ICT 진흥 및 혁신기반 조성사업 RS-2025-02173110

정책연구 25-28-10

디지털 저탄소 전환을 위한 기술적·제도적 방안

A Study on Technological and Institutional Measures for Digital Low-Carbon Transformation

2025. 12

주관연구기관 : 정보통신정책연구원

연구기관 : (사)대한전자공학회

과학기술정보통신부 정보통신기획평가원

ICT 진흥 및 혁신기반 조성사업 RS-2025-02173110
정책연구 25-28-10

디지털 저탄소 전환을 위한 기술적·제도적 방안

(A Study on Technological and Institutional Measures for Digital Low-Carbon Transformation)

김 현 외

2025. 12

주관연구기관 : **정보통신정책연구원**
연구기관 : **(사)대한전자공학회**

과학기술정보통신부 정보통신기획평가원

이 보고서는 2025년도 과학기술정보통신부 방송통신발전기금 ICT 진흥 및 혁신기반 조성사업의 연구결과로서 보고서 내용은 연구자의 견해이며, 과학기술정보통신부의 공식입장과 다를 수 있습니다.

제 출 문

과학기술정보통신부 장관 귀하

본 보고서를 『**디지털 대전환 메가트렌드 연구: 디지털 저탄소 전환을 위한 기술적·제도적 방안**』의 연구결과보고서로 제출합니다.

2025년 12월

연 구 기 관 : (사)대한전자공학회
총괄책임자 : 김 현 교 수
참여연구원 : 이승일 박사과정
김남준 박사과정
신 진 박사과정
구광현 박사과정
최다훈 박사과정
장지훈 박사과정
이길하 박사과정
황인성 박사과정
정상범 박사과정
강범진 박사과정

목 차

표 목 차

그 림 목 차

요 약 문

1. 제 목

디지털 저탄소 전환을 위한 기술적·제도적 방안

2. 연구 배경 및 목적

본 연구의 배경은 인공지능(AI)과 디지털 기술이 산업 혁신을 주도하는 동시에, 기하급수적인 에너지 소비 증가와 탄소 배출을 야기하는 중대한 역설에 맞닿아 있다. 국제에너지기구(IEA)에 따르면 2022년 기준 전 세계 데이터센터의 전력 소비는 세계 전력 수요의 약 1~1.3%를 차지했으며, 2030년까지 AI 확산으로 인해 그 수요가 폭발적으로 증가할 전망이다. 이는 단순한 미래의 우려가 아닌 현실적 위협으로, 유럽연합(EU) 등 주요 경제권은 이미 데이터센터의 에너지 효율성 공개를 의무화하는 등 저탄소 전환을 새로운 무역 및 기술 경쟁의 기준으로 삼고 있다. 글로벌 ICT 기업들은 와트당 성능(Performance per Watt)을 핵심 경쟁력으로 내세우며 저전력 AI 가속기 개발과 재생에너지 전환에 사활을 걸고 있다. 그러나 국내에서는 아직 고성능 AI 개발과 탄소중립 전략이 유기적으로 결합되지 못하고 있어, 국제 규제와 시장 변화 속에서 경쟁력이 약화될 위험에 처해 있다.

이에 본 연구는 디지털 경제의 탄소 발자국을 종합적으로 진단하고, 고성능과 저탄소를 동시에 달성할 수 있는 통합적이고 다층적인 전략을 제시하는 것을 목표로 한다. 구체적으로는 ▲디지털 기술의 탄소 영향을 정량적으로 진단하고 ▲주요국의 저탄소 전환 정책을 비교·분석하며 ▲산업별 디지털 탄소 감축 성공사례를 발굴하고 ▲AI 모델 경량화부터 맞춤형 반도체 설계까지 구체적인 저탄소 기술 체계를 정립하며 ▲이를 뒷받침할 연구개발, 규제, 공공부문 선도 등 정책·제도적 추진체계를 설계함으로써, 지속가능한 디지털 생태계 속에서 대한민국의 미래 디지털 리더십을 확보하기 위한 실질적인 로드맵을 제시하

고자 한다.

3. 연구의 구성 및 범위

본 연구의 구성은 디지털 저탄소 전환 문제를 다각적으로 분석하고 통합적 해법을 제시하기 위해 총 5개의 장으로 설계되었다. 각 장은 '진단-전략-기술-정책'의 유기적 흐름을 따른다.

첫째로, 연구의 배경과 목표를 제시하고, OECD의 디지털 경제 계량화 프레임워크를 기반으로 연구 범위를 설정한다. AI 모델 학습·추론, 데이터센터, 통신 인프라, ICT 단말기 제조 등 디지털 기술의 각 단계에서 발생하는 에너지 소비와 탄소 배출 현황을 정량적으로 진단하고, EU·미국 등 주요국의 규제 및 정책 동향을 분석하여 국내 상황과의 격차를 규명한다.

둘째로, 디지털 기술을 활용한 탄소 감축의 두 가지 측면을 다룬다. 먼저 데이터센터, 클라우드, 통신망 등 디지털 인프라 자체의 에너지 효율을 최적화하는 저탄소 전환 방안을 국내외 사례를 통해 분석한다. 다음으로 제조, 에너지, 물류 등 핵심 산업 분야에서 AI, 디지털 트윈 등 디지털 기술을 적용하여 공정 효율화와 에너지 절감을 달성한 구체적인 탄소 감축 적용 사례를 심층적으로 제시한다.

셋째로, 저탄소화를 위한 핵심 기술적 해법을 소프트웨어와 하드웨어로 나누어 체계적으로 분석한다. 소프트웨어 기반 솔루션으로는 모델 경량화(양자화, 프루닝, 병합), 에너지 지향적 알고리즘(선형 어텐션, 신경망 아키텍처 탐색, 적응형 디코딩) 등을 다룬다. 하드웨어 기반 솔루션으로는 범용 GPU의 한계를 극복하기 위한 맞춤형 AI 가속기(ASIC, FPGA), 메모리 병목 현상을 해결할 PIM 아키텍처, 그리고 하드웨어-소프트웨어 공동설계 방안을 제시한다.

마지막으로, 앞서 제시된 기술적 해법들이 산업 현장에 성공적으로 안착하고 확산될 수 있도록 뒷받침하는 정책 및 제도적 프레임워크를 제안한다. 저탄소 ICT 분야 연구개발(R&D) 지원 및 금융·세제 인센티브, 데이터센터 에너지 효율 기준 강화 등 규제 체계, 공공부문의 선도적 기술 도입, 그리고 국가 탄소중립목표(NDC)와 연계된 거버넌스 구축 방안 등을 포괄적으로 다룬다.

4. 연구 내용 및 결과

본 연구에서는 디지털 저탄소 전환을 위한 다층적 분석을 통해 다음과 같은 핵심 내용과 결과를 도출하였다. 첫번째로, 디지털 기술의 이중성과 탄소 영향의 심각성을 강조하기 위해 AI 모델 학습(예: LLaMA 3.1 모델 학습 시 8,930톤 이상의 CO_2 배출)과 데이터센터 운영이 에너지 소비 급증의 핵심 원인임을 정량적으로 분석하였다. 특히 단위 기술의 효율성 개선이 전체 데이터 및 연산량의 폭발적 증가를 따라가지 못해 총에너지 소비가 오히려 증가하는 '효율성 역설' 현상을 확인하였다. 또한 AI 하드웨어 제조 공급망이 화석연료 의존도가 높은 지역에 집중되어 있어, 운영 단계(Scope 2)뿐만 아니라 제조 단계(Scope 3)의 탄소 배출 관리 또한 시급한 과제임을 밝혔다.

두번째로, 산업별 저탄소 전환의 실증적 효과를 보이기 위해 데이터센터는 고효율 냉각 기술과 AI 기반 에너지 관리, 재생에너지 직접구매계약(PPA)을 통해 PUE(전력효율지수)를 1.1 이하로 낮출 수 있음을 확인하였다. 클라우드 전환은 온프레미스 대비 최대 98%의 탄소 감축 잠재력을 가지며, 통신망은 AI 기반 기지국 전력 최적화와 레거시 네트워크 종료를 통해 상당한 에너지 절감이 가능함을 국내외 사례를 통해 입증하였다. 또한 제조(AI 공정 최적화), 에너지(스마트그리드), 물류(경로 최적화) 등 주요 산업에서 디지털 기술이 탄소 감축의 핵심 도구로 작용하고 있음을 실증적으로 분석하였다.

세 번째로, 소프트웨어-하드웨어 통합 기술 스택 기술사례를 통해 저탄소화를 위해서는 개별 기술이 아닌, 소프트웨어, 하드웨어, 그리고 이 둘의 공동설계가 상호의존적으로 결합된 다층적 '솔루션 스택'이 필수적임을 제시하였다. 소프트웨어 단에서는 연산량을 근본적으로 줄이는 모델 경량화(양자화, 프루닝)와 에너지 지향적 알고리즘(선형 어텐션, 적응형 디코딩)을, 하드웨어 단에서는 특정 연산에 최적화된 맞춤형 AI 가속기(ASIC)와 메모리 병목을 해결하는 PIM 아키텍처로의 전환을 핵심 전략으로 제시하였다. 이를 통해 하드웨어-소프트웨어 공동설계는 이 둘을 통합하여 에너지 효율을 극대화하는 핵심 원칙임을 밝혔다.

마지막으로, 기술 혁신과 시장 창출을 위한 제도적 기반을 수립하기 위한 정부 주도의 R&D 투자 확대 및 세제·금융 인센티브 제공을 제안하였다. 동시에, 데이터센터 PUE 목표 의무화, 온실가스 배출량 모니터링 및 공시 제도, '친환경 ICT 인증제' 도입 등 명확한 시장

신호를 제공하는 규제 및 표준화 방안을 제시하였다. 결론으로서, 공공부문이 인증된 친환경 클라우드 서비스 및 데이터센터를 의무적으로 도입하여 초기 시장을 창출하고 모범 사례를 확산시키는 리더십을 발휘해야 함을 강조하며, 국가 온실가스 감축목표(NDC) 달성을 위한 ICT 부문의 법적 역할 명확화를 촉구하였다.

5. 기대효과

본 연구 결과는 다음과 같은 다방면의 기대효과를 창출할 수 있다. 먼저, 정책적 기여로는 정부가 '한국형 디지털 저탄소 전환 전략'을 수립하는 데 있어 체계적이고 통합적인 로드맵을 제공한다.

본 연구에서 제시한 기술별·산업별 진단과 정책 제언은 국가 탄소중립 목표 달성을 위한 ICT 부문의 역할을 구체화하고, 관련 법·제도 개선의 실증적 근거 자료로 활용될 수 있다. 산업적 기여로서 국내 기업들이 탄소 감축을 더 이상 비용 부담이 아닌 새로운 성장의 기회이자 미래 경쟁력의 핵심 요소로 인식하는 전환점을 마련해준다. 제시된 구체적인 기술 솔루션과 산업별 적용 사례는 기업들이 저탄소 디지털 인프라에 대한 투자를 확대하고 관련 기술 R&D 방향을 설정하는 데 실질적인 가이드라인을 제공할 것이다. 기술적 기여로서 알고리즘부터 하드웨어 아키텍처에 이르는 저탄소 기술 스택을 체계적으로 정립함으로써, 국내 저전력·고효율 AI 기술 생태계 조성을 촉진하고 미래 기술 R&D 방향성을 제시하는 데 기여한다. 이는 향후 글로벌 기술 표준 경쟁에서 국내 기술 주도권을 확보하는 기반이 될 수 있다. 마지막으로, 사회적 기여로서 디지털 혁신이 가져오는 편익과 환경적 책임을 균형 있게 조망함으로써, 기술 발전과 환경 보존이 공존하는 지속가능한 사회로 나아가기 위한 사회적 공감대를 형성하는 데 기여한다. 이를 통해 미래 세대를 위한 지속가능한 성장 기반을 확보하고, 디지털 강국을 넘어 '지속가능한 디지털 선도국'으로서의 국가 비전을 정립하는 데 이바지할 것이다.

SUMMARY

1. Title

A Study on Technological and Institutional Measures for Digital Low-Carbon Transformation

2. Research Objectives

The background of this research lies in the critical paradox that artificial intelligence(AI) and digital technologies are driving industrial innovation while simultaneously causing exponential increases in energy consumption and carbon emissions. According to the International Energy Agency(IEA), global data center electricity consumption accounted for about 1-1.3% of world electricity demand as of 2022, and this demand is projected to increase explosively by 2030 due to the spread of AI. This is not just a future concern but a current threat. Major economies like the European Union(EU) are already making energy efficiency disclosures mandatory for data centers, setting low-carbon transformation as a new standard for trade and technological competition. Global ICT companies are staking their survival on developing low-power AI accelerators and transitioning to renewable energy, promoting "Performance per Watt" as a core competency. However, in Korea, high-performance AI development and carbon-neutral strategies are not yet organically linked, placing the nation at risk of weakened competitiveness amidst international regulations and market changes.

Therefore, this study aims to comprehensively diagnose the carbon footprint of the digital economy and propose an integrated, multi-layered strategy to achieve both high performance and low carbon. Specifically, it seeks to: ▲ quantitatively diagnose the

carbon impact of digital technologies, ▲ compare and analyze the low-carbon transition policies of major countries, ▲ identify successful cases of digital carbon reduction by industry, ▲ establish a concrete low-carbon technology framework, from AI model lightweighting to custom semiconductor design, and ▲ design a policy and institutional implementation system—including R&D, regulations, and public sector leadership—to support these efforts. Through this, the study intends to present a practical roadmap for securing Korea's future digital leadership within a sustainable digital ecosystem.

3. Contents and Scope of the Research

The composition of this study is designed with a total of five chapters to analyze the problem of digital low-carbon transformation from multiple perspectives and propose an integrated solution. Each chapter follows an organic flow of 'Diagnosis - Strategy - Technology - Policy.'

First, the background and objectives of the research are presented, and the research scope is set based on the OECD's framework for quantifying the digital economy. It quantitatively diagnoses the energy consumption and carbon emissions status at each stage of digital technology—including AI model training/inference, data centers, communication infrastructure, and ICT terminal manufacturing. It also analyzes regulatory and policy trends in major countries like the EU and the US to identify the gap with the domestic situation.

Second, it addresses two aspects of carbon reduction using digital technology. It first analyzes domestic and international cases of low-carbon transformation methods that optimize the energy efficiency of digital infrastructure itself, such as data centers, cloud services, and communication networks. Next, it presents in-depth case studies of specific carbon reduction applications where digital technologies like AI and digital twins have achieved process efficiency and energy savings in key industries such as manufacturing, energy, and logistics.

Third, it systematically analyzes key technological solutions for low-carbonization, dividing them into software and hardware. Software-based solutions cover model lightweighting(quantization, pruning, merging) and energy-oriented algorithms(linear attention, neural architecture search, adaptive decoding). Hardware-based solutions include custom AI accelerators(ASIC, FPGA) to overcome the limitations of general-purpose GPUs, PIM architecture to solve the memory bottleneck, and hardware-software co-design methods.

Finally, it proposes a policy and institutional framework to support the successful adoption and spread of the aforementioned technological solutions in the industry. This comprehensively covers R&D support and financial/tax incentives for the low-carbon ICT sector, regulatory systems like strengthening data center energy efficiency standards, pioneering technology adoption by the public sector, and governance-building measures linked to the national carbon-neutrality goals(NDC).

4. Research Results

In this study, the following key contents and results were derived through a multi-layered analysis for digital low-carbon transformation. First, to emphasize the duality of digital technology and the severity of its carbon impact, we quantitatively analyzed that AI model training(e.g., LLaMA 3.1 model training emitting over 8,930 tons of CO2) and data center operations are key drivers of the surge in energy consumption. We particularly identified the 'efficiency paradox,' where improvements in unit technology efficiency cannot keep up with the explosive growth in total data and computation, leading to an increase in total energy consumption. It was also revealed that the AI hardware manufacturing supply chain is concentrated in regions heavily dependent on fossil fuels, making the management of manufacturing-stage(Scope 3) carbon emissions, not just operational-stage(Scope 2), an urgent task.

Second, to show the empirical effects of low-carbon transformation by industry, it was confirmed that data centers can lower their PUE(Power Usage Effectiveness) to 1.1 or less through high-efficiency cooling technology, AI-based energy management, and renewable energy Power Purchase Agreements(PPAs). Cloud migration has the potential for up to 98% carbon reduction compared to on-premise systems. It was also demonstrated through domestic and international cases that communication networks can achieve significant energy savings through AI-based base station power optimization and decommissioning of legacy networks. Furthermore, it was empirically analyzed that digital technology acts as a key tool for carbon reduction in major industries such as manufacturing(AI process optimization), energy(smart grids), and logistics(route optimization).

Third, through technological case studies of the software-hardware integrated technology stack, we proposed that low-carbonization requires not individual technologies, but a multi-layered 'solution stack' where software, hardware, and their co-design are interdependently combined. We presented model lightweighting(quantization, pruning) and energy-oriented algorithms(linear attention, adaptive decoding) at the software level to fundamentally reduce computations, and a shift to custom AI accelerators(ASIC) optimized for specific operations and PIM architecture to solve memory bottlenecks at the hardware level as key strategies. This revealed that hardware-software co-design is the core principle for maximizing energy efficiency by integrating these two.

Finally, to establish an institutional foundation for technological innovation and market creation, we proposed the expansion of government-led R&D investment and the provision of tax and financial incentives. Simultaneously, we suggested regulatory and standardization measures that provide clear market signals, such as mandating data center PUE targets, a greenhouse gas emissions monitoring and disclosure system, and the introduction of an 'Eco-friendly ICT Certification.' In conclusion, we emphasized that the public sector must demonstrate leadership by mandating the adoption of certified eco-friendly cloud services and data centers to create an initial market and spread best

practices, urging the clarification of the ICT sector's legal role in achieving the national greenhouse gas reduction targets(NDC).

5. Expectations

The results of this study can create the following multifaceted expected effects.

First, as a policy contribution, it provides a systematic and integrated roadmap for the government to establish a 'Korean-style Digital Low-Carbon Transformation Strategy.' The diagnosis by technology and industry, along with the policy proposals presented in this study, can specify the role of the ICT sector in achieving national carbon-neutrality goals and be used as empirical data for improving related laws and institutions.

As an industrial contribution, it provides a turning point for domestic companies to recognize carbon reduction not as a cost burden, but as a new growth opportunity and a core element of future competitiveness. The specific technology solutions and industry-specific application cases presented will offer practical guidelines for companies to expand investment in low-carbon digital infrastructure and set R&D directions for related technologies.

As a technological contribution, it systematically establishes a low-carbon technology stack ranging from algorithms to hardware architecture, thereby promoting the creation of a domestic low-power, high-efficiency AI technology ecosystem and contributing to presenting future technology R&D directions. This can serve as a foundation for securing domestic technology leadership in the global technology standards competition.

Finally, as a social contribution, it helps form a social consensus for moving toward a sustainable society where technological advancement and environmental preservation coexist by providing a balanced perspective on the benefits of digital innovation and its environmental responsibilities. Through this, it contributes to securing a sustainable growth foundation for future generations and establishing a national vision not just as a digital powerhouse, but as a 'Sustainable Digital Leading Nation.'

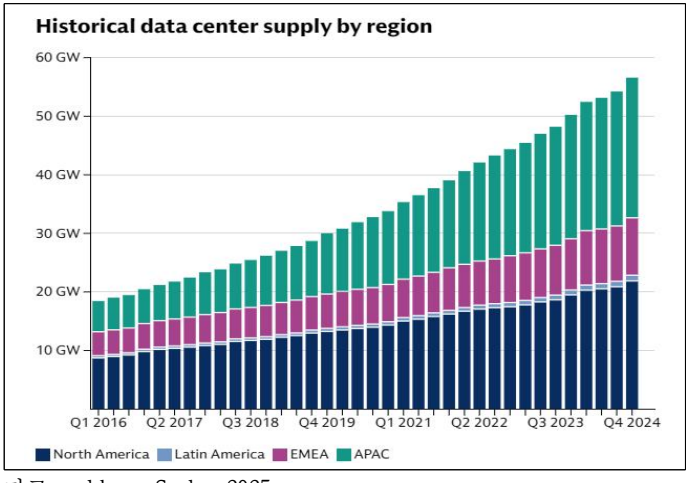
CONTENTS

제1장 서 론

제1절 연구의 배경

AI는 정말 무한한 가능성만을 남길까, 아니면 새로운 한계를 드러낼까. 초거대 언어모델(LLM)과 생성형 AI는 산업과 사회 전반을 혁신하는 동력으로 자리 잡았다. 그러나 그 뒤편에는 눈에 보이지 않는 대가가 존재한다. 학습과 추론을 위해 끊임없이 돌아가는 연산 장치, 이를 식히고 유지하기 위해 하루 24시간 가동되는 데이터센터, 그리고 그 모든 것을 움직이는 막대한 전력 소비가 그것이다. 기술의 비약적 도약은 곧바로 전력 수요의 폭발로 이어지고, 이는 탄소 배출의 가속으로 연결된다. 국제에너지기구(IEA)에 따르면, 2022년 기준 전 세계 데이터센터에서 소비한 전력은 세계 전력 수요의 약 11.3%에 해당한다(IEA, 2023). 이러한 규모는 일본의 연간 전력 소비량에 육박하는 수준이며, Goldman Sachs는 2030년까지 글로벌 데이터센터 전력 수요가 크게 증가할 것이라고 전망했다〔그림 1-1〕.

〔그림 1-1〕 데이터센터 공급량 변화



자료: goldman Sachs, 2025

세계 각국은 이미 이 변화의 파급력을 감지하고 있다. 유럽연합은 데이터센터의 에너지 사용과 효율성을 공개하도록 의무화하며, 디지털 인프라의 지속 가능성을 법제화했다. 미국 또한 전력 수급과 탄소 배출을 동시에 관리하는 규제를 준비 중이다. 글로벌 ICT 기업들은 성능과 효율이라는 두 축을 모두 잡기 위해 AI 가속기 개발과 재생에너지 전환 계획을 경쟁적으로 내놓고 있다. 이는 기술 경쟁의 새로운 무대가 성능 향상뿐 아니라 '에너지 효율 확보'로 이동하고 있음을 보여준다.

글로벌 ICT 기업들의 대응은 이미 시작됐다. 구글과 마이크로소프트는 2030년까지 데이터센터 전력의 100%를 무탄소 에너지로 전환하겠다고 선언했고, 엔비디아와 AMD는 차세대 AI 가속기의 성능지표로 와트당 처리 효율을 핵심 경쟁 요소로 삼고 있다(Microsoft, 2024; The Guardian, 2025). 이는 단순한 친환경 홍보가 아니라, 국제 규제 강화와 전력 비용 상승, 그리고 투자자·소비자의 ESG[1] 요구에 대응하기 위한 필연적 선택이다. 다시 말해, 기술 경쟁의 무대는 더 이상 연산 속도와 모델 크기만으로 결정되지 않으며, 동일한 성능을 더 적은 에너지로 구현하는 능력이 곧 미래 시장의 주도권을 좌우하게 될 것이다.

그렇다면 우리는 이 흐름에 어떻게 대응해야 할까? 국내에서는 고성능 AI 개발과 디지털 인프라 확충이 여전히 탄소중립 전략과 유기적으로 결합되지 못하고 있다. 전력 생산 구조의 제약, 하드웨어 설계의 비효율, 알고리즘 경량화 수준의 격차, 네트워크 운용에서의 에너지 부담 등 해결해야 할 과제가 여전히 산적해 있다. 지금 이 시점에서 성능과 효율이라는 두 축을 모두 잡을 한국형 디지털 저탄소 전환 전략을 마련하지 않는다면, 국제 규제와 시장 속에서 경쟁력이 약화될 위험이 크다.

본 연구는 이러한 문제의식에서 출발한다. AI·데이터센터·네트워크 전반의 에너지 소비와 탄소 배출 구조를 면밀히 진단하고, 성능을 유지한 채 효율성을 극대화할 수 있는 기술·정책 조합을 제시하는 것. 이는 단순한 절감 노력이 아니라, 미래 기술 패러다임의 주도권을 확보하기 위한 전략적 선택이다. 이제 우리는 물어야 한다. "고성능과 저탄소, 두 마리 토끼를 동시에 잡을 준비가 되었는가?"

---

1) ESG: 환경(Environment), 사회(Social), 지배구조(Governance)의 약자로, 기업의 지속가능성과 사회적 책임을 평가하는 비재무적 경영·투자 지표이다.

제 2 절 연구의 목표

본 연구의 목표는 급속히 확산되는 디지털 기술이 초래하는 에너지 소비와 탄소 배출의 구조를 다각도로 진단하고, 고성능과 저탄소를 동시에 실현할 수 있는 기술적·정책적 해법을 제시하는 것에 있다. 인공지능(AI), 데이터센터, 통신 인프라, ICT 제조 등 디지털 산업 전반은 4차 산업혁명의 핵심 동력으로 자리 잡았으나, 그 이면에서는 전력 수요의 폭증과 온실가스 배출 증가라는 새로운 환경적 부담이 심화되고 있다. 특히 초거대 언어모델(LLM)과 생성형 AI는 연산 집약적 특성으로 인해 에너지 사용량을 기하급수적으로 끌어올리고 있으며, 데이터센터의 냉각과 유지에 필요한 전력까지 포함할 경우 그 탄소 영향은 산업 전반에 파급력을 미치고 있다. 그럼에도 불구하고 국내에서는 아직 이러한 디지털 기술의 탄소 발자국[2]을 체계적으로 측정하고 관리하기 위한 기반이 충분히 마련되어 있지 않다.

