RAN (Telco Edge AI)", in the 15th International Conference on ICT Convergence (ICTC) Workshop, Oct. 2024

[15] SK Telecom and NTT DOCOMO, "Key Considerations for vRAN: Insights from SK Telecom and NTT DOCOMO", Feb. 2024

[16] SK Telecom and Intel, "Toward 6G Core Architecture Using an Inline Service Mesh", Jan, 2024

[17] SK Telecom, S1-242019, "6G Area of Interest", 3GPP SA1#107, Aug. 2024

[18] H. Kim et al. "Frequency-Aware and Telemetry Item-Wise In-band Network Telemetry for Per-flow Measurement", IEEE INFOCOM, May. 2024

Abbreviations

Abbr.	Full Name	Abbr.	Full Name
3GPP	3rd Generation Partnership Project	**LCM**	Life Cycle Management
AI	Artificial Intelligence	**LLM**	Large Language Model
AI DC	Artificial Intelligence Data Center	**LTE**	Long Term Evolution
AI-RAN	Artificial Intelligent-Radio Access Network	**ML**	Machine Learning
AIX	AI Transformation	**MLOps**	Machine Learning Operations
AP	Application Processor	**mMTC**	massive Machine Type Communications
APN	All Photonic Network	**NPU**	Neural Processing Unit
APN-T	All Photonic Network Transceiver	**NWDAF**	Network Data Analytics Function
APN-G	All Photonic Network Gateway	**PAA**	Personal AI Assistant
APN-I	All Photonic Network Interchange	**PoC**	Proof of Concept
AR	Augmented Reality	**QoS**	Quality of Service
B2B	Business-to-Business	**RAN**	Radio Access Network
B2C	Business-to-Customer	**RDMA**	Remote Direct Memory Access
CDMA	Code Division Multiple Access	**RoCE**	RDMA over Converged Ethernet
CPU	Central Processing Unit	**RU**	Radio Unit
DCI	Data Center Interconnect	**SDI**	Software Defined Infrastructure
DU	Digital Unit	**SLA**	Service Level Agreement
DX	Digital Transformation	**TCO**	Total Cost of Ownership
eMBB	enhanced Mobile BroadBand	**UALink**	Ultra Accessible Link
ETSI	European Telecommunications Standards Institute	**UAM**	Urban Air Mobility
FPGA	Field Programmable Gate Array	**UPF**	User Plane Function
GDR	GPU Direct RDMA	**URLLC**	Ultra Reliable Low Latency Communications
GDS	GPU Direct Storage	**V2X**	Vehicle to Everything
GPU	Graphic Processing Unit	**VM**	Virtual Machine
GTAA	Global Telco AI Alliance	**VR**	Virtual Reality
ICT	Information & Communication Technology		
IEEE	Institute of Electrical and Electronics Engineers		
IMT	International Mobile Telecommunication		
IOWN	Innovative Optical and Wireless Network		
IP	Internet Protocol		
ISAC	Integrated Sensing And Communication		
ITU-R	International Telecommunication Union Radiocommunication Sector		

Change History

Date	Descriptions	Version
2024-10-15	Initial Release	1.0


SK telecom