이에 본 연구는 첫째, 디지털 기술의 확산이 에너지 시스템과 탄소 배출 구조에 미치는 영향을 정량적으로 진단하고, 주요 기술별·산업별 배출 현황을 실증적으로 분석한다. 둘째, EU·OECD 등 주요국의 저탄소 디지털 전환 정책과 국제 표준화 동향을 비교·검토하여, 글로벌 규제 변화에 대응할 수 있는 국내 전략적 시사점을 도출한다. 셋째, AI·데이터센터·제조·물류 등 핵심 산업 영역에서의 디지털 기반 탄소 감축 사례를 발굴·평가하고, 에너지 효율 개선 및 운영 최적화를 통한 저감 효과를 구체적으로 제시한다. 넷째, 경량화 모델 개발, AI 가속기 최적화, 메모리 아키텍처 개선 등 AI 기반 저탄소화 기술체계를 정립하여, 기술적 효율성과 지속가능성을 병행할 수 있는 방향을 탐색한다. 마지막으로, 연구개발 지원, 에너지 효율 기준 강화, 공공부문 선도 도입 등 정책·제도적 추진체계를 설계함으로써, 국가 차원의 디지털 탄소중립 거버넌스 모델을 제안한다.

결국 본 연구는 디지털 기술이 산업 경쟁력의 원천이자 새로운 환경 부담의 주체로 동시에 작용하는 시대적 전환점에서, 기술 혁신과 탄소 저감이 상충되지 않는 지속 가능한

2) 탄소 발자국(Carbon Footprint): 제품, 서비스, 또는 사람의 활동이 원료 채취부터 폐기까지 전 과정에서 배출하는 온실가스의 총량을 나타내는 지표. 지구 온난화에 미치는 영향을 정량화하여 나타낸 것으로, 탄소와 발자국을 합성한 용어.

디지털 생태계의 방향을 모색하고자 한다. 나아가 본 연구를 통해 한국형 디지털 저탄소 전환 전략 근거를 마련하고, 향후 글로벌 탄소 규제와 기술 경쟁이 교차하는 지점에서 국가 산업이 주도적으로 대응할 수 있는 전략적 로드맵을 제시하는 것을 최종 목표로 한다.

제3절 연구의 프레임워크

본 연구는 디지털 기술이 초래하는 탄소 배출 구조를 다차원적으로 진단하고, 산업별·기술별 저감 방안을 체계적으로 제시하기 위해 〔그림 1-2〕와 같은 연구 프레임워크를 구성하였다. 이 프레임워크는 단순한 현황 분석을 넘어, 기술적 효율화와 제도적 실행 체계를 아우르는 통합적 접근을 지향한다. 연구의 전체 흐름은 크게 ① 탄소배출 현황 분석 → ② 제도적 방안 도출 → ③ 저탄소 기술 제시 → ④ 이행 로드맵 설계의 4단계로 구성되며, 각 단계는 상호 연계되어 순환적 개선 구조를 형성한다.

첫째, 탄소배출 현황 분석 단계에서는 디지털 경제 전반의 에너지 소비와 탄소 배출 구조를 정밀하게 진단한다. 이를 위해 국제 표준화 논의와 계량적 모델을 검토하고, 디지털 기술의 범위를 명확히 정의한 후, AI 학습·추론 과정과 데이터센터, 통신 인프라, ICT 단말 제조 등에서 발생하는 에너지 사용량과 배출 특성을 정량적으로 분석한다. 이 과정에서는 국제에너지기구(IEA), 글로벌 ICT 기업의 ESG 보고서, 국내 에너지 통계 등을 교차 활용하여, 국내 디지털 부문의 탄소 배출 현황과 글로벌 수준에서의 상대적 위치를 규명한다. 이러한 진단은 연구의 후속 단계인 기술적 대응 및 정책 설계의 자료로 기능한다.

둘째, 제도적 방안 도출 단계에서는 디지털 기술의 탄소 절감을 위한 제도적 기반과 정책적 수단을 모색한다. 유럽연합(EU)의 데이터센터 에너지 효율 공개 의무제, 미국의 그린 데이터센터 이니셔티브, OECD의 디지털 탄소 계량화 표준 등 해외 사례를 분석하여, 국내 정책 설계에 적용 가능한 시사점을 도출한다. 또한 한국형 표준화 정책(K-RE100 등)의 방향성을 검토하고, 에너지 효율 기준의 단계적 강화, ICT 부문 온실가스 모니터링, 친환경 인증제 도입 등 실행 가능한 제도적 수단을 제시한다. 이를 통해 디지털 기술 발전과 탄소 절감이 병행될 수 있는 제도적 틀을 마련하고자 한다.

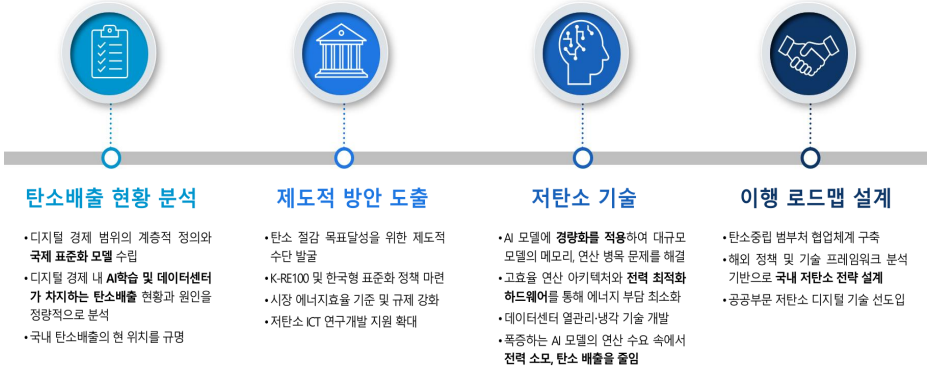
셋째, 저탄소 기술 단계에서는 AI 및 디지털 인프라의 효율성을 높이기 위한 구체적 기술적 접근 방안을 제시한다. AI 모델 경량화(Compression, Quantization, Pruning 등) 기술을 적용하여 연산 및 메모리 부담을 경감하고, 고효율 연산 아키텍처 및 전력 최적화 하드웨어를 통해 에너지 소모를 최소화한다. 또한 데이터센터의 냉각·열관리 시스템의 효율 향상, 클라우드 전환 및 분산형 에너지 관리 시스템을 활용한 전력 절감 기술 등을 탐색한다. 이러한 기술들은 단순한 에너지 절감이 아니라, 향후 대규모 AI 운용 환경에서 전력 대

비 성능 효율을 극대화하기 위한 핵심 수단으로 작용한다.

넷째, 이행 로드맵 설계 단계에서는 앞선 분석과 기술적 대안을 종합하여, 디지털 탄소중립 달성을 위한 실행체계와 협업 구조를 제시한다. 범부처 협력 거버넌스 구축, 공공부문 선도 도입, 민간 참여 확대 등을 중심으로 단계별 실행 로드맵을 설계하고, 해외 정책 및 기술 프레임워크와의 정합성을 검토한다. 또한 공공조달, 기술 실증사업, 인센티브 제도 등 정책적 촉진수단을 포함하여, 실질적 저탄소 전환이 가능한 실행 경로를 제시한다.

결과적으로, 〔그림 1-2〕의 프레임워크는 디지털 기술의 탄소 영향을 "진단-대응-혁신-실행"의 순환 구조로 통합한 체계적 접근 모델이다. 본 연구는 이를 통해 탄소 절감과 기술 경쟁력 제고가 병립하는 한국형 디지털 탄소중립 모델을 제시하고자 하며, 각 단계별 연구 결과는 상호 보완적으로 작동하여 정책적·산업적 실행력을 확보하는 것을 목표로 한다.

〔그림 1-2〕 연구의 프레임워크



자료: KISDI, 2025

제2장 디지털 기술의 탄소 영향 진단

제1절 AI 등 주요 디지털 기술의 에너지 소비 및 탄소 배출 현황

1. 디지털 기술 확산과 환경적 영향

가. 디지털 기술 확산에 따른 글로벌 에너지 소비 구조 변화와 온실가스 배출 증가

인공지능(AI), 클라우드 컴퓨팅, 사물인터넷(IoT), 5G 통신 등 첨단 디지털 기술은 고성능 반도체의 발전, 초고속 네트워크 보급, 데이터 중심 경제의 부상에 힘입어 빠르게 확산되고 있다. 특히 코로나19 팬데믹은 비대면·온라인 중심의 생활과 산업활동을 촉진하며 디지털 전환을 가속화하였다. 팬데믹 종료 이후에도 이러한 확산세는 둔화되지 않고, 오히려 팬데믹 기간에 구축된 디지털 인프라의 지속 활용, 인공지능과 데이터 분석의 산업 표준화, 각국 정부와 기업의 전략적 투자, 그리고 소비자들의 디지털 서비스 기대 수준 상승 등이 맞물려 고속 성장 구간을 유지하고 있다. 이러한 변화 속에서 디지털 기술은 현대 사회의 핵심 인프라로 자리 잡았지만, 동시에 에너지 소비 증가와 탄소 배출 확대라는 새로운 환경 문제를 야기하고 있다.

국제전기통신연합(ITU)과 세계 벤치마킹 얼라이언스(WBA)의 공동 연구 보고서에 따르면, 디지털 산업 내 200개 주요 기술 기업이 배출한 온실가스는 약 2억 9천만 톤으로, 전년 대비 1.4% 증가했으며 전 세계 에너지 관련 배출량의 약 0.8%를 차지한다.[3] 이는 아르헨티나, 볼리비아, 칠레의 연간 배출량을 합한 규모와 같다. 이러한 결과는 AI 구동을 위한 데이터센터, 통신망, AI 연산 인프라와 같은 다양한 ICT 기반 시설의 전력 수요의 급증이 주요 원인이다. AI의 급속한 성장은 에너지 수요와 배출량 추이에 영향을 미칠 것으로 예상되며, 이 부문에서 과학에 기반한 강력하고 시급한 전략을 채택해야 할 필요성이 커지고 있다.

---

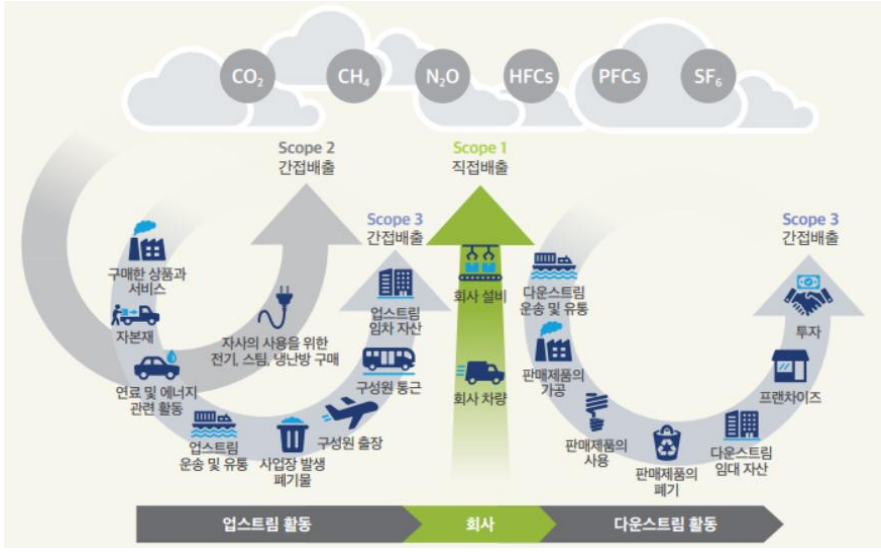
3) ITU and WBA, Greening Digital Companies Report, 2025

나. 온실가스 배출량 측정 기준

온실가스 배출량 산정은 세계자원연구소(WRI)와 세계지속가능기업협의회(WBCSD)가 공동 개발한 국제 표준인 온실가스 프로토콜(GHG Protocol)에 기반하며, 이는 파리 협정 등 국제 기후 협약뿐 아니라 기업의 배출 관리와 감축 목표 수립에도 활용된다〔그림 2-1〕.

Scope 1은 기업이 보유 및 통제하는 설비에서 직접 배출되는 온실가스를 의미하며, 예를 들어 화석연료를 사용한 공정에서 발생하는 배출이 이에 해당한다. Scope 2는 외부에서 구매한 전기, 열, 스팀 등을 사용하는 과정에서 간접적으로 발생하는 배출을 포함한다. Scope 3는 밸류체인 전반에서 발생하는 기타 간접 배출로, 부품 생산, 물류, 제품 사용 및 폐기 등 기업의 사업장 밖에서 발생하지만 기업 활동과 직접적으로 연관된 배출을 의미한다.

〔그림 2-1〕 온실가스 프로토콜



자료: 탄소중립녹색성장위원회

다. 디지털 기술 확산의 환경적 파급효과

디지털 기술의 확산은 전력 수요 증가를 통해 직·간접적인 환경 영향을 초래한다. 특히 데이터센터는 막대한 전력을 소모할 뿐 아니라 냉각 과정에서도 상당한 전력과 수자원을 사용한다. 이러한 냉각 과정의 추가 전력 소모는 전체 에너지 소비를 증가시켜 탄소 배출

확대에 기여하며, 동시에 지역 수자원 관리에도 부담을 준다. 또한 서버와 반도체 칩의 제조, 운송, 폐기 등 전 주기 단계에서 상당한 온실가스가 배출된다.

국제에너지기구(IEA)는 2030년까지 데이터센터 전력소비는 2024년 소비의 두 배가 넘는 약 945TWh으로 전망했다.[4] 또한 일부 지역에서는 전력망 제약과 발전 인프라 부족으로 인해 데이터센터 전력 수요를 충당하는 과정에서 화석연료 발전 비중이 높아지고 있다. 예를 들어, 미국과 중국은 2030년까지 데이터센터 전력 수요 증가분의 상당 부분을 가스와 석탄 발전으로 공급할 것으로 예상되며, 이에 따라 해당 지역의 전력 부문 탄소 배출이 단기적으로 확대되는 양상이 관측되고 있다.

또한 AI 산업은 데이터센터 운영뿐 아니라 제조 공급망 단계에서도 높은 탄소발자국을 남긴다. 최근 연구에 따르면, GPU, TPU, HBM 등 AI 가속기 전반의 생애주기(Life-Cycle)를 고려했을 때 제조 및 자재 생산 단계에서 약 25% 수준의 탄소 배출이 발생하며, 나머지 70~90%는 운영 단계에서의 전력 소비에 기인하는 것으로 나타났다(Schneider, et al.(2025)) 〔그림 2-2〕. 이는 AI 가속기와 데이터센터의 확산이 단순한 운용 효율 개선만으로는 환경적 부담을 줄이기 어렵고, 제조 단계의 배출까지 포괄하는 전주기적 관리가 필요함을 시사한다. AI 구동에 필수적인 고성능 GPU와 HBM 등 반도체는 에너지 집약적 제조 공정을 거치는데, 주요 생산 거점인 대만(TSMC), 한국(삼성전자, SK 하이닉스), 일본·미국(마이크론 등)의 전력망은 화석연료 의존도가 높아 탄소 집약도가 상대적으로 크다. 2023년 기준 전력망의 화석연료 발전 비중은 한국 58.5%, 일본 68.6%, 대만 83.1%에 달했다. 대규모 반도체 파운드리는 시간당 최대 100MWh에 이르는 전력을 소비하기도 하며[5], 이러한 제조 과정에서 발생하는 배출은 AI 기업의 Scope 3 항목 중 '상품 및 서비스 구매' 부문에 반영된다. 실제로 2023 회계연도 기준 주요 AI 빅테크 기업들의 공급망 배출량은 운영상의 직/간접 배출량보다 더 큰 비중을 차지하고 있으며〔그림 2-3〕, 이는 AI 산업 전반에서 공급망 저탄소화 전략이 시급함을 보여준다.

데이터센터와 ai 외에도, 다른 디지털 기술의 확산 역시 환경적 부담을 키우고 있다. 예

4) IEA(2025.04), Energy and AI

5) Greenpeace(2025), 인공지능(AI) 시대의 그림자: 반도체 제조산업의 전력 소비량과 온실가스 배출량 분석

를 들어 블록체인과 암호화폐 채굴은 24시간 고성능 연산 장비를 가동해야 하며, 일부 국가에서는 국가 단위 전력 소비의 수 퍼센트를 차지할 정도로 에너지 소모가 크다. 사물인터넷 기기의 대량 보급은 배터리 생산과 교체에 따른 자원 채굴 및 폐기 부담을 증가시키고, 수십억 대의 장치가 상시 네트워크에 연결되면서 누적 전력 소비가 커진다. 또한 5G 통신망은 기존 4G 대비 기지국 수와 밀도가 높아 운영 전력 수요가 늘어나며, 통신 장비 제조 과정에서도 온실가스가 배출된다[그림 2-4]. 이와 더불어, 5G 시대에는 데이터 트래픽의 급증으로 인해 전력 소비가 크게 증가하고 있다. 5G 기지국은 약 3,700 W 수준의 전력을 소모하며, 이는 4G 대비 약 2.5~3.5배 높은 수준으로 보고된다.[6] 이러한 전력 증가는 비단 통신 인프라뿐 아니라, 데이터 전송량의 증가로 이어지며 네트워크 전체의 에너지 부담을 가중시킨다.

한편, 비디오 스트리밍 서비스의 보급 확대는 인터넷 트래픽 구조를 송두리째 바꾸고 있다. DemandSage 통계에 따르면, 2025년까지 인터넷 트래픽의 약 82%가 영상 콘텐츠가 될 것으로 예측되고 있으며, 사람들은 하루 평균 약 100분을 온라인 영상 시청에 소비한다고 한다. 이러한 현상은 스트리밍 중심의 서비스가 일반화되었음을 보여준다.[7].

나아가 5G 기술은 비트당 에너지 효율성 측면에서는 4G보다 향상되었지만, 전송되는 데이터의 양이 폭발적으로 늘면서 결과적으로는 더 많은 전력을 요구하는 역설적인 현상이 나타나고 있다. 즉, 비디오 스트리밍 기반 서비스의 확산과 해상도 향상으로 인해, 에너지 효율의 개선 폭보다 트래픽 증가 폭이 더 커져 총 전력 수요가 오히려 확대되는 구조가 형성된 것이다.[8] 이처럼 디지털 기술 전반이 에너지 소비와 자원 사용을 동반하면서, 지속가능성을 확보하기 위한 기술적·정책적 대응의 중요성이 커지고 있다.

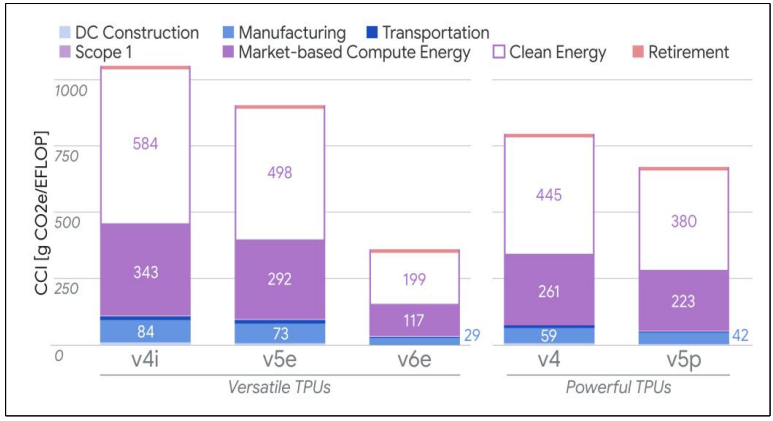
---

6) YiNGDA(2025), Why does 5g base station consume so much power and how to improve it?

7) DemandSage(2025) 93 Video Marketing Statistics 2025

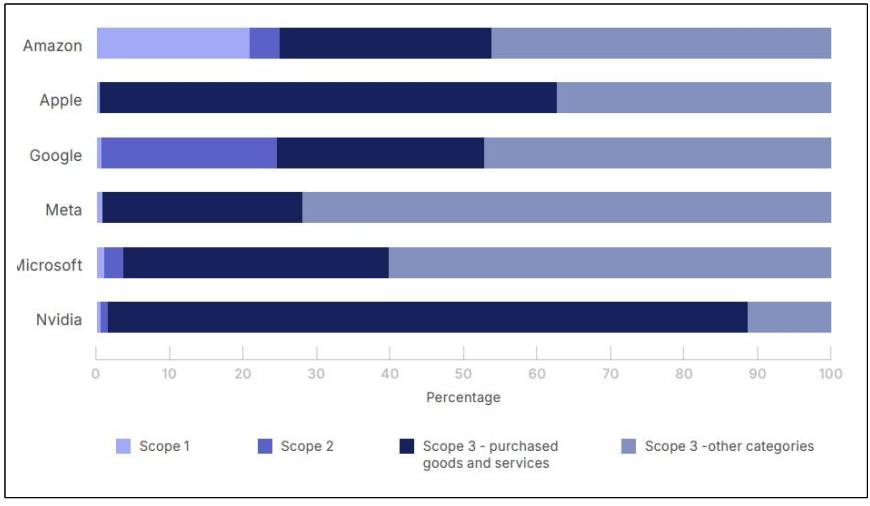
8) RCR Wireless News(2022) Parsing the 5G power equation: Is 5G actually greener?

〔그림 2－2〕 AI 하드웨어의 전 생애주기별 계산 탄소집약도(CCI) — 단계별 배출 구성비



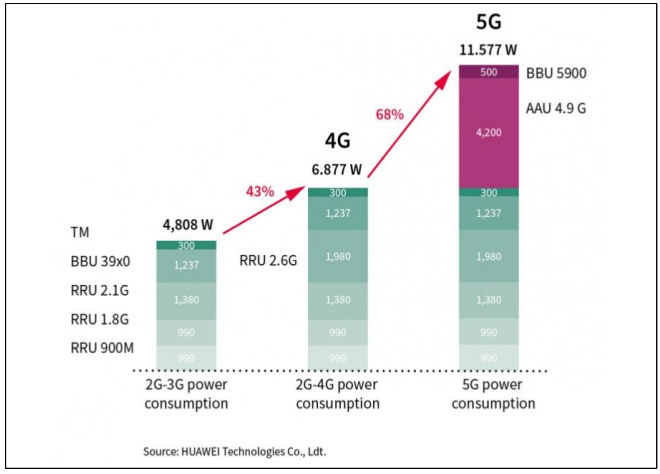
자료: Life-Cycle Emissions of AI Hardware: A Cradle-To-Grave Approach and Generational Trends(Schneider, et al.(2025))(arXiv'25) p.5

〔그림 2－3〕 2023 회계연도 기준 주요 AI 빅테크 기업들의 배출량 구성



자료: Greenpeace(2025)

〔그림 2－4〕 세대별 통신 기지국의 전력 소비 비교



자료: 화웨이 테크놀로지스

2. AI · 데이터 인프라의 에너지 소비 및 탄소 배출 현황

가. AI 모델의 학습 및 추론 단계의 에너지 소비와 환경 영향

AI 기술의 급속한 발전과 확산은 전 세계 전력 수요를 폭발적으로 증가시키고 있다. 향후 5~10년 내 글로벌 데이터 전력 수요는 두 배 이상 확대될 것으로 전망되며, 이에 따라 전체 전력 공급에서 AI 인프라가 차지하는 비중도 급격히 높아질 것으로 예상된다. IEA는 전 세계 데이터센터의 전력 소비 비중이 2022년 1%에서 2030년 3%로 상승할 것이라고 전망하고 있다. 또한 Statista에 따르면 2022년 당시 전 세계 데이터센터의 전력 소비량은 프랑스 전체 소비량과 비슷한 수준이었으며, 2026년에는 일본의 전력 소비량에 근접할 것으로 예상된다〔그림 2－5〕.

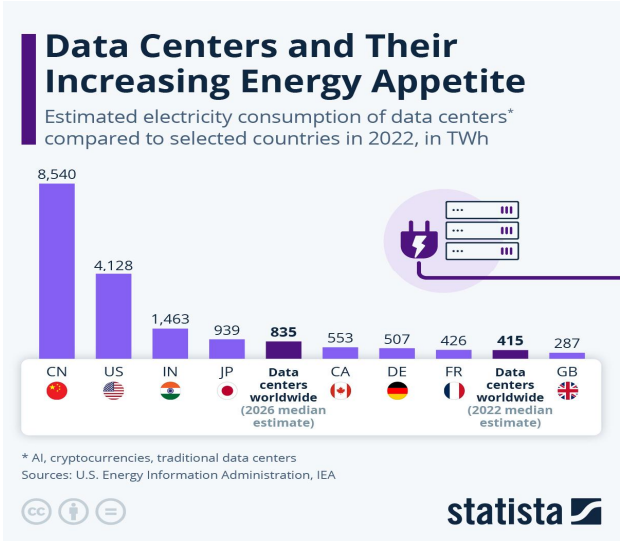
AI 연산의 에너지 소비는 학습 단계와 추론 단계에서 뚜렷한 차이를 보인다 〈표 2－1〉. 딥러닝 모델의 규모가 커질수록 학습 과정에서 요구되는 연산량과 전력 사용량은 기하급수적으로 증가한다. 예를 들어, 1750억 개의 파라미터를 가진 GPT-3 학습에는 약 1,287 MWh의 전력이 소모된 것으로 추정되는데, 이는 평균적인 미국 가정이 약 120년 동안 사용하는 전력량과 비슷하다.[9] 스탠퍼드 대학 연구에 따르면 이 과정에서 약 502톤의 이산화탄소가 배출되었으며, 이는 개인 연간 배출량(5.51톤)의 약 91배, 뉴욕-샌프란시스코 왕복

편 승객 1명이 배출하는 양(0.99톤)의 약 507배에 해당한다〔그림 2-6〕. 이처럼 전력 소비는 곧 탄소 배출로 이어진다.

추론 단계는 단일 요청당 전력 소모는 상대적으로 적지만, 서비스 요청 규모가 방대할 경우 상시 전력 부하가 크게 증가한다. 예를 들어 GPT-4o의 경우 단일 쿼리당 평균 0.3Wh를 사용하지만[10], 월 10억 건 이상의 요청이 발생하는 서비스 환경에서는 연간 수천 MWh에 달하는 전력을 소비하게 된다. 실제로 AI 서비스 전체 수명주기에서 약 80%의 배출량이 추론 단계에서 발생하며, 학습 단계는 약 20%를 차지한다.[11] 즉, 개별 요청의 효율성이 높더라도 대규모 서비스 환경에서는 추론 과정이 무시할 수 없는 환경 부담을 초래한다.

이처럼 AI 관련 컴퓨팅은 기존 IT 워크로드 대비 훨씬 높은 연산 밀도와 전력 소모를 수반하며 이는 탄소 배출량 확대와 직결된다.

〔그림 2-5〕 전세계 데이터 센터와 각국의 소비 전력량 비교



자료: Statista(2024), How Energy Intensive Are Data Centers?

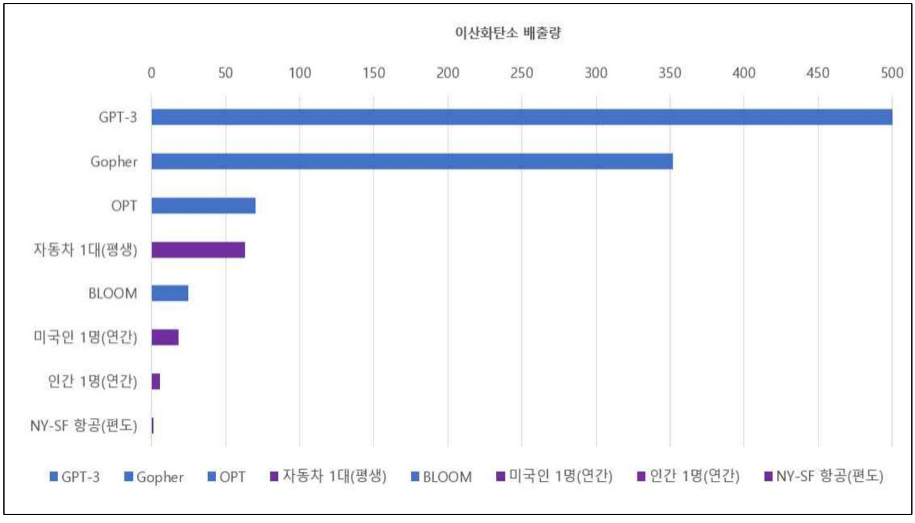
9) ADaSci(2024), How Much Energy Do LLMs Consume? Unveiling the Power Behind AI
10) EPOCH AI(2025), How much energy does ChatGPT use?
11) EMERGING TECHNOLOGIES(2024), How to manage AI's energy demand-today, tomorrow and in the future

〈표 2－1〉 모델 크기에 따른 예상 에너지 소비량

Model Size (Parameters)	Computational Resources	Training Duration (Hours)	Infrastructure	Training Energy (MWh)	Evaluation Energy (MWh)
7B	8 GPUs (NVIDIA V100)	336	Cloud	50	5
40B	64 GPUs (NVIDIA V100)	672	Cloud	200	10
100B+	1024 GPUs (NVIDIA A100)	1344	Cloud	1,287	50

자료: ADaSci(2024), How Much Energy Do LLMs Consume? Unveiling the Power Behind AI

〔그림 2－6〕 인공지능 모델과 실생활 이산화탄소 배출량 비교(톤 단위)



자료: 한국지능정보사회진흥원(NIA, 2025), 디지털 분야의 탄소배출과 대응 노력

나. 데이터센터 전력 구성과 탄소 집약

대규모 AI 학습 및 추론 워크로드는 주로 클라우드와 온프레미스 데이터센터에서 처리된다. 데이터센터의 전력 사용은 크게 IT 장비(서버·스토리지·네트워크), 냉각 시스템, 전

력 변환·분배 장치로 구분된다. PUE(Power Usage Effectiveness) 지표는 데이터센터 효율성을 평가하는 핵심 지표로, 총 전력 소비량을 IT 장비 소비량으로 나눈 값이며 1에 가까울수록 효율이 높다. 2021년 기준, 글로벌 대규모 데이터센터의 평균 PUE는 1.57로 보고되었다.[12)]

하지만 효율이 높아도 전체 전력 사용량이 늘어나면 탄소 배출이 함께 증가한다. 탄소 배출량은 국가별 전력 생산 구조(재생에너지·화석연료 비중 등)에 따라 달라지므로, 동일한 전력 사용량이라도 국가별 탄소 집약도는 큰 차이를 보인다〔그림 2-7〕.

다. 글로벌 데이터센터 전력 소비 추세와 AI로 인한 전력 소비 증가

IEA에 따르면 2024년 기준 전 세계 데이터센터의 연간 전력 소비량은 415TWh로, 이는 전 세계 전력 소비량의 약 1.5%에 해당한다. 최근 5년간 데이터 센터의 전력 소비는 연평균 12%의 빠른 증가세를 보였으며, 2030년에는 약 945TWh에 이를 것으로 전망된다.[13)]

세계 주요 기관들도 공통적으로 AI의 급성장이 데이터센터 전력 수요를 가속한다고 분석한다 〈표 2-2〉. 특히 중국과 미국은 데이터센터 전력 소비 증가율이 가장 높은 지역으로, 2030년까지 전 세계 전력 소비 증가분의 약 80%를 차지할 것으로 예상된다. 미국과 중국의 전력 소비량은 2024년 대비 각각 130%, 170% 증가하고, 유럽도 70% 증가할 것으로 전망된다〔그림 2-8〕.

AI 모델의 대규모 학습과 고빈도 추론 서비스 확산은 GPU 클러스터 확장을 촉발하고, 이에 따라 냉각·전력 인프라 강화가 필수적으로 뒤따른다. 결과적으로 AI 산업의 확장은 데이터센터 전력 소비 구조 전반에 직접적인 영향을 미친다.

12) 디지털데일리(2022), [데이터센터 퀀텀점프] 우후죽순 늘어나는 데이터센터, '지속가능성'이 화두

13) IEA(2025.04), Energy and AI

〔그림 2－7〕 주요국의 탄소집약도 현황 및 개선속도 비교

(단위: CO_2 kg per USD)

구분	한국	일본	이탈리아	프랑스	미국	독일	영국
집약도	0.14	0.12	0.09	0.08	0.11	0.09	0.09
개선속도	2.4%	2.7%	3.7%	4.6%	4.9%	5.5%	5.9%

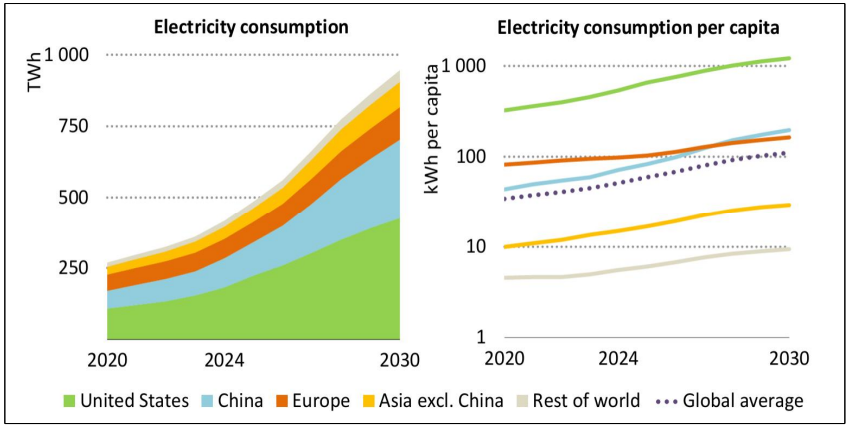
자료: 한국경제인협회(2024), 미국 청정경쟁법의 국내 파급효과 및 시사점

〈표 2－2〉 인공지능 관련 글로벌 전력 수요 전망

구분	기준	현재 수요	예상 수요	성장률
국제에너지기구 (IEA)	글로벌 데이터센터	460TWh(2022)	1,000~2,000TWh(2030)	117~335%
세미애널리시스 (SemiAnalysis)	글로벌 AI 데이터센터	49GW(2023)	74~132GW(2028)	196%
보스턴 컨설팅그룹 (BCG)	글로벌 AI 데이터센터	60GW(2023)	127GW(2028)	212%
맥킨지 (McKinesy)	글로벌 AI 데이터센터	55GW(2023)	171~219GW(2030)	311~398%
골드만삭스 (Goldman Sachs)	글로벌 AI 데이터센터 (암호화폐 제외)	400TWh(2023)	1,040TWh(2030)	260%
전략국제 문제연구소 (CSIS)	미국 AI 데이터센터	4GW(2024)	84GW(2030)	2,100%

자료: 한국과학기술기획평가원(KISTEP,2025) AI로 인한 전력 수요의 폭발적 증가와 대응방안

〔그림 2-8〕 지역별 세계 데이터센터 전력 소비량 전망



자료: IEA(2-25.04), Energy and AI, p65

3. 통신 인프라의 에너지 소비 및 탄소 배출 현황

가. 통신 인프라의 역할과 전력 소비 구조

통신 인프라는 AI, 클라우드, IoT, 메타버스 등 모든 디지털 서비스의 데이터 전송을 가능하게 하는 핵심 기반이며, 유·무선 네트워크 장비, 전송망, 기지국, 코어망 등으로 구성된다. ITU에 따르면 2024년 인터넷 사용자 수는 약 55억 명으로 세계 인구의 68%를 넘어섰으며[14], 2027년까지 셀룰러·고정형·와이파이 전 분야에서 네트워크 데이터 소비량이 지속 증가할 것으로 전망된다. 특히 와이파이가 전체 전송량에서 가장 큰 비중을 차지할 것으로 예상된다〔그림 2-9〕. 통신 산업의 주요 전력 소비원은 RAN 안테나, 라디오 유닛, 기지국 장비이며, 여기에 코어·엣지 컴퓨팅, 냉각 시스템, 전력 증폭기, 광전송장비(OTN)까지 포함된다. 이러한 구성 요소들은 전력과 냉각 부하를 동시에 증가시켜, 통신 인프라의 Scope 2(구매 전력) 배출뿐 아니라 Scope 3(장비 제조 및 폐기) 배출에도 기여한다.

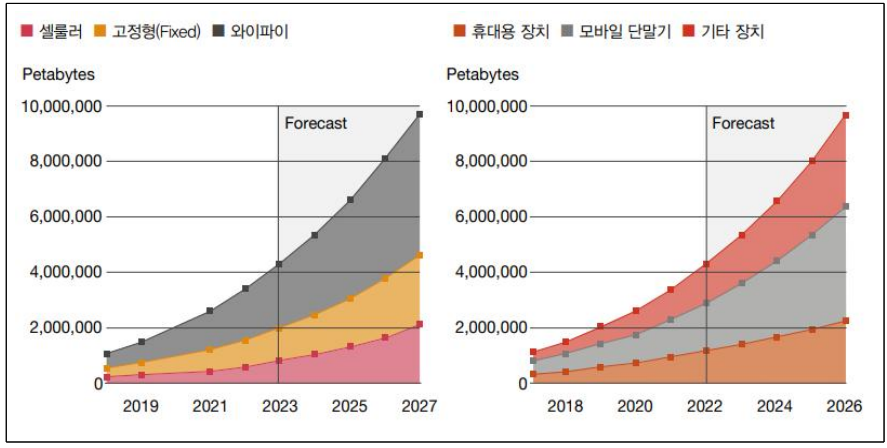
나. 유/무선 통신망의 탄소 배출

유선 통신망은 인터넷과 네트워크 연결의 근간이며, 사용되는 전송 매체의 종류에 따라 에너지 효율과 탄소 배출량이 큰 차이를 보인다. 초기 유선 통신망에서 널리 사용된 동축

14) ITU(2024), Facts and Figures 2024

케이블은 설치 및 운영 과정에서 14kg CO2/km를 배출하는 반면, 광 케이블은 2.3kg CO2/km로 훨씬 낮은 배출량을 기록한다. 또한, 전송 속도 50Mbps를 기준으로 할 때, 광케이블은 연간 약 1.7톤의 이산화탄소를 배출하지만 동축 케이블은 약 2.7톤의 이산화탄소를 배출한다.[15] 광케이블은 제조 시 에너지 집약도가 낮고 내구성이 높아 유지보수 주기가 길다는 점에서 환경적 이점을 가지며, 노후 동축망을 대체하는 FTTH[16] 전환이 주요 통신사업자의 탄소 저감 전략으로 자리잡고 있다. 그러나 전송 거리가 길어질수록 라우터와 스위치 등 중계 장비의 전력 사용이 늘어나 전체 네트워크 소비 전력의 상당 부분을 차지하므로, 향후 네트워크 인프라 효율화와 재생에너지 기반 운영이 병행되어야 한다.

〔그림 2－9〕 2018-2027 네트워크별 전 세계 총 데이터 소비량



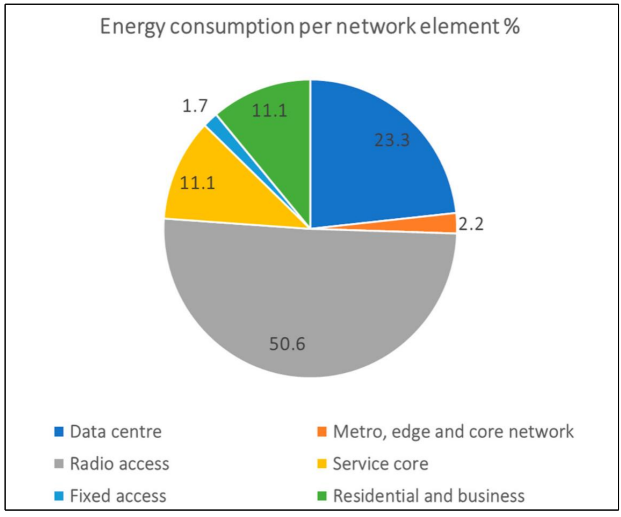
자료: PwC's Global Telecom Outlook 2023-2027, Omdia

무선 통신망은 AI 서비스 및 모바일 데이터 전송의 핵심이며, 전체 통신망 탄소 배출에서 큰 비중을 차지한다〔그림 2－10〕. 2021년 기준 전 세계 탄소 배출량 중 약 1.6%가 무선 통신과 관련되어 발생했으며, 이는 약 6억 톤의 이산화탄소에 해당한다[17]. 국제에너지기구

15) Hutchinson, How optical fibre will help create a sustainable future for data centres
16) Fiber To The Home의 줄임말로 가입자 개별 가정까지 광섬유를 직접 연결하는 초고속 유선 통신망 구조를 의미
17) Boston Consulting Group(2021), Putting Sustainability at the Top of the Telco Agenda

(IEA)는 세계 무선 통신망의 소비 전력이 2015년 최대 253TWh에서 2021년 최대 370TWh로 약 46% 증가했다고 분석했다. 또한 유·무선 인터넷 통신망과 여기에 연결될 IoT 기기까지 합친 에너지 소비량은 2026년 세계 전체 에너지 소비량의 약 6%에 이를 것으로 전망된다.

〔그림 2－10〕 네트워크 요소별 에너지 소비



자료: Energy consumption breakdown by network element, (Chochliouros, et al.(Energies 21)) p.5

4. ICT 단말 및 주요 부품 제조 분야의 에너지 소비와 탄소 배출 현황

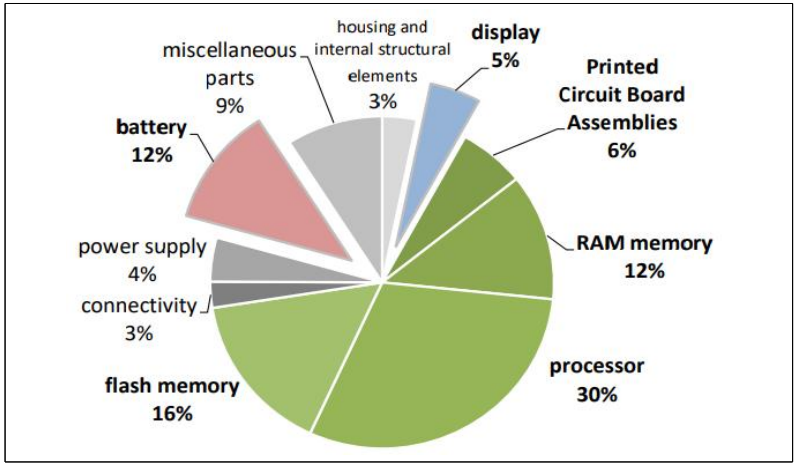
가. 스마트폰 제조 및 보급으로 인한 탄소 배출 현황

Greenpeace는 스마트폰 온실가스 배출량 중 제조 및 운송 과정이 전체 배출량의 83%를 차지한다고 분석했다. 스마트폰 제조 과정에 대한 전 생애 주기 분석(Life Cycle Analysis, LCA) 결과, 프로세서, flash 메모리, RAM 메모리, 배터리 순으로 많은 온실 가스를 배출한다〔그림 2－11〕. 스마트폰 1대 제조시 평균 85kg의 이산화탄소에 달하는 온실가스가 배출되며[18], 탄소 배출량은 2010년 4%에서 2020년 11%로 크게 증가하였다. ICT 분야 시장 조사

18) Le Monde(2023), Smartphones' carbon footprints are largely underestimated.

전문 기관인 IDC에 따르며 2025년 글로벌 스마트폰 출하량을 12억 4,000만 대로 전망하며 스마트폰으로 인한 온실가스 발생량은 꾸준히 증가할 것으로 추산된다. 스마트폰 제조와 운송 단계에서의 높은 배출 비중은 ICT 산업 전반의 공급망 탄소 관리와 지속가능한 생산 체계 구축의 중요성을 보여준다. 향후 산업 성장에 따라 배출 규모가 확대될 가능성이 있는 만큼, 저탄소 전환을 촉진하는 거시적 정책 방향과 산업별 이행 체계 마련이 요구된다.

〔그림 2－11〕 스마트폰 부품별 온실가스 배출 정도



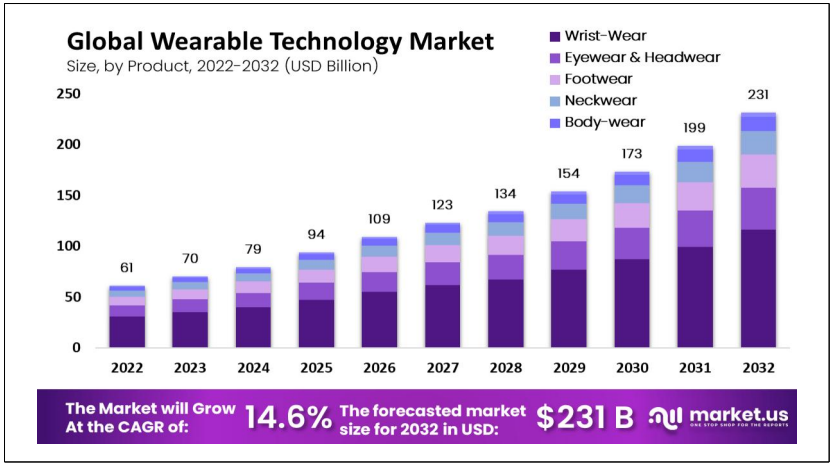
자료: an analysis of the design evolution and environmental impact of smartphones

나. 웨어러블 디바이스 제조 및 보급으로 인한 탄소 배출 현황

IoT, 인공지능 등의 발전으로 웨어러블 기기 출하량의 지속적인 성장이 예상되며, 스마트 워치는 웨어러블 디바이스 중 가장 큰 비중을 차지하고 있으며, 2023년 1억 5천만 대 이상의 스마트 워치가 출하되었다〔그림 2－12〕. 스마트워치 1기 제조 시 36.7kg의 이산화탄소에 해당하는 온실가스가 발생하며, 연간 약 550만 톤의 이산화탄소가 발생하는 것으로 추정된다.[19)]

19) Apple(2023), Product Environmental Report-Watch Series 9

〔그림 2－12〕 디바이스 유형별 글로벌 웨어러블 기술 시장 규모 전망

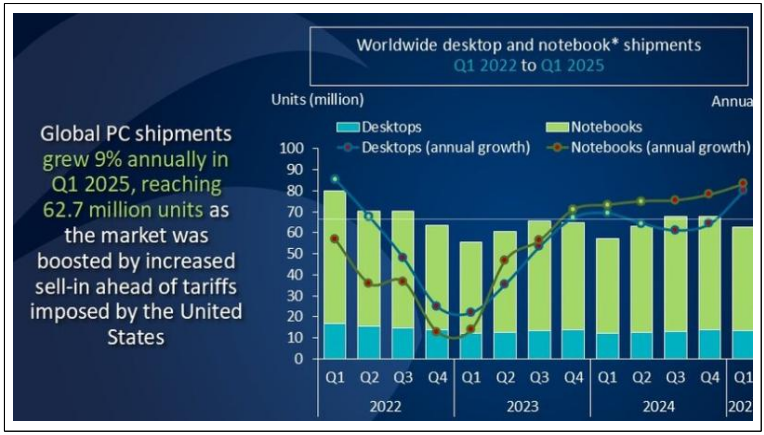


자료: Electro IQ(2025), Wearable Technology Statistics By Market, Consumers And Region

다. PC 제조 및 보급으로 인한 탄소 배출 현황

Canalys의 보고서에 따르면, 2025년 1분기 데스크톱, 노트북, 워크 스테이션의 총 출하량은 6,270만 대를 기록하였으며, 전체 PC 출하량 중 노트북 출하량은 4,940만 대로 전년 대비 10% 증가하였고, 데스크톱 출하량은 8% 증가한 1,330만대를 기록하였다〔그림 2－13〕. 데스크톱 1대로 인한 탄소 배출량은 약 422kg에 이르며, 노트북 1대로 인해 발생하는 온실가스는 약 305kg에 달한다.

〔그림 2－13〕 2022년 1분기-2025년 1분기 전 세계 PC 출하량



자료: Canalys(2025)

라. AI 관련 칩 제조로 인한 에너지 소비 및 탄소 배출 현황

2025년 4월에 Greenpeace는 보고서를 통해 전 세계 AI 시장이 성장함에 따라 2030년이 되면, 전 세계 AI 칩 제조를 위한 전력 수요는 최대 3만 7,238GWh에 달할 것으로 추정한다. 이는 2023년 대비 약 170배 증가한 수치이며, 현재 아일랜드의 총 전력 소비량보다 많은 양이다.[20] 특히 동아시아는 상업용 AI 모델의 핵심 하드웨어 제품을 주도적으로 생산해 왔으며, 이에 따른 전력 수요 증가는 동아시아, 특히 한국과 대만의 전력망에 막대한 부담을 주고 있다. AI 칩 제조로 인한 전 세계 전력 소비량은 2023년 218GWh에서 2024년 984GWh로 늘어났다. 증가율로 따지면 350%가 넘는다〔그림 2－14〕. 2024년 대만의 AI칩 제조에 사용된 전력망은 375.8GWh에 달했다. 이는 전년 대비 350.6% 증가한 양으로, 대만에서 약 9만 3,000가구가 소비하는 전력량과 맞먹는다. 대만 정부는 반도체 산업의 전력 수요 급증으로 인해 2030년까지 전력 사용량이 매년 12~13% 늘어날 것으로 예측하고 있다. 한국에서는 AI 칩 제조로 인한 전력 소비량이 2023년 134.6GWh에서 2024sus 315.2GWh로 두 배 이상 증가했다. 2050년이 되면, 반도체 생산 시설이 밀집한 용인이 수도권 전체 전력 수요의 25%인 10GWh 이상의 전력을 소비할 것으로 예상된다. 이는 국가

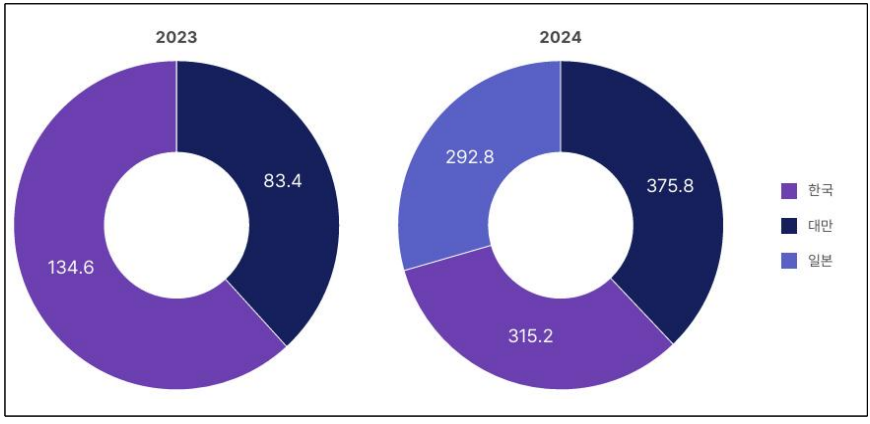
20) Greenpeace(2025), 인공지능(AI) 시대의 그림자: 반도체 제조산업의 전력 소비량과 온실가스 배출량 분석

전력망에 커다란 부담으로 작용할 것이다.

AI 칩 제조와 관련된 전력 소비로 인한 전 세계 배출량은 이산화탄소 환산량(CO2e)으로 2023년 9만 9,200톤에서 2024년 45만 3,600톤으로 4배 이상 증가했다〔그림 2－15〕. 이러한 배출량은 대부분 동아시아 전력망의 화석 연료에 대한 의존도가 높았기 때문이다. 따라서 ICT 단말 및 주요 부품 제조 부문의 탄소 저감을 위해서는 탈탄소화 및 에너지 효율 기술 도입이 시급하다.

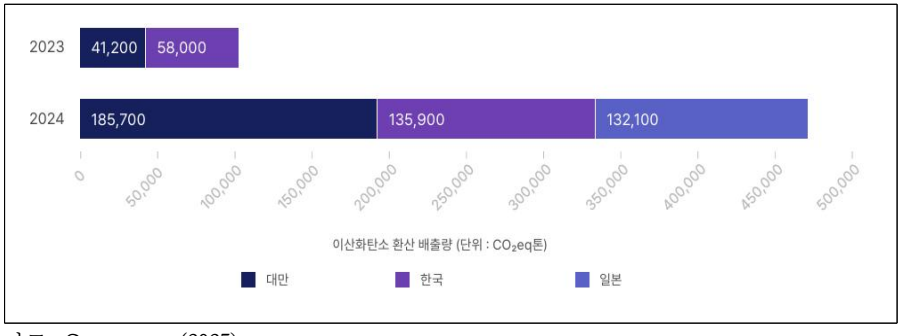
〔그림 2－14〕 지역별 AI 관련 칩 제조에 소요된 전력 소비량 추정치

(단위: GWh)



자료: Greenpeace(2025)

〔그림 2－15〕 지역별 AI 관련 칩 제조 관련 탄소 배출량 추정치



자료: Greenpeace(2025)

5. 기타 디지털 기술의 에너지 소비와 탄소 배출 현황

가. 블록체인 기술로 인한 에너지 소비와 탄소 배출 현황

1) 블록체인의 역할과 전력소비 구조

블록체인은 데이터를 분산된 여러 컴퓨터에 복제하여 저장하는 분산형 디지털 장부이다. 데이터가 블록 단위로 체인처럼 연결되어, 한 번 기록된 데이터는 사실상 위·변조가 불가능하다. 이러한 보안성과 투명성 덕분에 블록체인은 암호화폐를 넘어 금융, 물류, 저작권 관리 등 신뢰성이 중요한 다양한 분야의 핵심 기술로 주목받고 있다. 실제로 글로벌 블록체인 기술 시장은 빠르게 성장하여, Grand View Research에 따르면 2030년에는 14,315억 달러 규모에 이를 것으로 전망된다〔그림 2－16〕.[21] 하지만 블록체인의 보안성을 유지하는 방식에 따라 막대한 에너지를 소비하는 문제가 발생한다. 특히 비트코인 등이 사용하는 작업증명(Proof-of-Work, PoW) 합의 방식은 새로운 거래를 검증하기 위해, 전 세계 수많은 컴퓨터가 막대한 연산력으로 복잡한 수학 문제를 푸는 경쟁, 즉 채굴(Mining)을 수행한다. 이 에너지 집약적인 활동은 블록체인 기술의 지속 가능성에 대한 심각한 우려를 낳았다.

2) 블록체인 기술의 에너지 소비 및 탄소 배출 현황

UN University(UNU)의 연구 결과에 따르면, PoW방식의 비트코인 채굴 네트워크는 2020-21년 기간 동안 상위 10개국에서만 162.79TWh의 전력을 소비했다〔그림 2－17〕.[22] 이는 파키스탄의 연간 전력 소비량을 넘어서는 수치이다. 당시 비트코인 채굴 에너지의 45%가 석탄 발전을 통해 공급되는 등 화석 연료 의존도가 높았다〔그림 2－18〕. 또한 이로 인해 발생한 상위 10개국의 이산화탄소 배출량은 83.12Mt에 달했다〔그림 2－17〕. 이처럼 막대한 에너지 소비와 탄소 배출 문제는 블록체인 기술의 지속 가능성에 대한 근본적인 과제로 지적되었다.

---

21) Grand View Research(2025), Blockchain Technology Market(2025 - 2030) https://www.grandviewresearch.com/industry-analysis/blockchain-technology-market

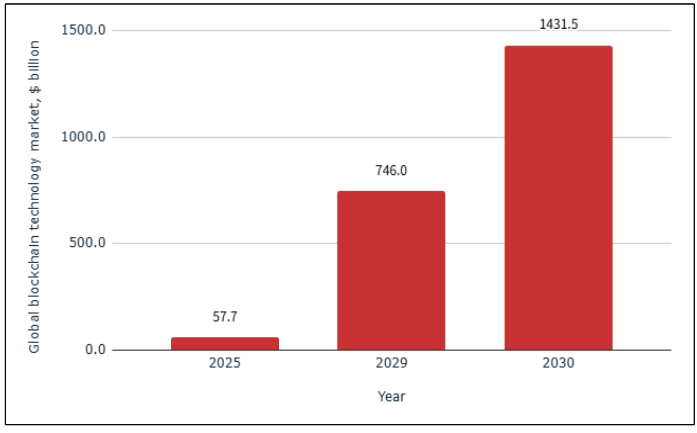
22) CHAMANARA et al., The environmental footprint of bitcoin mining across the globe: Call for urgent action. Earth's Future, 2023

3) 블록체인 기술의 에너지 문제 해결을 위한 기술적 대안

블록체인 업계는 에너지 문제를 해결하기 위해 지분증명(Proof-of-Stake, PoS) 방식을 핵심 대안으로 도입하고 있다. PoS는 막대한 컴퓨팅 경쟁 대신, 해당 암호화폐를 많이 보유한 검증인에게 블록을 생성할 기회를 부여하는 방식이다.

대표적인 사례로, 이더리움은 2022년 9월 머지(The Merge) 업데이트를[23] 통해 기존의 PoW에서 PoS로 성공적으로 전환했다. 이 전환을 통해 이더리움 네트워크의 에너지 소비량은 99.95% 이상 획기적으로 감소된다〔그림 2-19〕.[24] 이는 블록체인 기술이 환경 문제를 해결하고 지속 가능성을 확보하기 위해 어떻게 진화하고 있는지를 보여주는 중요한 전환점이 되었다.

〔그림 2-16〕 글로벌 블록체인 기술 시장 규모



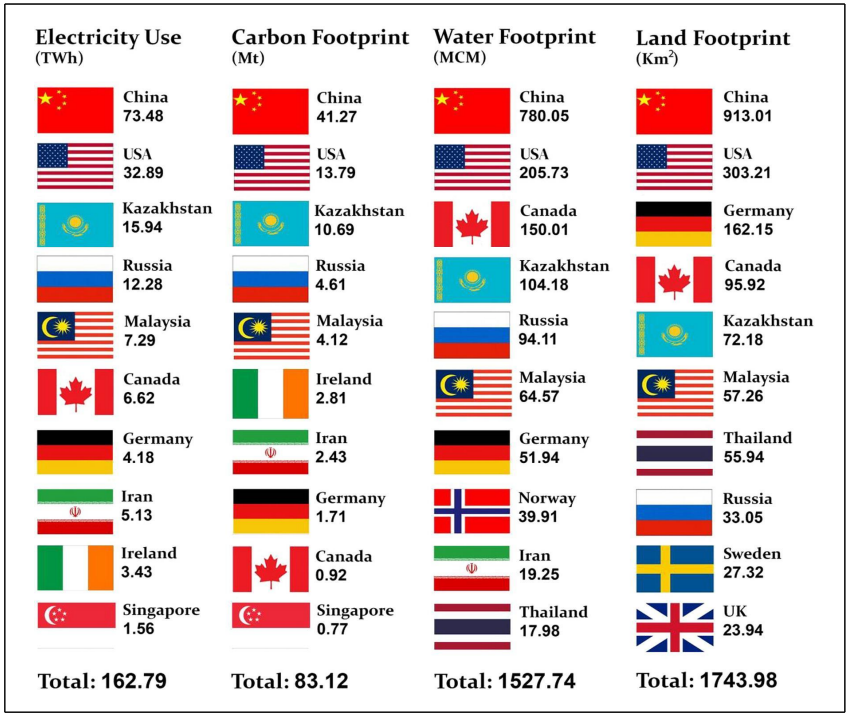
자료: Grand View Research(2025)

23) 머지 업데이트: 이더리움 블록체인의 엔진을 교체한 역사적인 대규모 업그레이드를 뜻한다. 2022년 9월 15일에 완료되었으며, 이더리움의 운영 방식을 완전히 바꾸었다.

24) CoinLaw(2025), Bitcoin Energy Consumption Statistics 2025: Efficiency, Regulation & Green Tech
https://coinlaw.io/bitcoin-energy-consumption-statistics/

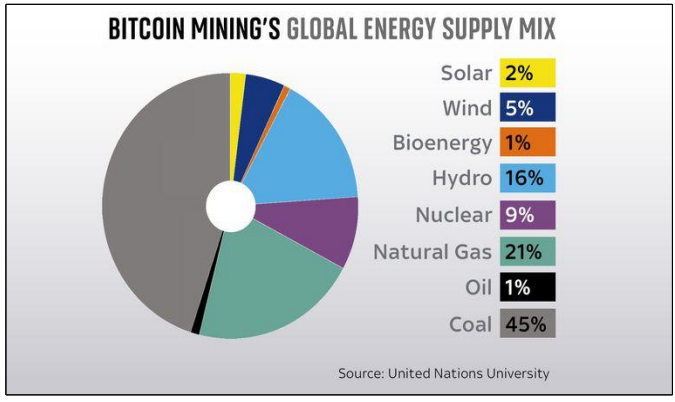
〔그림 2－17〕 2020-21년 비트코인 채굴 작업에 사용된 국가별 전력 소비량

(단위: TWh)



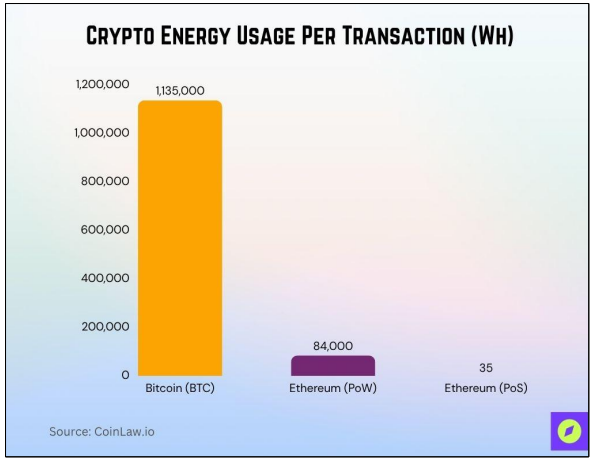
자료: CHAMANARA et al., The environmental footprint of bitcoin mining across the globe: Call for urgent action. Earth's Future, 2023

〔그림 2－18〕 2020-21년 비트코인 채굴 작업에 공급된 에너지원



자료: sky news(2023)

〔그림 2－19〕 암호화폐 거래 종류별 에너지 사용량(Wh)



자료: CoinLaw(2025)

나. 사물인터넷(IoT) 기술로 인한 에너지 소비와 탄소 배출 현황과 과제

사물인터넷(IoT) 기술은 우리 삶의 편의성을 높이고 산업 효율성을 증대시키는 핵심 동력으로 자리 잡았지만, 그 이면에서는 에너지 소비와 탄소 배출이라는 환경적 문제를 심화시키고 있습니다. 시장조사기관 IoT Analytics의 보고서에 따르면, 2023년 전 세계 활성 IoT 연결 장치 수는 166억 개에 달했으며, 2025년에는 215억 개로 급증할 것으로 전망됩니다〔그림 2－20〕. 이러한 IoT 기기의 폭발적인 증가는 단순히 개별 기기의 전력 소비를 넘어, 이를 뒷받침하는 데이터 전송 네트워크와 데이터센터의 에너지 부담을 직접적으로 가중시키고 있습니다.

개별 IoT 기기의 전력 소모는 미미할 수 있으나, 수백억 개의 장치가 24시간 내내 데이터를 생성하고 전송하는 과정에서 막대한 에너지 소비가 발생합니다. 특히, IoT 기기 제조 단계에서부터 상당한 에너지와 자원이 소모되며, 제품의 수명 주기가 짧아지면서 발생하는 전자 폐기물(e-waste) 문제 또한 심각한 환경 부담으로 작용합니다.

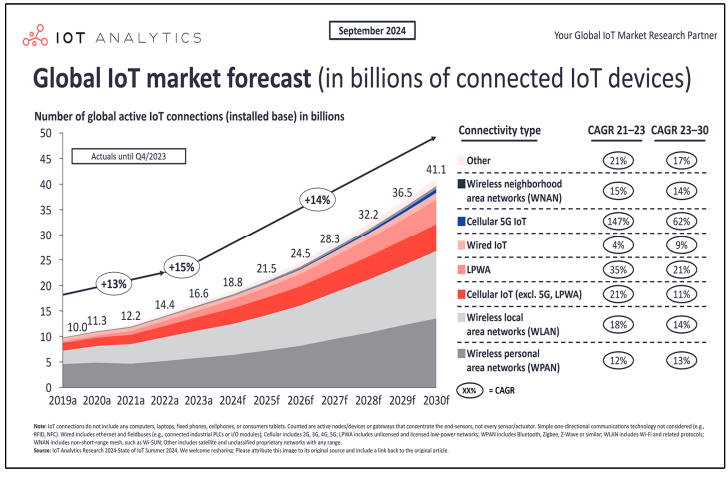
더 큰 문제는 IoT 생태계를 유지하기 위한 핵심 인프라의 에너지 소비입니다. 수많은 기기가 생성하는 방대한 데이터 트래픽을 처리하기 위해서는 5G와 같은 고성능 통신망이 필수적이며, 이 네트워크를 운영하는 데 막대한 전력이 소모됩니다. 또한, 이렇게 수집된 데이터는 에너지 집약적인 데이터센터로 전송되어 저장, 분석, 처리됩니다. 또한, 데이터센

터는 서버는 발생하는 열을 식히기 위한 냉각 시스템에 엄청난 에너지를 사용합니다.

국제에너지기구(IEA)의 분석에 따르면, 디지털화를 뒷받침하는 데이터센터, 데이터 전송 네트워크, 그리고 연결된 장치를 포함한 전체 디지털 생태계는 전 세계 에너지 관련 온실가스 배출량의 약 2%에 해당하는 약 700메가톤(Mt)의 이산화탄소 환산량(COe)을 배출했습니다.[25] 이는 IoT 기술이 단순히 편리함을 제공하는 것을 넘어, 지구 온난화에 미치는 영향이 결코 작지 않음을 보여줍니다.

결론적으로, IoT 기술의 혜택을 지속 가능하게 누리기 위해서는 기술 발전과 더불어 환경적 영향을 최소화하기 위한 노력이 시급합니다. 저전력 IoT 기기 설계, 에너지 효율적인 통신 기술 개발, 그리고 신재생 에너지를 사용하는 친환경 데이터센터 구축 등 기술적, 정책적 차원의 다각적인 접근이 필요합니다.

〔그림 2－20〕 글로벌 IoT 시장 예측



자료: IOT ANALYTICS(2024)

25) ZCA(2022), With the Internet of Things set to near 30 billion devices by 2025, tech giants are facing pressure to address the growing environmental impact their products pose
https://www.zerocarbonacademy.com/posts/with-the-internet-of-things-set-to-near-30-billion-devices-by-2025-tech-giants-are-facing-pressure-to-address-the-growing-environmental-impact-their-products-pose

제 2 절 국내외 메가트렌드 및 동향 분석

디지털 전환의 확산 속에서 탄소감축의 방향성을 확인하고자, 본 절에서는 기술적·산업적·정책·거버넌스 축을 따라 국제 및 국내 동향을 체계적으로 점검한다. 우선 OECD와 EU의 계량·정책 프레임을 정리하고, 이후 국내 동향을 동일 좌표계에서 재배치함으로써 글로벌 기준과의 정합성 및 정책적 시사점을 도출한다.

1. 국제 표준화에 따른 디지털 경제 계량화와 탄소 영향 분석

〔그림 2-21〕 디지털 경제의 계층적 정의

<table>
<tr><td colspan="2" rowspan="2">Included in GDP production boundary</td><td colspan="3">Outputs</td></tr>
<tr><td>Digital</td><td>Non-Digital</td><td>Society</td></tr>
<tr><td rowspan="3">Digital inputs as a factor of production</td><td>High</td><td>Core measure:
Economic activity from producers of digital content, ICT goods and services</td><td>Narrow measure:
Economic activity from producers reliant on digital inputs</td><td rowspan="2">Digital society</td></tr>
<tr><td>Medium</td><td>Core measure:
Economic activity from producers of digital content, ICT goods and services</td><td>Broad measure:
Economic activity from producers significantly enhanced by digital inputs</td></tr>
<tr><td>Low / none</td><td>Core measure:
Economic activity from producers of digital content, ICT goods and services</td><td>Traditional economy</td><td>Traditional society</td></tr>
</table>

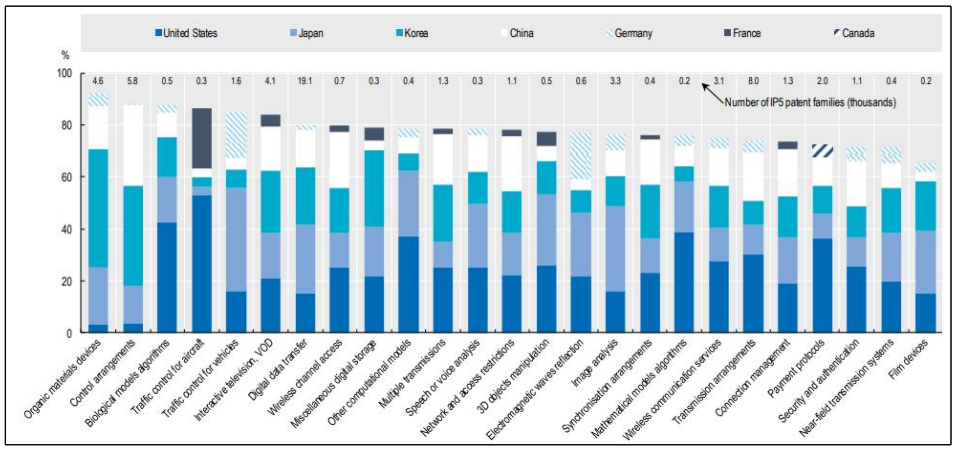
자료: OECD(2020d). p.40.

디지털 전환이 경제 전반으로 파급되면서, "무엇을 디지털 경제로 볼 것인가"에 대한 합의와 계량 체계의 정교화가 선결과제로 부상했다. OECD는 이 불확실성을 줄이기 위해 2020년 '디지털 경제 측정을 위한 공통 프레임워크 로드맵'을 제안하여, 디지털 경제를 핵심 및 협의 범위로 구분·층위화하고 각 범주가 국내총생산에서 차지하는 부가가치 비중을 국제적으로 비교 가능한 형태로 제시하였다(OECD, 2020d). 해당 정의는 〔그림 2-21〕에 제시되어 있다. OECD의 공통 프레임워크는 국가별로 상이했던 디지털 경제 비중 추정

의 정밀도를 높였으며, 핵심 및 협의 범위를 구분해 각 범주가 국내총생산에서 차지하는 부가가치 비중을 국제적으로 비교할 수 있도록 하였다. 나아가 이러한 국제 비교 결과는 디지털 부문의 확대가 에너지 소비와 온실가스 배출 구조에도 중요한 영향을 미칠 수 있음을 시사하며, 이를 정밀하게 평가하기 위한 기술적·제도적 계정체계의 정비가 중요함을 보여준다. 동시에, 이러한 계정체계는 향후 탄소 영향 진단과 저탄소 전환 전략을 수립하는 데 기초자료로 기능할 수 있다.

이를 뒷받침하기 위해 OECD는 '디지털 공급 및 사용표(Digital Supply Use Tables)'를 도입하는 방안을 추진하고 있다. 이는 디지털 전송 여부 등 거래 특성을 포착하여 소프트웨어, 데이터베이스와 같은 디지털 산출물의 흐름을 체계적으로 계정화하려는 시도다. 이러한 데이터 체계는 디지털 경제의 활동 범위를 정밀하게 파악하는 기반이 되지만, 현재는 미국·캐나다 등 일부 국가를 제외하면 기초자료 부족으로 실제 작성에 제약이 존재한다. 이러한 한계 인식은 향후 디지털 기술의 탄소 영향 진단에도 직결된다. AI 학습·추론, 데이터센터, 블록체인과 같은 세부 기술 활동을 디지털 공급·사용표의 활동 분류에 연계하고, 전력 사용량과 배출계수를 결합하는 방식으로 정합적인 진단 틀을 마련해야 한다. 이는 AI·데이터센터·블록체인의 배출 현황을 정량적으로 분석하고, 분야별 효율 개선 목표를 설정하기 위한 국제 비교 가능 통계 기반의 출발점이 될 수 있다.

〔그림 2-22〕 디지털 경제 부문별 비중(국가별)

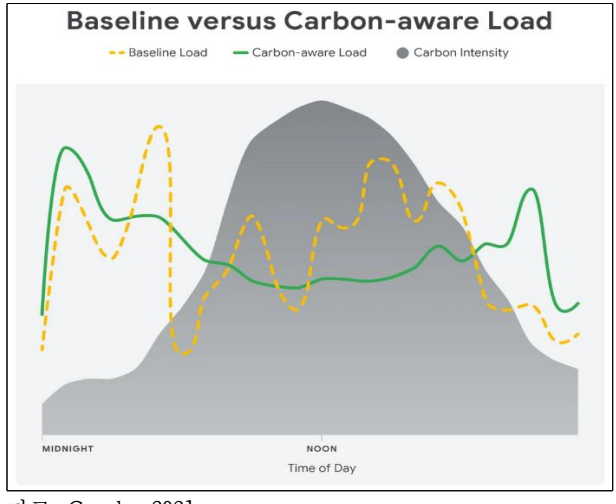


자료: OECD, STI Micro-data Lab: Intellectual Property Database

나아가 OECD 프레임은 단순한 규모 측정을 넘어, 저탄소 전환을 위한 정책 비교와 우선순위 설정에도 활용될 수 있다. 동일한 계정체계에서 디지털 전송 여부와 부가가치 비중이 파악되면, 대규모 클라우드·엣지 인프라와 같이 에너지 소비가 높은 활동의 상대적 위치를 식별하고, 에너지 효율화나 재생전력 조달(PPA) 확대와 같은 정책 수단이 가장 큰 효과를 발휘할 수 있는 부문을 도출할 수 있다. 이러한 분석은 데이터센터의 전력 사용 효율(PUE) 개선, 폐열 회수, 재생전력 조달 확대 등 본 과제의 인프라 전환 로드맵을 국제 동향과 정합적으로 설계하는 근거가 된다. 아울러 OECD 표준화 논의 참여와 RE100 등 글로벌 추진체계와의 연계는 진단과 정책 실행 간의 선순환을 강화한다. 특히, 국제 지표와 호환되는 데이터 체계를 내부화할수록 해외 감축 정책 및 협력 동향을 정량적으로 분석하고 이를 국내 제도 개선에 신속히 반영할 수 있다. 특히, 한국은 OECD 주요국 대비 디지털 경제 비중이 상대적으로 높은 수준을 보이고 있다〔그림 2-22〕. 이는 국내 디지털 부문이 경제 전반에서 차지하는 구조적 비중과 잠재력이 크다는 것을 의미하며, 그만큼 해당 부문의 에너지 소비와 온실가스 배출 특성을 정밀하게 평가할 필요성을 시사한다.

2. OECD 권역 기술 트렌드

〔그림 2-23〕 탄소 지능형 컴퓨팅



자료: Google, 2021

OECD 권역에서 주목할 변화는 '측정-감축-최적화'의 기술 통합이다. 배출량 측정의 표준화 범위가 소프트웨어·클라우드 계층까지 확대되었으며, 2024년 ISO/IEC 21031로 제정된 그린소프트웨어재단(GSF)의 '소프트웨어 탄소강도(SCI)' 규격은 애플리케이션 단 배출을 기능 단위로 관리하는 공통 기준을 제시한다. 이 표준은 위치별 전력계통 탄소강도와 하드웨어 내재배출을 포함해 전 단계의 의사결정을 지원한다. 아울러, 탄소강도가 낮은 시간대·지역으로 워크로드를 분산시키는 SDK와 API가 클라우드 환경과 CI/CD 파이프라인에 연계되면서 탄소 대응형(Carbon Aware) 최적화 사례가 확산되고 있다.

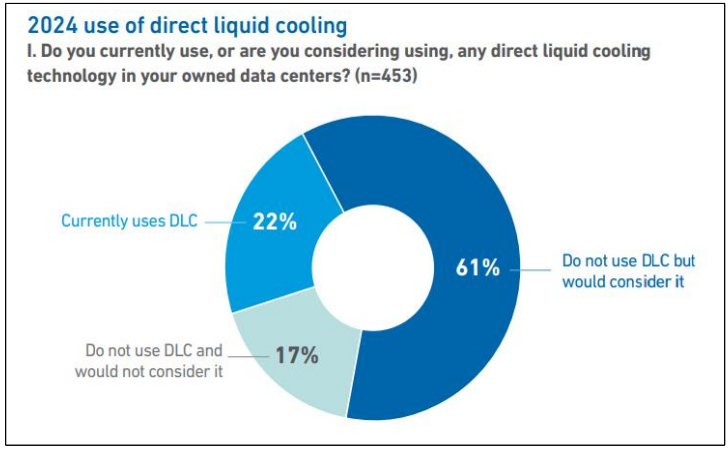
둘째, 데이터센터와 클라우드 인프라 분야에서는 탄소 대응형 운용과 액체냉각과 같은 하드웨어 혁신이 병행되고 있다. 구글이 2020년부터 운영 중인 탄소 대응형 컴퓨팅은 유연한 작업을 저탄소 시간대에 배치하고〔그림 2-23〕, 전력망 수요와 연계해 피크 전력 소비를 낮추는 방식으로 활용된다. 이를 통해 동일한 작업이라도 실행 시점과 위치 조정만으로 배출을 실질적으로 줄일 수 있음을 산업 차원에서 입증하였다. 한편, 고밀도 AI 워크로드를 처리하는 데이터센터의 냉각·전력 설계는 액체냉각 중심으로 전환되는 추세다. 업타임 인스티튜트(Uptime Institute)의 2024년 조사에 따르면 응답 기관의 22%가 직접 액체냉각(DLC)을 이미 도입했으며〔그림 2-24〕, 절반가량은 일부 랙에만 적용 중이지만 확대 의사가 뚜렷하다〔그림 2-25〕. 북유럽의 LUMI 슈퍼컴퓨터처럼 100% 수력 전기를 사용하면서 폐열을 지역난방에 공급하는 사례는, 열 순환 개념을 통신·HPC 인프라로 확장한 대표적인 예로 꼽힌다.

셋째, 네트워크 계층의 효율화는 5G-Advanced 표준과 함께 본격화되고 있다. 3GPP 릴리즈 18에는 기지국·단말의 저전력 모드, 신호 최소화, AI/ML 기반 에너지 절감 기능 등 RAN 전반의 에너지 최적화를 위한 규격이 포함되었다. 실제 현장에서는 심야 시간대 대용량 안테나를 '딥 슬립' 모드로 전환해, 트래픽이 적은 구간에서 에너지 사용을 절감하는 시범 결과가 보고되고 있다(FCST, 2024). OECD는 이러한 기술이 데이터센터와 네트워크의 직접적인 탄소 발자국을 줄이는 동시에, 디지털 트윈·IoT·AI 기반 수요관리와 같은 다른 산업 부문의 효율화까지 촉진하는 이중 효과를 가진다는 전제하에 측정 및 정책 프레임을 강화하고 있다.

마지막으로, AI 자체의 에너지 집약도를 낮추기 위한 알고리즘·아키텍처 연구가 빠르게

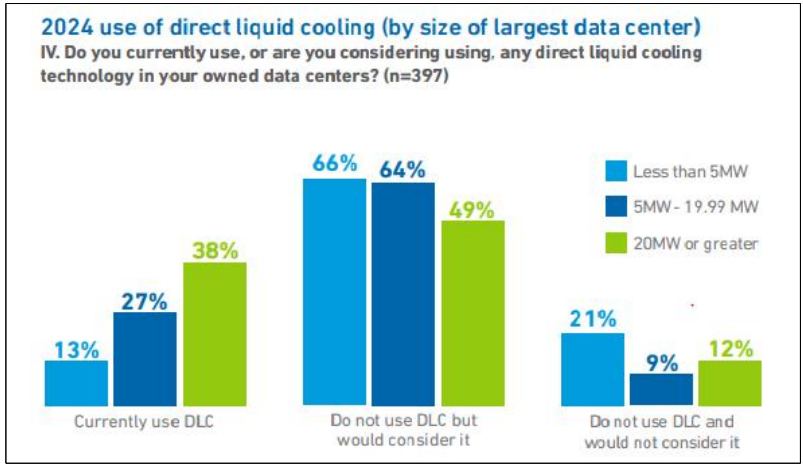
진전되고 있다. 조건부 연산을 활용하는 Mixture-of-Experts(MoE)는 파라미터 용량을 확대하면서도 활성 연산만 제한적으로 수행해 효율을 높이는 방식으로, 스위치 트랜스포머 등에서 계산량 절감 효과가 확인되었다. 추론 단계에서는 8·4비트 저정밀 양자화 및 모델 프루닝, 캐시 및 메모리 최적화와 같은 시스템 계층의 개선이 병행되어 전력과 지연을 동시에 줄이는 방향으로 발전하고 있다. 종합하면, OECD 권역의 기술 트렌드는 소프트웨어의 탄소 대응 기능 강화, 데이터센터와 인프라의 액체냉각·폐열 회수 확대, 네트워크의 표준 기반 절전 기술 도입, AI의 희소 연산 및 저정밀화와 같은 흐름이 상호 연계되며 하나의 통합 방향으로 전개되고 있다.

〔그림 2－24〕 액체 냉각(DLC) 사용률



자료: Uptime Institute(2024), p.1

〔그림 2－25〕 DLC 사용 현황과 향후 고려 여부



자료: Uptime Institute(2024), p.1

3. 국내 권역 기술 트렌드

국내 권역 기술 트렌드도 유사하게 측정-감축-최적화의 방향성을 따르고 있지만, 아직은 소프트웨어 계층보다는 데이터센터 등 물리적 인프라에 더 집중되어 있으며, 이제 막 소프트웨어 계층으로 확산되는 초기 단계에 있다.

측정 측면에서 정부가 행정·공공기관 정보자원 클라우드 전환·통합 추진계획에 따라 공공 부문의 IT 시스템을 민간 클라우드로 전환할 때, 서비스 기업의 탄소 배출량을 주요 평가 항목으로 포함하는 방안을 추진하고 있다. 이는 클라우드 서비스 자체의 탄소 성적표를 요구하기 시작했다는 의미로, 국내 클라우드 제공사(CSP)들이 자신들의 서비스가 얼마나 친환경적인지 계량하고 증명해야 하는 직접적인 동기가 된다.

감축 측면에서 네이버의 데이터센터 각 세종은 단순히 전력을 적게 소비하는 것을 넘어, 자연풍과 수자원을 이용한 냉각, 서버에서 발생하는 스마트팜이나 직원 온수 공급에 재활용하는 등 에너지 순환 개념을 도입했다. 네이버는 이를 통해 연간 1만 3천 MWh 수준의 전력을 절감해 6천 톤의 온실가스 배출을 줄일 수 있을 것으로 예상하고 있다.[26] 이외에도

26) 네이버(2024), 데이터센터 각 세종까지 LEED 플래티넘 획득… 친환경 분야도 글로벌 최고 수준 입증 https://www.navercorp.com/media/pressReleasesDetail?seq=31837

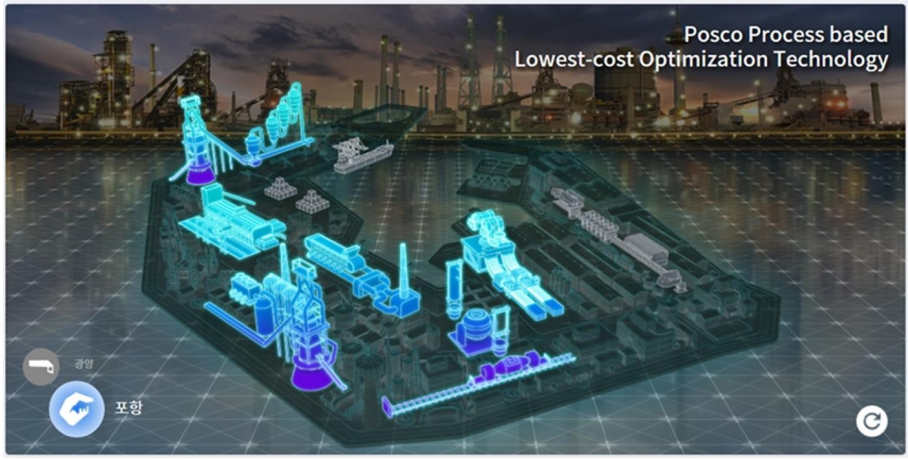
대부분의 국내 신규 데이터센터는 OECD 트렌드에서 언급된 액체 냉각의 도입을 검토 및 적용하고 있다.

최적화 측면에서 OECD의 탄소 대응형 SDK나 API처럼 개발자가 직접 활용할 수 있는 도구가 상용화된 사례는 드물지만, 그 기반이 되는 기술은 이미 적용되고 있다. 국내 클라우드 제공사들은 특정 시간대에 전력 사용량이 급증하지 않도록 AI를 이용해 워크로드를 분산하고, 외부 기온에 따라 냉각 시스템을 최적으로 자동 제어한다. 이는 사용자에게 API 형태로 제공되지는 않지만, 클라우드 내부적으로는 이미 탄소 대응형과 유사한 최적화가 이루어지고 있음을 의미한다. 또한 앞서 언급한 공공 클라우드 전환 사업에서 탄소 배출량 평가가 본격화되면, 국내 클라우드 기업들은 경쟁력을 확보하기 위해 탄소 대응형 API 상품 출시를 고려할 가능성이 높다.

이러한 AI 기반 최적화 기술은 IT 인프라를 넘어 국가 전력망과 주요 산업으로 확산되고 있다. 정부와 기업이 협력하는 AI 기반 수요반응(Auto DR) 국민 캠페인은 스마트 가전을 활용해 가정의 전력 소비를 최적화하며, 여러 곳에 흩어진 재생에너지를 AI로 통합 관리하는 가상발전소(VPP) 기술은 2024년 6월부터 시행된 분산에너지 활성화 특별법 시행과 함께 본격적인 확산이 기대된다. 또한, 포스코 등 주요 제조 기업들은 디지털 트윈 기술을[27] 통해 실제 공정과 동일한 가상 모델에서 에너지 효율을 극대화하는 방안을 찾아내고 있다〔그림 2-26〕. 또한 도시에서는 AI가 교통 흐름을 제어하는 차세대 지능형 교통 시스템(C-ITS)이 수송 부문의 탄소 배출을 줄이는 데 기여하고 있다.

27) 디지털 트윈 기술: 현실 세계의 물리적인 사물이나 시스템을 컴퓨터 속 가상 세계에 똑같이 복제하여 만드는 것을 의미한다. 단순한 3D 모델링과 다른 점은, 현실의 쌍둥이가 실제 사물에 부착된 센서 등을 통해 수집되는 데이터와 실시간으로 연동된다는 것이다.

〔그림 2－26〕 공정 프로세스 정보를 디지털(digital)로 구현한 디지털 트윈 포스플롯 (Digital Twin PosPLOT) 예시 이미지

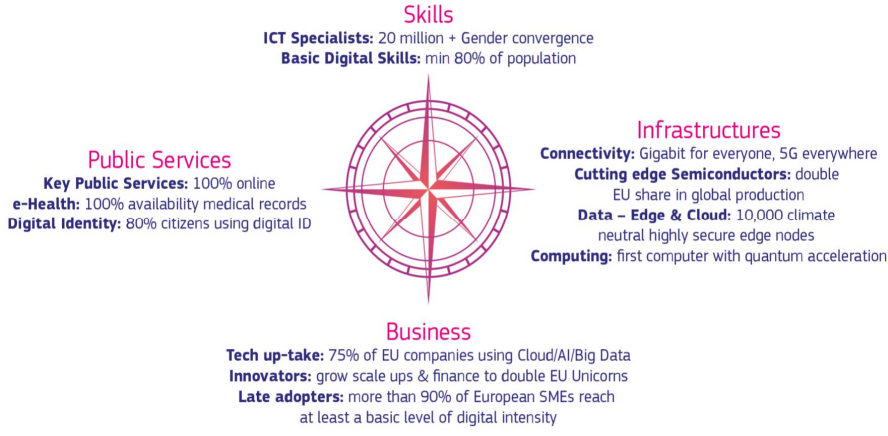


자료: 포스코그룹(2021)

4. 국제 저탄소 메가트렌드 및 동향: EU 및 미국

가. EU 디지털 및 경쟁력 나침반: 지속가능한 산업 전환 전략

〔그림 2－27〕 디지털 나침판(Digital compass) 2030



자료: European Union, 2020

EU는 유럽의 디지털 미래 구상에 이어, 팬데믹 이후 위험과 기회 요인을 재평가하여 〔그림 2-27〕과 같이 '2030 디지털 나침판(2030 Digital Compass)'을 발표하고, 인력·인프라·산업·공공의 네 축을 중심으로 2030년 달성 목표와 이행지표를 확정하였다. 이 전략의 특징은 디지털 경쟁력 강화와 지속가능성의 결합에 있다. 모든 가구의 초고속 기가급 연결망 구축과 인구 밀집지역 5G 커버리지 확대, 첨단 반도체의 역내 생산 비중 20% 달성, 2025년까지 양자 가속 컴퓨터 확보 및 2030년 최첨단 역량 확보 등 공급 측 투자를 명확히 하는 동시에, 1만 개의 기후 중립적이면서 고도 안전성을 갖춘 엣지 노드를 설치한다는 인프라 목표를 제시하여 네트워크와 컴퓨팅 자원의 에너지·탄소 성능을 전략의 핵심 요소로 설정하였다(European Union, 2020).

비즈니스 부문에서는 5G, IoT, 엣지, AI, 로보틱스 등 핵심 기술의 보급률을 2030년까지 질적으로 제고하고, 중소기업의 디지털화 수준 향상과 스케일업 생태계 조성을 목표로 한다. 공공 부문에서는 핵심 공공서비스의 100% 온라인 제공, 전자의무기록의 보편적 접근, 전체 시민의 80%가 디지털 ID를 활용하는 환경을 제시하여, 서비스 전달방식의 완전한 디지털화를 통한 효율성과 접근성 향상을 도모한다. 그러나 현재 수준과 비교할 때 5G 네트워크, 클라우드·빅데이터·AI 도입, 유니콘 기업 수, 반도체 생산 비중 등에서 여전히 상당한 격차가 존재하는 것으로 평가된다. 이러한 격차 인식은 곧 '에너지 효율적 디지털화'라는 실행 과제로 귀결된다. 즉, 커버리지 확대와 데이터 집약 가속이 불가피하다면, 반도체·엣지-클라우드 전 주기에 걸쳐 지속가능성을 전제한 설계가 우선되어야 한다.

이러한 디지털 나침반과 함께, EU 집행 위원회는 2025년 1월 29일 향후 5년간의 EU의 경쟁력 회복과 기업 규제 간소화 전략을 구체적으로 제시한 "경쟁력 나침반(Competitiveness Compass)"을 공식 발표했다. 이 계획은 EU의 지속가능산 성장 기반을 강화하기 위해 3대 필수 혁신 과제와 이를 실행하기 위한 5대 기반 조치를 중심으로 구성되어 있다. 그중에서도 핵심 축으로 제시된 과제 중 하나는 탈탄소화와 경제 성장의 동시 추진을 목표로 한 청정 산업 협약이다. EU는 탄소중립(Net Zero) 목표를 유지하면서도 산업 경쟁력을 저하시키지 않는 방식으로 친환경 전환을 추진하기 위해, 제조업 중심의 청정에너지 전환 및 에너지 효율성 증대를 위한 전략을 수립하였다. 이러한 전략은 에너지 비용 절감을 통한 청정에너지 공급망 구축과 산업 전환 지원, 저탄소 생산과 청정기술 도입을 유도하기 위한 유인체계 마련, 그리고 철강·금속·화학 등 에너지 집약 산업을 대상으로 한 맞춤형 탈탄

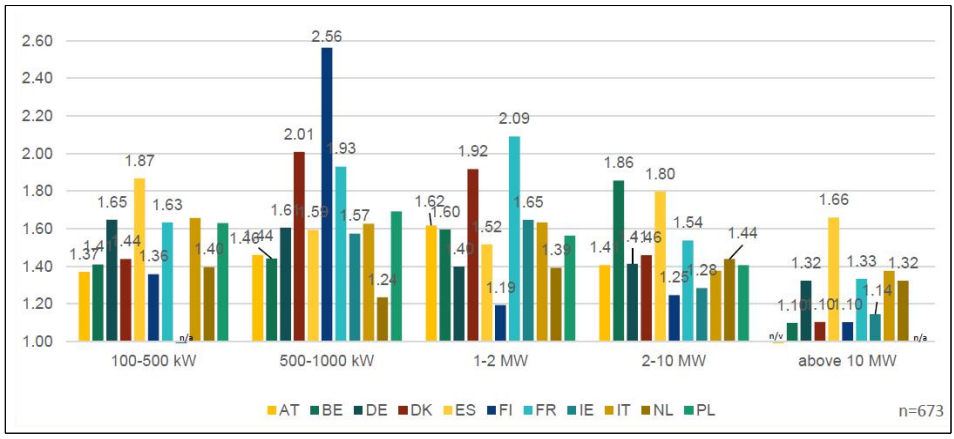
소 전략 추진을 포함한다. 또한 순환경제법 제정을 통해 재활용 역량을 확대하고 지속가능한 자원순환 투자를 촉진함으로써 산업 전반의 탄소 배출 저감을 도모하고 있다.

EU의 이러한 방향성은 데이터센터의 에너지 효율 향상, 냉각·폐열 회수, PUE 개선과 재생전력 조달, 클라우드 전환을 통한 자원 이용률 극대화와 같이 저탄소 인프라 전환을 핵심 축으로 삼고 있다. 여기에 저전력 반도체, AI 모델 경량화, 알고리즘 효율화와 같은 기술적 해법은 반도체 생산의 지속가능성과 AI 도입의 에너지 집약성 문제를 완화하는 수단이 된다. 이러한 접근은 산업과 공공 부문 전반에서 디지털화를 총 전력·총 배출 관점에서 최적화하고, 국제 탄소 감축 목표와 조화를 이루는 데 기여할 수 있다. 궁극적으로, 기술 혁신과 에너지·탄소 성과를 연계하는 전략이야말로 디지털 전환의 지속가능성을 담보하는 핵심 과제임을 시사한다.

나. 지속가능 디지털 인프라 전환과 정책 동향

EU는 규제와 민간 협력 프로그램을 결합해 기술 투자의 방향을 명확히 하고 있다. 2024년 개정된 에너지효율지침(EED) 은 IT 전력 수요 500 kW 이상 데이터센터에 대해 PUE, WUE, 에너지 재사용, 재생에너지 사용 등 핵심 성능지표를 〔그림 2－28〕과 같이 보고 및 공개하도록 의무화했으며, 이를 기반으로 유럽 데이터센터 데이터베이스가 가동되었다. 업계 주도의 기후중립 데이터센터 협약(CNDCP) 은 신규 센터에 대해 2025년까지 냉량 지역 PUE 1.3, 온난 지역 1.4, 기존 센터는 2030년까지 이 목표 달성을 요구하며, 2025년까지 전력 수요의 75%, 2030년까지 100%를 재생 또는 탄소 없는 전력으로 매칭할 것을 자율 목표로 제시했다(European Union, 2020). 이러한 규범들은 궁극적으로 기술 적용의 방향을 변화시키고 있다. 데이터센터에서는 고밀도 랙 구성을 전제로 한 직접 액체냉각(DLC)과 침지냉각의 도입이 확산되고 있으며, 자유냉각과 하이브리드 냉각 방식의 적용도 병행되고 있다. 특히 유럽의 HPC 인프라에서는 열회수를 도시 단위의 에너지 자산으로 통합하는 방식이 표준으로 자리잡아 가고 있으며, 이는 인프라 설계와 에너지 운영 전략 전반에 변화를 촉진하고 있다.

〔그림 2－28〕 데이터센터 크기에 따른 PUE 평균 수치



자료: Assessment of the energy performance and sustainability of data centres in EU

네트워크 분야에서는 5G-Advanced의 에너지 절감 기능을 토대로, 사업자와 장비 벤더가 AI 기반 절전 기술(딥 슬립, 지능형 셀 셰이핑 등)을 상용망에 적용하며 피크 외 시간대 전력 부하를 눈에 띄게 줄이고 있다. 이러한 기능은 표준에 기반하기 때문에 유럽 전역으로의 확산 가능성이 높다. 한편, 제품·부품 측면에서는 디지털 제품여권(DPP)을 포함한 지속가능제품규정(ESPR)이 2024년 발효되어, 서버·네트워크 장비·배터리 등 규제 대상에 대해 수명, 수리, 재사용, 소재, 탄소 정보를 디지털화하는 작업이 진행되고 있다. 이는 ICT 장비의 순환성과 수명 연장을 위한 데이터 인프라를 제공하며, 제조-조달-운영 각 단계에서 모듈 교체, 재제조, 부품 회수 등 기술 개선의 경제성을 높이는 기반이 되고 있다.

AI와 반도체 분야에서도 저전력·저자원 설계에 대한 투자 흐름이 뚜렷하게 나타나고 있다. 칩스 공동사업(Chips Joint Undertaking, Chips JU)과 중요프로젝트(Important Projects of Common European Interest, IPCEI)는 에너지 효율형 마이크로컨트롤러, 엣지 AI 프로세서, 질화갈륨(GaN)·탄화규소(SiC) 기반 전력반도체, 고집적 패키징 공정 등 저전력 반도체 설계를 중심으로 다수의 프로젝트를 추진 중이다. 이들은 AI 특화 하드웨어와 소프트웨어의 공동 최적화(Co-Design)를 통해 연산당 전력 소모를 줄이는 로드맵을 제시하고 있다. 또한 유럽그린디지털연합(EGDC)은 ICT 솔루션의 '순 탄소 영향(Net Carbon Impact)[28]' 평가 방법론을 공개하였다. 이 방법론은 데이터센터와 단말 장치의 직접 배출뿐 아니라 전

력 사용 등에서 발생하는 간접 배출까지 함께 추적하는 과학적 프레임을 제공한다.

EU의 기술 트렌드는 규제와 자율 협약을 결합하는 방향으로 진화하고 있다. 규제 측면에서는 측정·보고·성능지표 공개를 의무화하고, 자율 협약에서는 시간대별 무탄소 전력 매칭, PUE·WUE·열회수 목표를 설정한다. 이러한 정책 조합은 데이터센터의 액체냉각과 열회수, 재생전력 조달, 모바일망의 표준화 절전 기능과 AI 기반 운영, 제품·부품의 디지털 여권 기반 순환 설계, 저전력 칩과 엣지 AI 도입을 가속하고 있다. 결과적으로, 이러한 전환은 설계와 조달 단계부터 요구조건으로 내재화되며, 사업성 평가와 기술 선택의 기준을 동시에 변화시키고 있다.

5. 국내 동향: 한국의 저탄소 디지털 전환 현황과 과제

〈표 2－3〉 국가표준시행계획 추진사업 내역

4대 분야	12대 중점추진과제	예산
세계시장 선점을 위한 표준화	디지털기술 표준화, 국가유망기술 표준화, 저탄소기술 표준화	405억 원
기업 혁신을 지원하는 표준화	맞춤형 시험인증 서비스 확대, 국내외 기술규제 애로 해소, 新측정표준 개발보급	1517억 원
국민이 행복한 삶을 위한 표준화	생활밀착 서비스 표준화, 사회안전 서비스 표준화, 공공민간데이터 표준화	433억 원
혁신 주도형 표준화체계 확립	R&D-표준-특허 연계체계 확보, 개방형 국가표준체계 확립, 기업 중심 표준화 구축	389억 원

자료: 2023 산업통상자원부

OECD가 추진하는 디지털 경제 계량·탄소 영향 표준화 흐름에 발맞춰, 한국 역시 저탄소 디지털 전환을 위한 제도와 기술 기반을 확립해 가고 있다. 이러한 전환은 전력 조달 제도, 인프라 기술, 통신망 운영, AI·반도체의 네 축이 맞물리며 하나의 흐름으로 전개되고 있다. 전력 조달 측면에서는 한국형 RE100(K-RE100)이 정착되면서 기업이 녹색요금,

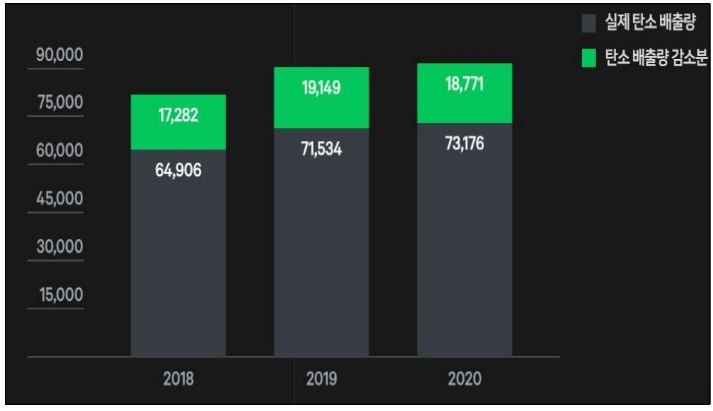
28) 순 탄소 영향(Net Carbon Impact): 특정 기술이나 솔루션이 발생시키는 온실가스 배출량에서, 이를 통해 절감되는 온실가스 감축량을 차감한 값. 기술의 전 과정에서의 환경 영향을 종합적으로 평가하는 지표이다.

REC 구매, 직접·제3자 PPA, 지분 참여, 자체 건설 등 다양한 수단을 조합해 재생전력을 확보하고, 사용 실적을 검증·인정받을 수 있는 체계가 마련되었다. 2022년 직접 PPA 제도 시행으로 조달 수단이 더욱 다양해졌으며, 이를 통해 데이터센터·통신망·클라우드 등 대규모 전력 수요를 '측정 가능한' 저탄소 전력으로 전환하는 기반이 강화되었다.

또한 분산형 전원과 수요지 인근 발전을 촉진하는 「분산에너지 활성화 특별법」이 2024년 6월 시행되면서, 대규모 수요시설의 입지와 계통 연계에 '지역 분산' 원칙이 적용되기 시작했다. 수도권 집중 완화와 전력계통영향평가 강화라는 정책 방향은, 고밀도 컴퓨팅 수요를 지역으로 분산시키고 신재생·ESS 연계형 데이터센터 모델의 확산을 촉진하는 설계 변수로 작용하고 있다. 이는 허용 가능한 총전력 한도 내에서 효율·재생·분산을 동시에 달성하도록 기술 선택에 구조적 압력을 가하며, OECD·EU가 추진하는 분산형·저탄소 인프라 전환 흐름과 궤를 같이한다.

고밀도 디지털 연산 수요는 데이터센터의 저탄소 설계와 직결되고 있다. 네이버의 하이퍼스케일 데이터센터 '각(GAK)'은 초기 춘천 센터에서 PUE 1.1~1.3 수준을 달성했으며, 이는 EU의 기후중립 데이터센터 협약(CNDCP)이 제시한 신규 센터 목표치(냉량 지역 1.3, 온난 지역 1.4)를 상회하는 수준이다〔그림 2－29〕. 이러한 효율 개선은 연간 약 30년생 소나무 152그루가 흡수하는 온실가스 저감 효과에 해당하며(산림청, 2013), 세종 센터로의 확장 과정에서는 재생에너지와 고효율 설비를 결합해 차세대 AI 워크로드 운용을 본격화했다. 국내에서는 민간 자율인증 '그린데이터센터 인증'이 ISO 30134-2 측정 방식을 준용해 PUE와 그린활동지표를 종합 평가하고, 설계·예비·본인증 트랙을 통해 액체냉각·열회수·수자원 관리 등 설계 변수를 정량화하는 체계를 확산시키고 있다. 이는 EU형 공개 및 보고 체계와 연계 가능한 민간 표준이자, 녹색금융과 한국형 녹색분류체계(K-택소노미) 기반의 혜택으로 이어지는 시장 규범으로 기능한다. 특히 열 순환은 구상을 넘어 사업 단계로 진입했다. 한국지역난방공사는 대형 데이터센터 폐열을 지역난방에 재활용하는 프로젝트를 추진 중이며, 반도체 공장의 폐열을 집단에너지로 회수해 난방·급탕에 활용하는 시범사업도 병행하고 있다. 고밀도 AI 연산으로 증가한 랙당 발열을 지역 에너지원으로 전환하는 이 모델은 전력에서 열로 이어지는 선형 소비 구조를 지역 에너지망의 순환 구조로 전환해, 데이터센터와 반도체 생산시설의 총배출 저감 효과를 높인다.

〔그림 2－29〕 네이버 '각' 데이터센터의 탄소 배출량 대비 감소분



자료: Naver, 2020

통신망 부문에서는 5G 및 5G-Advanced 표준에 포함된 절전 기능과 AI 기반 운용 기술이 상용망 전반에 확산되고 있다. 국내 사례에서는 RAN 에너지 효율 지표를 기반으로 한 기지국 슬립 모드, 저전력 운용, O-RAN 기반 절전 제어 등이 적용되고 있으며, 사업자는 심야 저부하 구간에서 기지국 장비와 안테나의 전원을 차단하거나, AI로 냉각 운전을 최적화해 전력 소모를 줄이고 있다. 예를 들어, AI 기반 냉각 제어를 도입한 경우 냉각 전력 소모를 대폭 절감했고, 일부 액체냉각 적용 사례에서는 서버 냉각 전력을 약 93% 줄이고 서버 전력도 10% 이상 절감한 결과가 보고됐다(SKT, 2023). 이러한 망 설계·운영 단계의 AI 최적화와 장비 저전력화는 디지털 트래픽 증가 속에서도 동일 품질의 서비스를 더 적은 전력으로 제공하게 하며, 데이터센터의 저탄소화와 함께 국내 ICT 인프라 전력 수요를 억제하는 핵심 축으로 기능하고 있다.

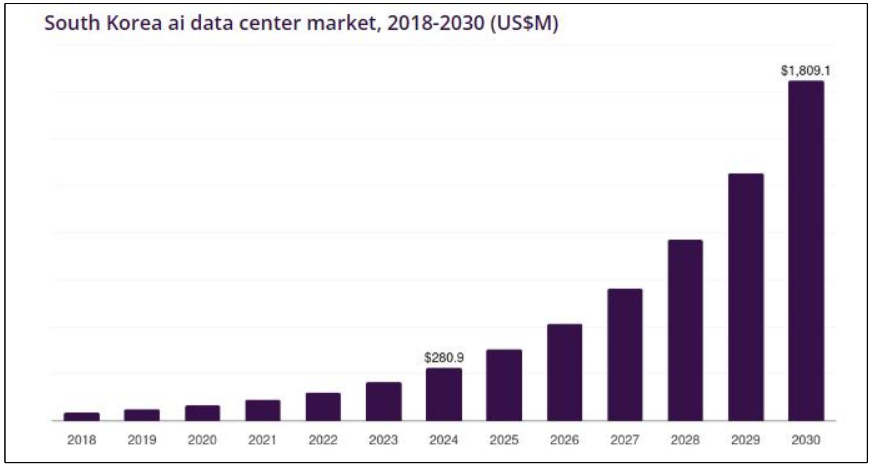
AI·반도체 부문에서는 '저전력·저자원' 중심의 한국형 기술 축이 뚜렷하게 형성되고 있다. SKT 계열 사피온(현 리벨리온)은 X330을 공개하며 전세대 대비 4배의 성능과 2배의 전력 효율을 달성했다고 밝혔다(연합뉴스, 2023). 이 칩은 고성능 AI 추론을 유지하면서도 전력 소비를 최소화하도록 설계되어, 국내 통신·공공·기업 워크로드 전반으로 적용이 확대되고 있다. 리벨리온은 초저지연·저전력 연산에 최적화된 ATOM 칩을 통해, INT8 기준 약 0.5 TOPS/W의 에너지 효율을 제시하며, 실시간 AI 처리, 엣지 컴퓨팅, 금융 거래 감시 등 지연 민감도가 높은 환경에서 경쟁력을 강화하고 있다(연합뉴스, 2023). 퓨리오사AI는

자체 아키텍처(TCP)와 온칩 메모리 최적화를 결합해 에너지 집약적인 LLM 및 비전 추론에서 와트당 성능을 향상시키는 로드맵을 제시하고 있다(EEJournal, 2025). 이러한 행보는 단순한 칩 성능 향상을 넘어, 해외 GPU 대비 전력 효율과 총소유비용(TCO)의 경쟁 구간을 확장하며, 데이터센터 전력 계약, 냉각 방식, 랙 밀도 설계와 같은 핵심 인프라 의사결정 전반에 직접적인 변화를 유도하고 있다. 특히 고밀도 AI 학습·추론이 데이터센터 전력 수요의 핵심 요인으로 부상하는 상황에서, 이들 칩은 전력 효율성과 성능을 동시에 충족시키는 대안으로 자리매김하고 있다.

시장 측면에서는 기술 혁신과 제도 변화가 맞물리며 데이터센터 수요 확대를 가속하고 있다. AI와 클라우드 서비스 확산에 따라 국내 데이터센터 시장은 2024년 약 41억 달러에서 2030년 99억 달러 규모로 성장할 것으로 전망되며, 연평균 성장률은 12%를 웃돈다(buisinesswire, 2025). 특히 AI 연산 수요를 전제로 한 AI 데이터센터 시장은 2024년 약 2억 8천만 달러에서 2030년 18억 달러로, 연평균 36.5%의 고성장이 예상된다〔그림 2－30〕.

이러한 수요 증가는 단순한 시설 확충을 넘어, 전력·냉각·수자원 등 복합 자원 제약을 해소할 수 있는 설계와 운영 전환을 촉발한다. 국내 주요 사업자는 액체냉각, 열회수, 재생전력 매칭을 결합한 고효율 설계 패키지를 신규 센터뿐 아니라 기존 센터 리모델링에도 적용해, 전력 밀도 증가와 냉각 효율 한계 문제를 동시에 해결하고 있다. 나아가 이러한 변화는 에너지 집약적 AI·반도체 워크로드의 지속가능 운용 기반을 마련하고, 데이터센터를 단순한 컴퓨팅 인프라에서 '저탄소·고효율' 플랫폼으로 재정의한다. 이는 국가 디지털 탄소중립 목표 달성에 기여할 뿐 아니라, 국제 경쟁 환경 속에서 국내 데이터 인프라의 신뢰성과 투자 매력도를 높이는 전략적 효과를 가져온다.

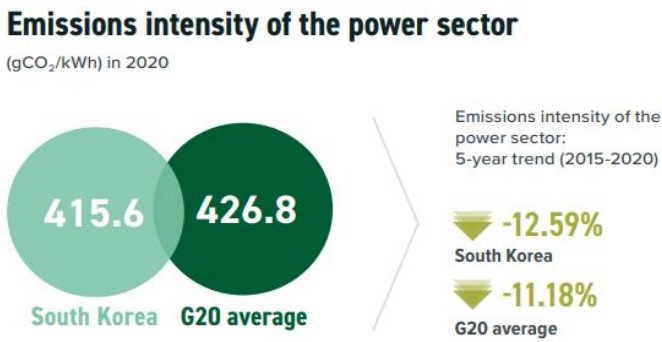
〔그림 2－30〕 한국의 AI 데이터센터 시장 성장률



자료: grandviewresearch

국내 디지털 탄소중립 체계에서 측정 부문은 여전히 미완의 과제로 남아 있다. 현재 전력 부문 온실가스 배출량은 기준 연료별 국가 단위 계수를 중심으로 산정되고 있으며, 2021년 기준 전력 1 kWh당 약 415.6g의 탄소가 배출되는 것으로 보고된다〔그림 2－31〕. 이는 G20 평균을 상회하는 수준이지만, 상업 시스템 전반에 실시간 또는 시간대별 탄소 강도를 적용하는 단계에는 이르지 못했다. 학술·정책 연구에서는 시간·지역별 배출계수 산정과 공개의 필요성이 지속적으로 제기되고 있으며, OECD와 EU에서는 이미 이러한 데이터의 표준화와 공개를 탄소중립 이행의 핵심 기반으로 논의하고 있다. 한국 역시 탄소 강도 기반의 스케줄링, 데이터센터 입지 최적화, 연산 워크로드 조정 등을 확산하기 위해서는 전력시장 데이터와 배출계수의 실시간 공개 체계를 구축해야 한다. 이를 통해 측정-감축-거버넌스의 선순환을 실현하고, 디지털 인프라의 탈탄소화를 가속할 수 있을 것이다.

〔그림 2－31〕 한국과 G20의 배출량 비교



자료: Climate Transparency

한국의 흐름은 K-RE100과 분산에너지법으로 저탄소 전력 접근성과 지역 분산을 강화하고, 데이터센터·네트워크·AI·반도체 전반에서 효율을 높이는 방향으로 수렴한다. 향후 실시간 탄소 데이터 인프라와 탄소 대응적 운영이 상용화되면, OECD·EU 수준의 지속가능 인프라 선순환에 한층 근접하게 될 것이다.

제3장 산업별 디지털 기반 탄소감축 적용 사례

제1절 디지털 인프라의 저탄소 전환 방안

1. 데이터 센터 에너지 효율 최적화

가. 그린 데이터센터 정의 및 필요성

그린 데이터센터는 데이터센터의 전력 사용과 냉각 과정에서 발생하는 에너지 낭비를 최소화하고, 자원 효율성을 극대화하여 환경에 미치는 부정적인 영향을 줄이기 위해 설계·운영되는 시설이다. 이러한 데이터센터는 물리적 인프라부터 운영 절차까지 전 과정에서 고효율 장비와 설계를 적용하며, 운영 중 발생하는 폐열을 재활용하거나 에너지 관리 소프트웨어를 통해 지속적으로 효율성을 모니터링한다.

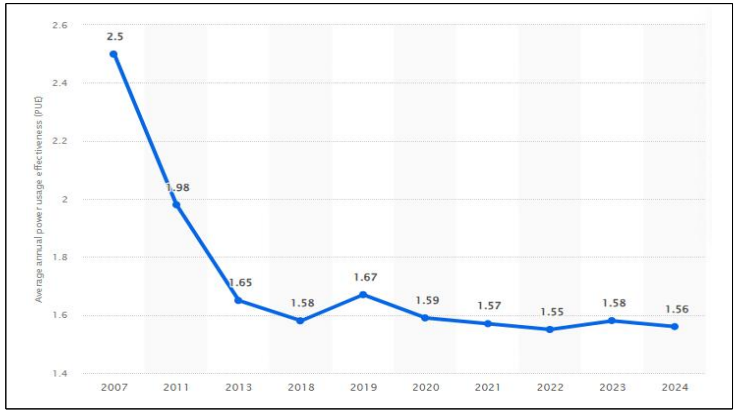
데이터센터의 에너지 효율을 객관적으로 평가하기 위해서는 단순한 전력 사용량뿐만 아니라, 운영 전 과정에서의 효율성 지표를 정량화하는 체계적 접근이 필요하다. 이에 따라 국제적으로 가장 널리 활용되는 지표가 PUE, Water Usage Effectiveness(WUE), Carbon Usage Effectiveness(CUE) 등이다. 이 지표들은 각각 전력, 물, 탄소 배출 측면에서 데이터센터의 효율성을 다각적으로 평가할 수 있게 해준다. 우선 PUE는 데이터센터 전체 소비 전력 대비 IT 장비가 직접 사용하는 전력의 비율을 의미한다.[29] 이 값이 1.0에 가까울수록 효율적임을 의미하며, 일반적으로 1.1~1.3 수준이면 고효율 데이터센터로 분류된다. 2007년 평균 2.5 수준이던 글로벌 데이터센터의 PUE는 2024년 기준 약 1.55 수준까지 개선되었으며, 구글, 마이크로소프트 등은 자체 AI 제어 냉각 시스템을 통해 1.1 이하를 달성하였다[30]〔그림 3-1〕. 다음으로 WUE는 데이터센터의 냉각 과정에서 사용된 물의 양(리터)을

29) Vertiv(2024), "What Is the PUE(Power Usage Effectiveness) Data Center Definition", https://www.vertiv.com/en-us/about/news-and-insights/articles/educational-articles/what-is-the-pue-data-center-definition/

30) Statista(2024), "Average annual power usage effectiveness(PUE) of data centers world

소모한 전력(kWh)으로 나눈 값으로, 단위는 L/kWh이다[31]. WUE는 기후 조건에 따라 달라지며, 물 사용이 많은 지역에서는 재활용 냉각수나 공랭식 설비를 활용해 절감 효과를 높이는 것이 중요하다. CUE는 데이터센터의 탄소 배출량을 소모한 전력으로 나눈 값으로, 전력원 구성의 청정도와 재생에너지 사용 비율을 반영한다[32]. CUE가 낮을수록 동일한 연산량에 비해 탄소 배출이 적은 친환경 데이터센터임을 의미한다. 이러한 세 가지 지표를 통합적으로 관리함으로써, 단순한 전력 효율성 중심의 관리에서 벗어나 물과 탄소까지 포함한 종합적 지속가능성 평가 체계로 전환이 가능하다. 특히 최근 한국의 Green Data Center 인증제도나 EU의 Energy Efficiency Directive 개정안 등은 PUE뿐 아니라 WUE와 CUE를 포함한 다차원 평가 방식을 사용하고 있다.

〔그림 3－1〕 구글 PUE 개선 그래프(2007~2024)



자료: Average annual power usage effectiveness(PUE) of data centers worldwide

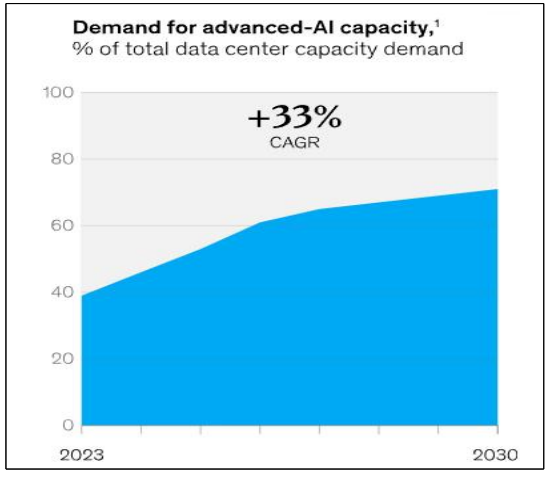
wide from 2007 to 2023", https://www.statista.com/statistics/1229367/data-center-average-annual-pue-worldwide/

31) Equinix(2024. 11. 13.), "What Is Water Usage Effectiveness(WUE) in Data Centers", https://blog.equinix.com/blog/2024/11/13/what-is-water-usage-effectiveness-wue-in-data-centers/

32) Stream Data Centers(2024), "Data Center Sustainability: Carbon Usage Effectiveness(CUE) Definition", https://www.streamdatacenters.com/resource-library/glossary/data-center-sustainability/

AI와 대규모 연산 수요의 확산 또한 데이터센터의 에너지 관리 패러다임을 근본적으로 바꾸고 있다. 2023년 이후 생성형 인공지능과 대규모 언어모델의 급격한 성장으로 인해 GPU 클러스터를 중심으로 한 고밀도 서버 인프라가 빠르게 증가하고 있다. Goldman Sachs의 보고서에 따르면, 2030년까지 전 세계 데이터센터 전력 소비량은 2023년 대비 약 165% 증가할 전망이다.[33] McKinsey & Company는 AI 기반 데이터센터 전력수요의 연평균 증가율이 약 33%에 달할 것으로 추정했으며, AI 전용 서버의 단위 전력 소모는 기존 8 kW 수준 CPU 서버에 비해 최대 40~60 kW 수준까지 상승하고 있다[34]〔그림 3-2〕. 이러한 급격한 증가세는 데이터센터 전력 구조를 근본적으로 변화시키고 있다. 과거에는 평균 부하 기반의 전력 관리가 가능했으나, 현재는 상시 최대부하 상태가 일반화되면서 냉각 에너지의 비중이 전체 소비 전력의 절반 이상을 차지하는 상황에 이르렀다.

〔그림 3-2〕 McKinsey의 AI용 전력수요 그래프



자료: Expanding Data Center Capacity to Meet Growing Demand

33) Goldman Sachs(2024), "AI to Drive 165% Increase in Data Center Power Demand by 2030", https://www.goldmansachs.com/insights/articles/ai-to-drive-165-increase-in-data-center-power-demand-by-2030

34) McKinsey & Company(2024), "AI Power: Expanding Data Center Capacity to Meet Growing Demand", https://www.mckinsey.com/industries/technology-media-and-telecommunications/our-insights/ai-power-expanding-data-center-capacity-to-meet-growing-demand

최근 데이터센터의 전력 사용량이 급증함에 따라, 특히 AI, 빅데이터, 클라우드 서비스 확산으로 인한 연산 수요 증가가 환경 부담을 가중시키고 있다. 이에 따라 에너지 효율성을 객관적으로 평가하고 인증하는 제도적 장치가 필수적으로 요구되고 있다.

각 국가는 자국의 기후 조건, 산업 구조, 에너지 정책 방향에 맞춰 데이터센터의 친환경 전환과 에너지 효율화를 위한 로드맵을 수립하고 있다〈표 3－1〉.

〈표 3－1〉 국가별 그린 데이터센터 로드맵

국가	로드맵	방안	차별점
한국	2050 탄소중립 로드맵, ZEB 인증 의무화(2025~), GDC 인증 도입 및 확산	데이터센터 지방 분산, PUE 개선, 재생에너지·친환경 기술 적용 확대	탄소중립·에너지자립률 중심, 지방 입지 확대
중국	AI 전력수요 대응 3대 축(공급확대·수요관리·기술혁신), 동수서산 전략	서부 내륙 대규모 허브 구축, 재생·원자력·석탄 혼합 전원, UHV 송전망	전국 규모 전력·컴퓨팅 자원 재배치, 규제·투자 병행
미국	국가 에너지 비상사태(2025), 전력 공급 확대 및 규제 완화, 대규모 AI 인프라 투자 유치(스타게이트 프로젝트)	송전 인프라 확충, SMR 조기 상용화, 자율적 설비 확대, 환경규제 완화	속도 우선, 산업 자율성 보장, 대규모 민간투자 유치
싱가포르	그린 데이터센터 로드맵(2024), BCA-IMDA 그린 마크 갱신, 에너지 효율 표준(2025)	PUE 1.3 이하 달성, 액체냉각·IT장비 효율화, 친환경 표준 제정, 인센티브 제공	고효율 서버·SW, 냉각수 재활용, 배출량 자원 효율 보조금
EU	에너지효율지침(EED) 개정, 최소 효율 기준 검토, 폐열 회수 의무	회원국별 규제와 EU 공동목표 병행, 재생에너지 전환, 폐열 지역난방	엄격한 환경·효율 규제, 폐열 활용, 회원국간 표준화

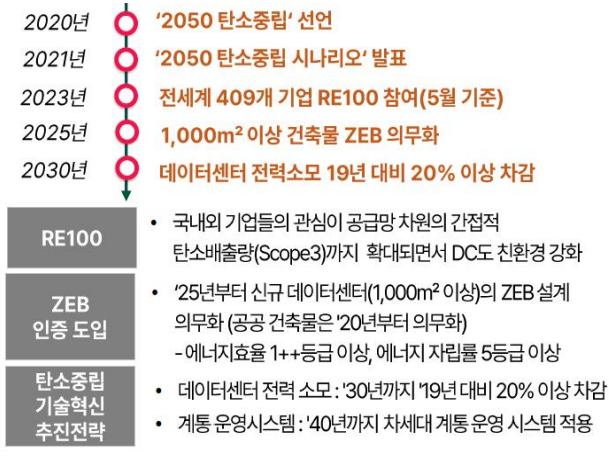
나. 국가별 그린 데이터센터 로드맵

한국은 그린 데이터센터 인증(GDC)을 통해 에너지 효율, 재생에너지 사용, 탄소배출 저감 등 요소를 종합적으로 평가한다[35]. GDC 인증은 데이터센터의 전력 사용량, 재생에너지 조달 비율, 냉각·전력변환 효율성, 온실가스 감축 실적 등을 객관적으로 검증해 부여되며,

35) 그린데이터센터, https://kdcc.or.kr/green/pageview.do?url=green01&keyvalue=green

지속 가능한 데이터센터 설계와 운영을 유도하는 제도다. 아울러 정부는 RE100 이행을 촉진하기 위해 고정가격계약 입찰시장 연계형 전력구매계약(PPA) 중개시장 제도 2차 시범사업을 운영하고 있다[36]. 이 제도는 소규모 재생에너지 발전사업자의 참여 문턱을 낮추고, 협의 기간과 계약 방식을 유연화함으로써 데이터센터를 포함한 RE100 수요기업의 재생에너지 조달 경로를 다변화한다. 이러한 전략은 탄소중립과 에너지자립률 제고라는 정책 기조에 재생에너지 직접조달 확대라는 요소를 결합한 것이 특징이다. 정부는 2025년 탄소중립 선언 이후 국내 ESG 목표와 데이터센터 관련 세부 규율을 강화하고 있다. 그 핵심 목표 중 하나는 2030년까지 데이터센터 전력소모를 2019년 대비 20% 이상 감축하는 것이다. 이를 위해 냉각 효율 개선, 저전력 서버 확산, 재생에너지 전환, 가상화 기술 고도화 등 다각적인 절감 방안을 추진하고 있다. 또한 2025년부터는 연면적 1,000m² 이상 신규 데이터센터에 ZEB(Zero Energy Building, 제로에너지건축물) 설계 의무가 적용된다〔그림 3－3〕. ZEB 인증은 건물의 연간 에너지 소비량을 최소화하고, 소비한 에너지와 동등하거나 그 이상의 재생에너지를 생산·조달하는 구조를 갖춘 경우 부여된다. 이를 위해 고단열·고기밀 설계, 고효율 장비, 태양광·지열 등 재생에너지 시스템의 통합이 필수적이다.

〔그림 3－3〕 국내 ESG 목표 및 데이터센터 관련 세부 규율



자료: 하나금융연구소(2024)

36) 한국에너지공단(2024), 『KEA 에너지 이슈 브리핑』, 제252호.

데이터센터에 ZEB 인증을 적용하면 전력 집약적인 운영 구조에서 발생하는 탄소 배출을 획기적으로 줄이는 동시에, 국제 친환경 인증 기준을 충족할 수 있다. 이와 함께 ESG (Environment·Social·Governance, 환경·사회·지배구조) 경영 기조도 데이터센터 운영 전략에 적극 반영되고 있다. ESG는 기업의 재무 성과뿐 아니라 환경 보호, 사회적 책임, 투명한 지배구조를 종합적으로 평가하는 개념으로, 글로벌 투자자와 고객사들이 기업을 평가하는 핵심 지표다. 이에 따라 데이터센터 사업자는 RE100 이행, GDC·ZEB 인증 획득, 폐열 재활용, 지역 사회 기여 활동 등을 통해 ESG 성과를 강화하고 있다.

싱가포르는 2024년 그린 데이터센터 로드맵[37]을 발표하고, PUE 1.3 이하를 목표로 설정하였다. 이에 따라 액체냉각 시스템, 고효율 IT 장비, 친환경 설계 표준을 제정하며 인센티브를 제공하고 있다. 2025년까지 BCA-IMDA 그린 마크 갱신과 IT 장비 에너지 효율 및 액체 냉각 관련 표준을 도입할 예정이다. 특히, 열대 기후 최적화 설계, 산업계와 공동 표준 개발이 주요 특징이다.

미국은 2025년 국가 에너지 비상사태[38]를 선포하고 데이터센터 전력 공급을 신속히 확충하기 위한 규제 완화, 송전 인프라 강화 및 SMR(소형모듈원자로) 조기 상용화를 추진 중이다. 하이브리드 전원(태양광·가스·배터리·SMR) 체계 구축 및 민간 주도의 대규모 AI 인프라 투자(예: 스타게이트 프로젝트 등을 통한 구축 확장)를 통해 '속도 우선 자율 확장' 전략을 택하고 있다.

중국은 'AI 전력수요 대응 3대 축' 전략(전력 공급 확대·수요관리·기술혁신)을 바탕으로, 국가 전략 '동수서산(東數西算)'을 통해 서부 내륙에 대규모 데이터센터 허브를 구축하고 있다[39]. 이를 UHV(초고압) 송전망과 결합해 동부에서 생성된 데이터 수요를 서부에서 처리하도록 하며, PUE 1.5 이하 유지, 재생에너지 비중 매년 +10%p 증가, AI 칩 전력 성능 기준 도입을 병행한다. 대규모 컴퓨팅 자원 재배치와 규제 및 투자 병행이 주요 특징이다.

EU는 2023년 개정된 에너지효율지침(EED)[40]을 기반으로, 데이터센터를 대상으로 한 보

37) IMDA(2024), Charting green growth pathways at scale for data centres in Singapore

38) U.S. Department of Energy(2025), Unleashing American Energy: Executive Order Implementation Report

39) KOTRA(2022. 3. 22.), "中 메가급 프로젝트 '동수서산'(東數西算)", https://dream.kotra.or.kr/dream/kotra/actionKotraShortUrl/aAs2Jlw4NPr6.do

40) European commission(2023), Energy Efficiency Directive

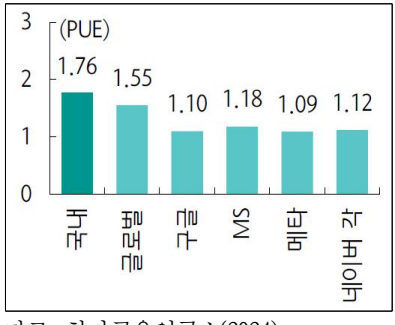
고 의무제도를 시행하고 있다. 2024년 9월부터는 IT 전력 수요 ≥ 500 kW인 데이터센터들이 데이터센터 내 설정온도, 용수 사용량, 재생에너지 비율, 폐열 재활용 등 KPI를 포함한 정보를 매년 보고하도록 의무화되었다. 해당 정보는 전 유럽 공통 데이터베이스에 집계된다. 이처럼 EU는 그린클라우드 및 그린데이터센터를 구축하여 엄격한 환경·효율 규제와 투명한 표준화, 회원국 간 조정체계를 통해 데이터센터 산업의 지속가능성을 주도하고 있다.

다. 기업별 데이터센터 에너지 효율화 방안

1) 고효율 장비, 첨단 냉각·열관리 기술 적용으로 전력 소모 최소화

국내 데이터센터의 평균 PUE는 1.76으로, 글로벌 평균 1.55 대비 여전히 비효율적인 수준이다. 반면, ESG 경영을 선도하는 구글과 마이크로소프트(MS) 등의 PUE는 약 1.0~1.2로, 글로벌 최상위 수준을 유지하고 있다. AI 확산에 따른 GPU 서버 확대와 고밀도 데이터센터 증가 추세는 발열을 심화시키며 PUE 개선에 부정적인 영향을 주고 있다〔그림 3-4〕. 실제로 데이터센터 전력 사용의 40 ~ 50%가 냉각에 소요되며, 향후 AI 워크로드 증가로 냉각 수요는 더욱 커질 것으로 전망된다. 이에 따라 글로벌·국내 기업들은 기존 공랭식 대비 효율성이 높은 수냉식 적용을 확대하고 있으며, 액침 냉각 등 유체 기반 냉각 기술도 시험·개발 중이다. 예를 들어, SK엔무브는 액침 냉각 시스템 개발을 위해 미국 GRC(Green Revolution Cooling)에 약 330억 원을 지분 투자하였다. 또한 SK에코플랜트는 연료전지를 보조 전원으로 채택한 데이터센터를 국내에 건립하여 안정성과 친환경성을 동시에 강화했다[41].

〔그림 3-4〕 글로벌 기업들의 데이터센터 평균 PUE 비교



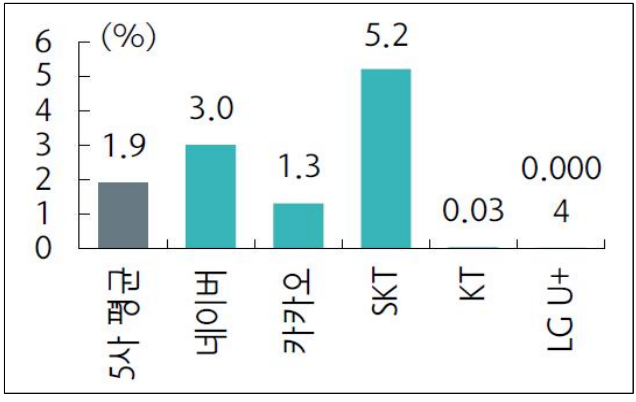
자료: 하나금융연구소(2024)

41) SK 에코플랜트, https://news.skecoplant.com/plant-tomorrow/18677

해외에서는 냉각 효율화를 위해 기후 조건이 유리한 북유럽 등 저기온 지역으로 데이터센터를 이전하는 움직임이 뚜렷하다. MS는 무수(Waterless) 침지 냉각 기술을 도입해 WUE(Water Usage Effectiveness)를 0.49 L/kWh에서 0.30 L/kWh로 39% 개선했으며[42], 해저 데이터센터 실험을 통해 서버 신뢰성을 기존 대비 8배 향상시켰다. 구글은 AI 기반 실시간 냉각 제어를 통해 냉각 전력 소모를 평균 30% 절감했다.

중국에서도 국유 전력망 기업(State Grid)이 AI 데이터센터 인근에 100MW급 ESS를 설치하여 전력망 부하를 완화하고, 알리바바, 바이두, 텐센트 등 주요 ICT 기업이 AI 기반 냉방 최적화 알고리즘, 난방 폐열 활용 등 에너지 절감 솔루션을 도입하고 있다.

〔그림 3-5〕 국내 ICT 기업의 재생에너지 사용 비중



자료: 하나금융연구소(2024)

2) 재생에너지 전력 조달(PPA 등) 확대로 데이터센터 전력 탄소중립 실현

국내 주요 ICT 기업 5곳의 2022년 탄소배출량은 382만 tCO_2 에 달하며, 이는 대부분 데이터센터 운영에서 발생한다. 그러나 재생에너지 사용 비중은 평균 1.9%에 불과하여, 탄소중립 달성을 위해서는 재생에너지 직접 조달 확대가 필수적이다〔그림 3-5〕.

정부는 RE100 이행을 위한 고정가격계약 입찰시장 연계형 전력구매계약(PPA) 중개시장

42) Microsoft(2024), Sustainable by design: Next-generation datacenters consume zero water for cooling

제도 2차 시범사업을 통해 소규모 재생에너지 발전사업자의 진입 장벽을 낮추고, 협의 기간 및 계약 방식을 유연화해 데이터센터를 포함한 RE100 수요기업의 재생에너지 조달 경로를 다양화하고 있다[43]. 이 제도는 최소 참여용량을 기존 1MW에서 300kW로 낮추고, 1:N·N:1 매칭 방식을 허용해 다수 발전사업자와 수요기업이 결합 계약을 체결할 수 있도록 하였다.

해외에서는 빅테크 기업들이 PPA를 적극 활용하고 있다. 10 ~ 22년 청정에너지 구매 글로벌 상위 기업 1 ~ 4위를 아마존, 메타 등의 빅테크 기업들이 차지하며 데이터 센터 관련 기업들의 친환경 개선 노력을 하고 있다〔그림 3－6〕. 알리바바는 내몽고·귀주 등에서 대규모 데이터센터를 운영하며 자체 풍력·태양광 발전을 연계했고, 허베이·장베이에서는 중국 최초의 탄소제로 데이터센터를 구축했다. 텐센트는 산시 화이라이에 풍력·태양광·배터리 결합 마이크로그리드를 구축해 인근 클라우드 데이터센터에 연간 7천만 kWh의 녹색전력을 공급하고 있다.

〔그림 3－6〕 10~22년 청정에너지 구매(PPA) 글로벌 상위 기업 현황

자료: 하나금융연구소(2024)

43) 전기신문(2025. 4. 28.), “입찰연계형 PPA 2차 시범사업, 용량 완화·계약 유연화로 재시동”, https://www.electimes.com/news/articleView.html?idxno=353993

미국에서는 마이크로소프트, 아마존, 구글 등의 기업은 핀란드·펜실베이니아·워싱턴 등지에서 데이터센터 폐열 활용, 원전 PPA 체결, SMR 투자 등을 통해 AI 전력 수요에 대응하는 청정에너지 인프라를 선제적으로 구축하고 있다. 또한 대부분의 미국 빅테크 기업에서는 탄소 인지 스케줄링(Carbon-Aware Scheduling)을 활용하여 전력망의 탄소 배출 강도가 낮은 시간대·지역으로 컴퓨팅 작업을 이동시키는 방식을 보편화하고 있다.

2. 클라우드 전환

가. 대규모 클라우드 데이터센터를 통한 전력소비 및 탄소배출 저감

전 세계 기업과 공공기관은 기존의 물리적 온프레미스(On-Premises) 서버에서 벗어나, 대규모 클라우드 데이터센터를 기반으로 한 인프라로 빠르게 전환하고 있다. 이러한 변화는 단순한 IT 인프라 구조 개선이 아니라, 에너지 효율 극대화와 탄소배출 저감을 동시에 달성하는 핵심 전략으로 자리 잡고 있다.

노스캐롤라이나대 연구진은 1997년부터 2017년까지 미국 내 57개 산업의 데이터를 분석해, 클라우드 기반 IT 서비스 도입이 산업 전반의 에너지 효율 개선에 실질적인 기여를 했음을 입증했다. 특히, 서비스형 소프트웨어(SaaS)는 모든 산업군에서 생산성과 에너지 효율성을 동시에 높였으며, 서비스형 인프라(IaaS)는 하드웨어 의존도가 높은 제조·통신·금융 분야에서 두드러진 효율 향상을 보였다[44]. 이는 클라우드 전략 수립 시 생산성과 비용 절감뿐 아니라, 에너지 절감 효과를 의사결정의 주요 변수로 고려해야 함을 시사한다.

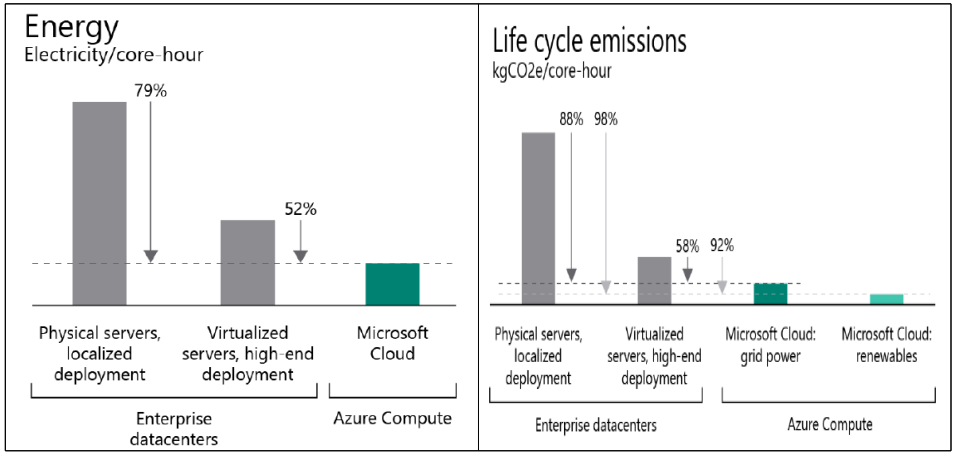
2020년 Microsoft-WSP 공동 연구는 이러한 경향을 실증적으로 뒷받침한다[45]. 연구에 따르면, 전통적인 기업 데이터센터에서 동일한 연산 기능을 수행하는 것 대비 Microsoft Azure Compute와 같은 대규모 클라우드 데이터센터로 전환 시 52~79% 더 높은 에너지 효율성을 제공한다〔그림 3-7〕. 또한 Microsoft는 전체 전력 소비의 95% 이상을 재생에너지로 조달하고 있으며, 이를 감안할 경우 탄소 배출량은 최대 92~98%까지 감소한다. 이 연구

44) 동아일보(2023. 7. 10.), "편리한 클라우드 서비스, 에너지 절약에도 '효자'", https://www.donga.com/news/Economy/article/all/20230709/120151296/1

45) Microsoft(2020), The carbon benefits of cloud computin: A study on the Microsoft Cloud in partnership with WSP

는 데이터센터 IT 장비, 인프라 설계, 운영 방식, 네트워크 트래픽 관리 등 전 과정의 환경 영향을 종합적으로 평가하여, 단순한 장비 효율 개선보다 재생에너지 도입과 결합될 때 클라우드 전환 효과가 극대화됨을 보여준다.

〔그림 3－7〕 Microsoft의 클라우드 데이터센터 이점 비교



자료: Microsoft(2020)

비슷하게, AWS(아마존 웹 서비스)의 분석 보고서[46]에 따르면, AWS 인프라가 온프레미스 대비 최대 4.1배 에너지 효율적이라는 사실을 밝혀냈다. 또한 워크로드 최적화를 병행할 경우 탄소 배출을 최대 99%까지 절감할 수 있다. 이러한 효과는 컴퓨트 중심 워크로드뿐 아니라 스토리지 중심 환경에서도 동일하게 적용되며, AWS 맞춤형 실리콘 칩(Graviton 시리즈)과 아키텍처 현대화를 도입하면 추가적인 절감 효과가 발생한다. 이는 랙 설계 최적화, 전력 분배 시스템 고도화, 고효율 냉각 기술, 탄소 무배출 전력 조달이 복합적으로 작용한 결과다. 실제 사례로, Qube Cinema와 Shyam Steel은 AWS로 이전해 스토리지·백업·머신러닝 워크로드를 클라우드로 이전함으로써 최대 80% 비용 절감과 데이터 처리 속도 개선을 달성했다[47].

46) AWS(2024), How moving onto the AWS cloud reduces carbon emissions
47) AWS(2022), 3개 회사가 AWS를 통해 클라우드를 도입하여 비용을 절약한 방법

국내에서도 유사한 성공 사례가 나타나고 있다. KT의 '엔지니어링 플랫폼 서비스'는 고성능 클라우드 컴퓨팅(HPC) 자원과 업계 표준 해석 솔루션을 SaaS 형태로 제공하는 모델로, 특히 제조업 분야에서 에너지 절감과 개발 효율화를 동시에 달성했다. 기존에는 4 ~ 5일 이상 소요되던 해석 시뮬레이션 작업이 하루 만에 완료되고, 비용은 최대 90% 절감되었다[48]. 이는 기업이 자체 HPC 인프라를 구축·운영할 때 발생하는 막대한 전력 소비와 냉각 비용을 대폭 줄인 대표 사례로, 클라우드의 탄소 절감 효과를 뚜렷하게 보여준다.

또한, 국내에서는 네이버의 'GAK 세종' 데이터센터가 대표적인 친환경 클라우드 인프라 전환 사례로 꼽힌다. 네이버는 경기도 춘천의 1센터 이후, 2023년 11월 세종시에 국내 최대 규모의 AI·클라우드 데이터센터 GAK 세종을 설립하였다. 이 데이터센터는 바람을 활용한 냉각 효율화 기술과 에너지 절감 시스템을 적용하여 일반 데이터센터 대비 약 70% 수준의 냉각 에너지 절감을 실현하였고 해당 센터는 평균 PUE에서 1.09를 달성하였다[49][50]. 또한, 전년 대비 약 22% 증가한 총 3,065톤의 온실가스를 재생에너지 사용으로 온실가스 사용을 대체했고 2023년의 온실가스 배출량은 89,505톤으로 전년 대비 3% 미만 소폭 증가하였으나, 집약도는 전년 대비 13% 감소하였다〔그림 3－8〕. 지속적인 감소세 기록 데이터센터 운영 전 과정은 네이버 클라우드 플랫폼으로 통합 관리되어 공공기관 및 금융 부문 SaaS 서비스까지 확장되고 있으며, 디지털 인프라 효율화와 탄소중립을 동시에 실현하고 있는 사례이다.[51]

48) ZDNET Korea(2024. 6. 16.), "제조업 설계도 클라우드로...KT, 비용 60% 절감", https://zdnet.co.kr/view/?no=20240616072215

49) DataCenterDynamics(2023. 11. 7.), "Naver launches 270MW data center campus in Sejong, South Korea."

50) TheLec(2023. 11. 3.), "네이버, AI 데이터센터 '각 세종' 공개...냉방에너지 73% 절감."

51) NAVER, 2023 통합보고서(Integrated Report 2023), 금융감독원 전자공시시스템(KIND), https://kind.krx.co.kr/external/2024/06/26/000486/20240625001176/NAVER_Integrated_Report_2023_ver3.pdf

〔그림 3－8〕 온실가스 배출량 및 감축량

자료: NAVER INTEGRATED REPORT 2023

나. 정부의 공공부문 클라우드 전환 및 클라우드 산업

정부는 2030년까지 기존 정보시스템의 90%를 클라우드 네이티브로 전환하고, SaaS 적용률을 70%까지 확대하는 목표를 설정했다[52]. 이는 2026년까지 클라우드 네이티브 50%, SaaS 35%였던 기존 계획을 대폭 상향한 것으로, 행정 효율성과 보안성, 서비스 품질을 동시에 향상시키겠다는 의지가 반영된 것이다. 클라우드 네이티브 전환이란 애플리케이션의 설계, 개발, 운영 전 과정에서 클라우드의 장점을 최대한 활용해 성능과 효율성을 극대화하는 접근 방식이다. 이 과정에서 확장성, 자동화, 장애 복구 능력이 크게 향상된다.

공공부문 클라우드 전환에서는 보안 인증과 안정성 확보가 필수적이다. 예를 들어, 네이버 클라우드 플랫폼은 다수의 보안 인증과 공공기관 심의 요건을 충족하며, 전국민 대상 공공 서비스 제공에 필요한 확장성과 편의성을 갖춘 솔루션을 제공하고 있다. 이는 민간 클라우드 사업자가 공공 클라우드 생태계 확장과 산업 경쟁력 강화에 기여할 수 있음을 보여준다. 또한 KT, NHN이 공공 및 금융 부문을 중심으로 맞춤형 프라이빗클라우드를 제공하고 있다.

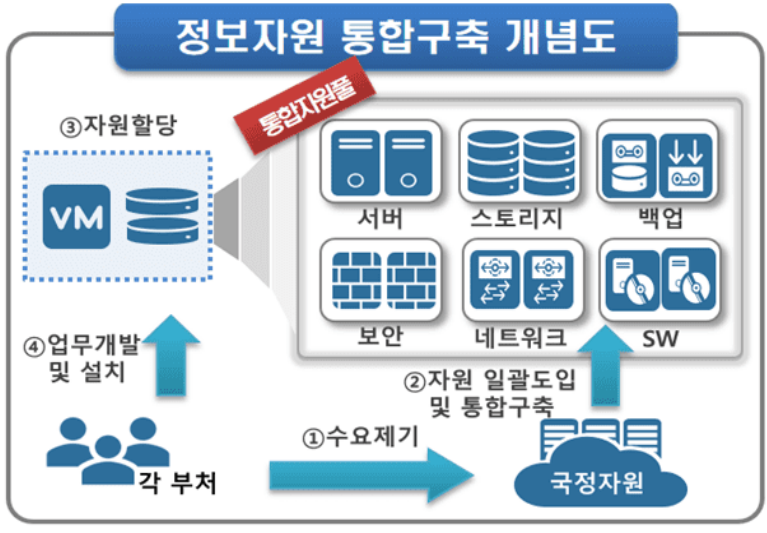
한국지능정보사회진흥원(NIA)의 사례는 공공기관 클라우드 전환의 구체적인 성과를 보

52) 전자신문(2024. 8. 12.), 정부, “2030년까지 기존 시스템 90% 클라우드 네이티브로 전환”, https://www.etnews.com/20240812000245

여준다. NIA는 자체 서버를 클라우드로 이전한 결과, 직접적인 성과로 약 1,254만 원의 에너지 절감과 961만 원 상당의 탄소 배출 저감을 달성했다. 더 나아가, 공공기관 전반으로 동일한 전환이 확산될 경우 간접 성과로 약 2억 4,210만 원의 에너지 절감, 1억 8,548만 원의 탄소 저감 효과가 발생할 것으로 분석되었다[53]. 이는 클라우드 이전이 단순히 물리 서버를 가상 서버로 대체하는 것을 넘어, 서버 집적화, 냉각 효율 개선, 전력 피크 부하 완화 등 다층적인 환경·경제적 효과를 창출함을 의미한다.

또 다른 사례로, 행정안전부는 정부 클라우드 인프라의 핵심 플랫폼인 G-클라우드를 기반으로, 공공기관의 분산된 정보시스템을 단계적으로 통합 및 전환하고 있다. 국가정보자원관리원은 2025년 기준 약 135억 원 규모의 정보자원 통합 구축 사업을 추진 중이며, 소방청 119시스템, 질병관리청 통합관리시스템, 행안부 공공데이터포털 등 39개 주요 업무 시스템을 클라우드 환경으로 이전하고 있다〔그림 3-9〕. 특히 2024년부터는 오픈소스 기

〔그림 3-9〕 정보자원 통합구축 개념도



자료: 국가정보자원관리원

53) 사회적가치연구원(2024), 『한국지능정보사회진흥원(NIA)의 환경성과 측정』, 제 4호.

반 관리체계와 자동화 운영 솔루션을 도입하여 컴퓨팅 자원 활용률을 향상시켰으며, 시스템 통합 관리로 데이터센터의 전력 효율과 자원 활용률이 개선되었다.[54]

조달청은 공공기관이 민간 클라우드 기반 디지털서비스와 SaaS를 신속하게 도입할 수 있도록 디지털서비스 전문계약제도를 운영하고 있다. 이 제도는 기존의 복잡한 입찰 과정을 간소화하였으며, 수개월이 걸리던 계약 절차를 평균 2주 이내로 단축시켰다. 2024년 기준 제도를 통해 체결된 누적 계약 금액은 약 5,000억 원, 계약 건수는 1,500건 이상으로 집계되었으며, 공공 SaaS 계약만 해도 2021년 26건에서 2024년 59건으로 증가하는 등 3년간 약 2.3배 성장하였다[55]. 또한 디지털서비스몰을 운영해 공공기관이 보안 인증을 받은 SaaS를 즉시 검색 및 도입할 수 있도록 지원하고 있으며, 이를 통해 중소 클라우드 기업의 공공시장 진입 장벽을 완화하였다. 대표적으로 클라우드 MSP 기업인 디딤 365는 동 제도를 활용해 누적 72건의 계약을 체결하여 공공부문의 클라우드 도입 확산의 촉진 역할을 수행하고 있다[56].

3. 통신망 인프라 효율화

가. AI를 활용한 기지국·통신장비 고효율화로 통신망 전력 소비 절감

통신망은 국가 전체 전력 소비에서 차지하는 비중이 점차 커지고 있으며, 5G와 차세대 네트워크의 확산에 따라 그 증가 속도는 더욱 가팔라지고 있다. 이에 따라 국내외 통신사들은 인공지능(AI)을 활용해 네트워크 전력 사용을 줄이고, 장비 운용 효율성을 극대화하는 방안을 적극 도입하고 있다.

SK텔레콤은 유선망과 5G/LTE망 전반의 제어·점검 과정을 중앙 집중식으로 자동화하는 'AI 오케스트레이터'를 자체 개발하여 도입했다[57]. 이 시스템을 통해 과거 며칠이 소요되던

54) 전자신문(2024. 8. 12.), 『정부, 2030년까지 기존 시스템 90% 클라우드 네이티브로 전환』, https://www.etnews.com/20240812000245

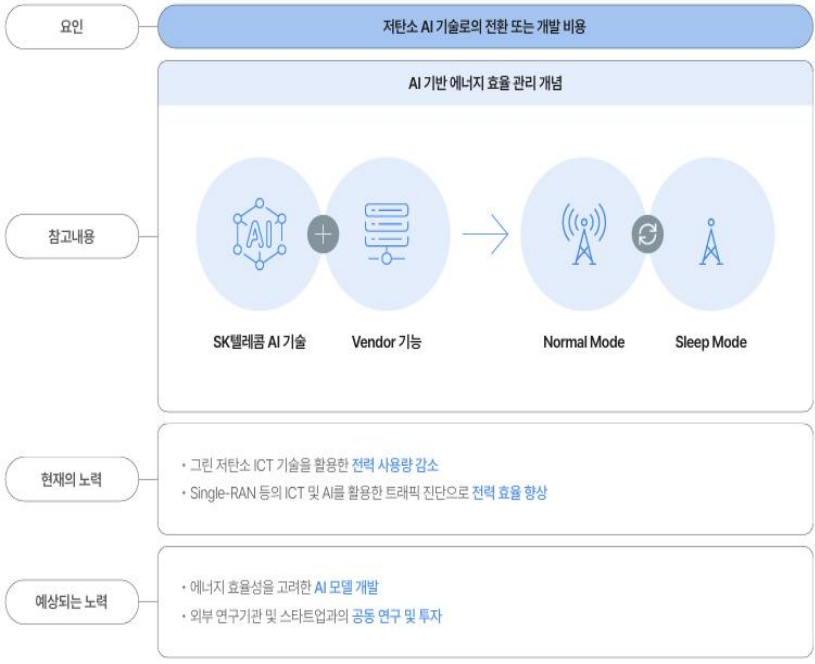
55) 한국지능정보사회진흥원(NIA)(2024.10.02.), 『디지털서비스 전문계약제도 계약 금액 5천억 원 돌파』, https://www.nia.or.kr/site/nia_kor/ex/bbs/View.do?bcIdx=27271&cbIdx=90549

56) 디딤365(2025. 7. 22.), 『클라우드 MSP 분야 최다 계약 달성』, https://didim365.com/press/newspress_20250722

57) SK텔레콤(2024), Annual Report 2024

네트워크 점검·최적화 작업을 하루 만에 완료할 수 있게 되었으며, 수동 작업에서 발생할 수 있는 오류 가능성을 대폭 줄였다. 또한 네트워크 전 구간의 장비 상태를 실시간으로 통합 모니터링하여 운용 안정성과 서비스 품질을 동시에 높였다.

〔그림 3－10〕 SK텔레콤의 저탄소 AI 기술



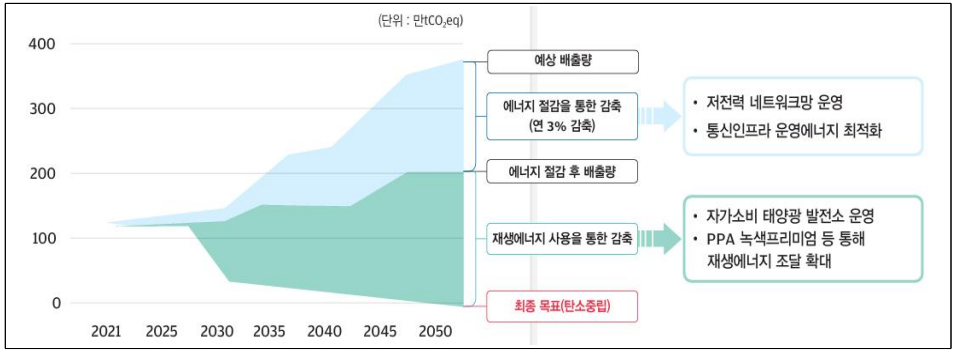
자료: SK텔레콤 Annual Report(2024)

특히 SK텔레콤은 저탄소 AI 기술로의 전환을 주요 과제로 삼고 있다〔그림 3－10〕. 그린 저탄소 ICT 기술을 활용해 네트워크 장비의 전력 사용량을 최소화하고, Single-RAN(다중 네트워크를 하나의 장비에서 운영하는 기술)을 적용해 중복 장비 운용으로 인한 불필요한 전력 낭비를 줄였다. 더 나아가, AI 모델의 연산 과정에서 발생하는 전력 소비까지 고려한 '에너지 효율형 AI 모델'을 연구·개발해 AI 자체의 탄소 배출까지 줄이고 있다. 이러한 기술 개발 과정에서 초기 R&D 비용 증가라는 부담이 존재하지만, 장기적으로는 전력 절감과 탄소 감축 효과가 투자 비용을 상쇄할 것으로 전망된다.

KT 역시 AI를 기반으로 한 에너지 효율화 기술을 다각도로 도입하고 있다. 'AI TEMS (Temperature of Equipment Management System)'를 통해 각 통신실 장비의 특성과 설치 환경, 위치별 온도를 분석하여 최적의 냉방 운전을 자동 수행함으로써 냉방 효율을 24% 개선했다. '에너지 절감 오케스트레이터'를 통해 기지국의 전파 출력을 실시간 트래픽 상황에 맞춰 자동 조정하고, '서버 전력 공급 최적화' 기술을 통해 CPU 부하 수준에 따라 전력 공급량을 조절하여 불필요한 에너지 낭비를 줄이고 있다. 이러한 기술적 노력과 함께 KT는 2021년 온실가스 배출량 대비 2030년까지 51.7%, 2040년까지 75.8% 감축, 2050년 탄소중립 달성을 목표로 하고 있다〔그림 3-11〕. 이를 위해 사용 전력의 100%를 재생에너지로 전환하는 RE100 목표를 추진하며, 탈탄소 재생에너지 전환, 가치사슬 친환경 혁신, 에너지효율성 제고, 기후경영 역량 강화 등 4대 핵심 분야를 선정하고 매년 세부 이행과제를 발굴·추진하고 있다[58].

LG유플러스는 AI 기반 모니터링 시스템을 도입해 네트워크 전력 사용량을 실시간 분석·관리하고 있으며, 이를 ESS(에너지저장시스템) 재가동과 결합해 연간 4,500만 kWh의 전력 절감을 추진하고 있다.

〔그림 3-11〕 KT의 2025 탄소중립 추진 목표



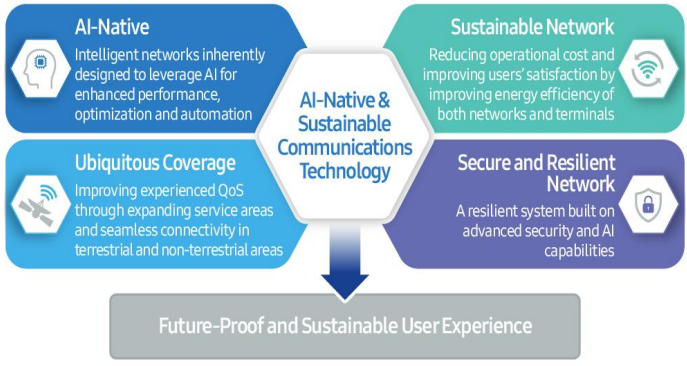
자료: KT ESG REPORT(2024)

한편, 삼성전자는 네트워크 장비의 전력 효율을 극대화하기 위해 AI 기반 에너지 세이빙 매니저 기술인 AI-ESM을 상용망에 적용하고 있다. 이 기술은 트래픽 부하를 인공지능이

58) KT(2024), KT ESG REPORT 2024

예측하여 시간대와 네트워크 부하 수준에 따라 무선 접속망의 전력 사용을 자동으로 최적화하는 방식이다. 특히 5G 기지국 장비에 적용되어, 트래픽이 적은 시간대에는 불필요한 송신 자원을 자동으로 절전 상태로 전환하고, AI가 네트워크 부하를 실시간으로 분석해 최적의 전력 운용 패턴을 찾아낸다. 실제 적용한 결과, 미국 Verizon 네트워크에서는 평균 약 15%의 에너지 절감 효과가 확인되었으며, 일부 구간에서는 최대 35%까지 절감이 가능함을 보였으며[59], 삼성전자는 이를 기반으로 차세대 6G 통신망에서도 'AI-Native & Sustainable Communication'을 핵심 비전으로 삼아 에너지 효율을 획기적으로 높이는 지속 가능한 통신 인프라 기술 개발을 이어가고 있다[60][61]〔그림 3－12〕.

〔그림 3－12〕 삼성전자의 AI-Native & Sustainable Communication



자료: Zero-Trust Architecture for 6G

59) Samsung Networks(2024. 2), "Powering a Greener Future: Inside Samsung's AI-Driven Network Energy Innovation", https://www.samsung.com/global/business/networks/insights/blog/0224-powering-a-greener-future-inside-samsungs-ai-driven-network-energy-innovation/?utm_source=chatgpt.com

60) Samsung Research(2024. 3. 15.), "Zero-Trust Architecture for 6G", https://research.samsung.com/blog/Zero-Trust-Architecture-for-6G

61) Samsung(2024. 2. 24.), "Verizon Infuses AI in Network, Accelerates Open RAN Innovation with Multi-Vendor RAN Intelligent Controller Deployment", https://www.samsung.com/global/business/networks/insights/press-release/0224-verizon-infuses-ai-in-network-accelerates-open-ran-innovation-with-multi-vendor-ran-intelligent-controller-deployment/

또한, 한국전력 KDN은 전력 계통의 통신 인프라 효율화를 위해 AI, 빅데이터, 클라우드 기술들을 융합한 AI 기반 전력 통신망 품질진단 시스템을 구축하였다[62]. 이 시스템은 전국 전력 통신망의 데이터 흐름을 실시간으로 분석하고, 장애 발생 가능성을 AI가 사전에 예측하여 자동 진단 및 조치가 가능한 체계를 구현하였다〔그림 3－13〕. 이를 통해 전력 통신망의 운영 안정성과 관리 효율성을 높였으며, 스마트그리드와 스마트시티 등 차세대 에너지 인프라와의 연계를 통해 통신 효율화와 에너지 절감을 동시에 달성하는 기반을 마련하였다.[63]

〔그림 3－13〕 KDN의 AI 기반 전력 통신망 품질진단 시스템

자료: 한전 KDN

나. 5G/6G 장비의 저전력 설계와 광전송망 고도화

전 세계 통신업계는 국제전기통신연합(ITU)이 제시한 2030년까지 이동통신망 온실가스 45%, 유선망 온실가스 62% 감축 목표를 달성하기 위해, 네트워크 전반의 저전력 설계와 고효율화를 핵심 과제로 삼고 있다[64]. 특히 5G와 향후 6G 네트워크의 상용화는 폭발적으

62) AI Times(2024. 5. 10.), “한전KDN, AI 기반 에너지 결합데이터 분석 플랫폼 개발”, https://www.aitimes.com/news/articleView.html?idxno=160637

63) 한양경제(2024. 5. 22.), “한전KDN, AI 기반 에너지 결합데이터 분석 플랫폼 개발”, https://www.hanyangeconomy.com/article/view/hye202405220001

64) ITU(2010), Recommendation ITU-T L.1470

로 증가하는 데이터 트래픽과 장비 확충으로 인한 전력 소비 부담을 수반하기 때문에, 기술 개발 초기부터 저전력·고효율 설계가 표준화 과정에 포함되고 있다.

첫 번째로, 레거시 네트워크 퇴출이 에너지 절감의 주요 전략 중 하나다. 전 세계적으로 3G 네트워크와 동축 케이블(HFC)망은 여전히 상당한 전력을 소모하고 있어, 이를 폐쇄하거나 고효율 광가입자망(FTTH)으로 교체하는 추세가 가속화되고 있다[65]. 실제로 지난 6년간 22개국 58개 통신사가 3G를 폐쇄했으며, 2024년에도 6개국 15개 사업자가 추가 폐쇄를 진행할 예정이다. 3G 폐쇄만으로도 약 15%의 에너지 절감 효과가, 동축망에서 FTTH로의 전환은 최대 85%의 효율 향상이 가능하다는 분석이 나왔다.

둘째, 광전송망 고도화가 필수다. 광케이블은 전력 소모가 낮고 유지보수가 용이해 장기적으로 탄소 배출을 줄이는 데 효과적이다. 싱가포르, 일본, 호주, 뉴질랜드 등은 국가 단위에서 동축망을 전면 교체하고 있으며, 한국 역시 일부 지역에서 단계적 전환이 진행 중이다.

셋째, 소프트웨어 정의 네트워크(SDN)다. 전통적인 네트워크가 라우터·스위치와 같은 하드웨어 장치로 제어되는 반면, SDN은 네트워크 자원을 가상화하여 중앙에서 통합적으로 제어·관리한다. 이를 통해 장치별 수동 제어를 없애고, 네트워크의 유연성과 자동화 수준을 크게 향상시킨다[66]. SDN 시장은 전 세계적으로 빠르게 성장 중이며, 2023년 28억 달러 규모에서 2032년까지 연평균 17% 성장할 것으로 전망된다[67]. 이러한 성장세는 단순한 네트워크 성능 개선을 넘어, 저전력 네트워크 구현과 지속가능한 통신 인프라 구축에 기여하는 중요한 요소로 작용한다.

또한 국내에서도 SK텔레콤은 3G와 LTE 장비를 하나로 통합 운영하는 싱글 RAN 기술을 상용화해 불필요한 중복 장비 운용을 줄이고, 연간 약 1만 톤의 탄소 배출을 절감하고 있다. KT는 경량형 5G 업무망을 개발해 기존 대비 인프라 구축 비용을 60% 절감했으며, 추가 장비 설치 없이 보안성과 망 효율화를 동시에 달성했다.

---

65) Deloitte Insights(2024), Dialing down the carbon: Telco sustainability surges on the back of four new trends

66) 한국전자통신연구원(2014),『스마트인터넷을 위한 SDN 및 NFV 표준기술 동향 분석』.

67) Global Market Insights(2024), Software Defined Networking Market

다. 정부 주도 에너지 절감 기술 개발 및 표준화 정책

과학기술정보통신부는 '디지털 탄소중립 민관협의회'를 확대 개편해 연 3회 이상 개최하며, 통신 3사·장비 제조사·AI 반도체 기업·학계·연구기관과 협력하여 기술·표준화 전략을 마련하고 있다.

첫 번째로, 기지국 저전력화를 핵심 의제로 설정했다. 현재 통신사 탄소 배출의 70% 이상이 네트워크 장비, 특히 기지국에서 발생한다. 정부는 이를 줄이기 위해 저전력 AI 반도체 탑재, 고효율 전력증폭기 소자 개발, 수동형 빔포밍 부품 경량화, 중계기 저전력 설계 등을 중점 지원하고 있다[68].

둘째, 차세대 이동통신(6G) 표준화 과정에서의 주도권 확보다. 국제이동통신표준화기구(3GPP)는 지난 14일 인천에서 열린 6G 워크숍·기술총회에서 6G 무선망 기술 연구과제를 승인했으며, 이번에 도출된 6G의 핵심 목표를 지속가능성, 효율성, 상호운용성으로 제시했다. 5G가 속도와 용량 등 기술 성능에 초점을 맞췄다면, 6G는 성능뿐 아니라 사용자 경험, 운용 효율, 구현 가능성까지 고려한 서비스 친화적 네트워크 설계와 인프라 가치 제고에 중점을 두고 있다. 이 과정에서 한국은 AI 내재화 네트워크, 에너지 절감 기능, 지상·위성 연동(비지상망, NTN)을 핵심 연구 항목으로 채택하는 데 주도적인 역할을 했다. 특히 SK 텔레콤, KT, LG유플러스 등 국내 통신사는 무선망 설계 시 6G와 AI 기술의 통합을 강조하며 6G AI·ML 프레임워크를 이끌었고, 3GPP는 이를 AI4NET(네트워크 최적화를 위한 AI)과 NET4AI(AI 서비스 지원을 위한 네트워크)로 구분해 추진하기로 했다〔그림 3-14〕.

68) 전자신문(2024. 4. 15). "정부·통신·장비사, 네트워크 저전력화 머리 맞댄다", https://www.etnews.com/20240415000319

〔그림 3－14〕 3GPP 6G 표준화 타임라인

자료: 한국정보통신기술협회(TTA)(2023)

이러한 논의는 6G 기지국의 '항상 켜져 있는 신호' 최소화, 기지국·단말·망 간 '딥슬립 모드' 확대 등 초기 설계 단계에서부터 에너지 효율성을 내재화하는 기반이 됐다. 생성형 AI 발전으로 통신 인프라에 더 높은 성능과 다양한 활용성이 요구되는 상황에서, 국내 통신사들은 AI를 결합한 네트워크를 통해 맞춤형 AI 서비스 지원이 가능한 차세대 인프라 아키텍처를 제시함으로써 6G 표준화에서 경쟁 우위를 확보했다[69].

69) 전자신문(2025. 3. 16.), "韓, AI 내재화 통신 연구 주도…6G 표준화 앞장선다", https://www.etnews.com/20250316000024

제 2 절 탄소 감축을 위한 디지털 기술 사례 및 대응 방안

1. AI 기반 제조 최적화

AI 기반 제조 최적화는 제조 공정 전반에 인공지능과 사물인터넷을 통합하여 설비 운영을 정밀하게 제어하고, 에너지 낭비를 최소화하는 혁신적인 접근법이다. 이러한 시스템은 센서와 IoT 기기를 통해 수집된 방대한 현장 데이터를 AI 분석 엔진에서 실시간으로 처리하여, 불량률 감소와 공정 스케줄 효율화를 동시에 달성한다 〈표 3-2〉. 이를 통해 원재료, 전력, 인력 등 자원의 사용량을 절감하며, 제품 1단위 생산 시 발생하는 에너지 소비와 탄소 배출량을 획기적으로 줄인다. 이 과정에서 단순한 비용 절감뿐 아니라 생산성 향상, 품질 안정성 확보, 예측 유지보수 기반의 설비 고장 예방 등 다양한 부가가치를 창출한다. 예측 유지보수는 설비의 진동, 온도, 전력 소모 패턴 등을 분석해 고장을 사전에 감지하고, 계획된 정비를 가능하게 하여 예기치 못한 가동 중단 시간을 줄인다. 이는 특히 에너지 집약적인 산업에서 큰 효과를 발휘하며, 생산 효율성을 극대화하는 동시에 탄소 감축 목표 달성에 직접 기여한다. 또한 머신러닝 기반 공정 스케줄링은 시장 수요, 재고 상황, 설비 가동률 등 복합적인 요인을 반영해 생산 계획을 실시간으로 조정함으로써 불필요한 대기 시간과 에너지 소모를 최소화한다. 이러한 AI 기반 제조 최적화는 정부의 스마트공장 보급 확산 정책, 탄소중립형 제조혁신 지원정책과 긴밀히 연계되어 산업 전반의 저탄소 전환을 가속화하고 있다. 궁극적으로는 제조 현장의 디지털 전환을 촉진하여, 에너지 효율성과 친환경성을 동시에 강화하는 지속가능한 생산 체계를 구현한다[70].

70) 정보통신정책연구원(2025), AI 기반 소프트웨어 정의 공장을 활용한 제조 혁신

〈표 3－2〉 AI 기반 제조 최적화의 주요 특징

구분	주요 특징	내용
1	디지털 트윈 기반 공정 최적화	생산 시설의 가상 모델을 생성해 물리적 프로토타입 전에 비효율성을 식별하고 최적화.
2	AI 기반 예측 유지보수	IoT 센서 데이터와 머신러닝 모델로 설비 상태를 실시간 분석해 고장을 사전 예측.
3	에너지 자원 최적화 제어	AI 제어 시스템이 가용 자원·수요·설비 상황을 종합 분석해 생산 및 에너지 사용 자동 최적화.
4	스마트 품질 관리	컴퓨터 비전·머신러닝을 활용한 실시간 불량품 검출 및 품질 예측. 결함률 감소, 품질 안정성 확보.
5	통합 데이터 분석, 운영 플랫폼	수만 개 센서 데이터를 통합·분석하여 공정 모니터링, 에너지 관리, 생산 스케줄링까지 지원.

자료: AI 기반 소프트웨어 정의 공장을 활용한 제조 혁신(2025)

가. 국외 기술 동향

1) Siemens

Siemens는 디지털 트윈 기술과 머신러닝 기반 예측 분석 및 Plant Simulation 도구를 결합하여 제조 공정을 최적화한다. 독일 Erlangen 전자공장에서 IoT 센서를 통해 실시간 데이터를 수집하고 AI 알고리즘으로 설비 효율을 극대화한 결과, 생산성이 69% 향상되고 에너지 소비는 42% 절감되었다. 구체적으로, 디지털 트윈은 생산 시설의 가상 모델을 생성하여 물리적 프로토타이핑 전에 비효율성을 식별하고 최적화하며, AI는 실시간 데이터 분석을 통해 시뮬레이션 시나리오를 실행하여 생산 처리량과 품질을 최적화한다[71]. 이를 통해 환기 시스템의 에너지 사용량은 70% 감소, 물류 공간 사용량은 50% 절감, 소재 순환량은 40% 감소하였다. 이러한 에너지 관리 혁신으로 탄소 배출량이 대폭 줄어들었으며, Siemens는 예측 유지보수와 공정 스케줄링을 통해 설비 가동률을 높이고 재생에너지와의 통합으로 친환경 경쟁력을 강화했다. 특히, 공장 시설의 포괄적인 디지털 트윈 내에서 생산 데이터를 기반으로 훈련된 AI는 지속적인 모니터링을 통해 수동 운영 대비 높은 수준의 에너지 최적화를 달성한다[72].

71) Siemens(2024), World Economic Forum: Siemens factory in Erlangen named Digital Lighthouse Factory

72) Siemens(2024), Getting the best out of production with the Digital Twin

2) Bosch

Bosch는 AI와 IoT를 결합한 스마트 제조 시스템을 적극 도입하여, 제조 공정 효율성 및 에너지 절감 성과를 크게 향상시켰다. 독일 Feuerbach 공장에서는 약 550개의 설비를 IoT로 연결하고 스마트 연결성을 활용한 데이터 기반 운영을 통해, 에너지 요구량을 2007년 대비 50% 이상 줄이는 성과를 달성했다. 또한, 사이클 타임은 약 10% 단축되었으며, 관리자 업무 부담(행정 오버헤드)는 50% 이상 감소했다[73]. 이에 더해 Blaichach 공장은 Bosch Connected Industry의 Nexeed Industrial Application System을 통해 한 차원 높은 자동화와 데이터를 활용한 운영을 구현하고 있다. Nexeed는 60,000개 이상의 센서를 통해 수집된 실시간 데이터를 기반으로 예측 유지보수(Predictive Maintenance)를 가능하게 함으로써, 제작 공정 중 불시적 다운타임을 약 25% 감소시켰다[74]. 또한, Nexeed는 공장 내 물류 최적화, 대기 재고 절감, 운송 경로 효율화, 리드타임 단축 등의 기능을 통해 전체적인 생산성 향상을 지원한다.

3) Toyota

Toyota는 미국 켄터키주 조지타운 공장에서 AI 기반 공정 최적화를 도입하여, IoT 센서를 통해 설비 상태를 실시간 모니터링하고 AI로 에너지 사용을 최적화하여 연간 3,500톤의 탄소 배출을 감축한다. Toyota는 불량률을 12% 줄이고 생산성을 18% 향상시켰으며, 미국 정부의 친환경 제조 지원 프로그램과 연계하여 지속 가능성을 강화한다. 구체적으로, 첨단 페인트 시설에 9억 2,200만 달러를 투자하여 100만 제곱피트 시설을 추가하여 탄소 배출을 30% 감소시키며 전기차 생산을 지원한다. AI 플랫폼을 통해 계획, 실행, 점검, 개선(Plan-Do-Check-Act, PDCA) 사이클을 가속화하여 수동 작업을 연간 10,000시간 절감한다. 또한, 생성형 AI를 활용하여 디지털 매뉴얼을 생성, 물리적 매뉴얼 생산의 재료를 줄여 환경 영향을 최소화하며, 대규모 언어 모델(Large Language Model, LLM)과 멀티모달 학습을 통해 매뉴얼의 텍스트와 이미지를 분석, 음성 인터페이스로 복잡한 쿼리에 답변하여 600페이지 이상의 종이 매뉴얼을 대체한다. 또한, AI 기반 예측 유지보수로 장비 고장을 사전 예측하여 다운타임을 최소화하고, 컴퓨터 비전으로 결함 검출 정확도를 92-95%로 높여 품

73) Bosch(2020), The Bosch plant in Feuerbach – where tradition meets high-tech

74) Bosch(2025), Nexeed — welcome to the smart factory

질을 강화하며, IoT 센서 데이터를 활용한 실시간 분석으로 에너지 소비를 최적화, 공급망 효율성을 높여 전체 생산 비용을 절감한다.[75]

4) Rockwell Automation

Rockwell Automation은 수처리, 제조 영역에서 AI와 실시간 데이터 분석을 활용해 에너지 사용과 운영비용을 동시에 줄이고 품질과 안정성을 높이는 접근을 적용한다. 수처리 시설의 경우 AI와 실시간 데이터 분석 기술을 미국 캘리포니아주의 동부 시립 수자원 지구 폐수처리장에서의 aeration 공정 최적화 및 제조현장 에너지 관리, 예지보전 체계로 확장한다. 구체적으로, 폐수에 공기를 불어 용존산소를 확보하는 과정에서 AI가 유입수의 양과 오염 정도를 학습해, 그때그때 필요한 산소량을 예측하고 송풍량을 자동 조절하는 방식을 적용한다. 결과적으로, Rockwell Automation은 수처리 시스템에 AI를 도입하면서 에너지 소비를 30-50%절감하는 동시에 공정 효율을 향상시킨다[76]. 제조 공정에서 Rockwell은 싱가포르의 아시아태평양 제조 거점에 에너지 사용을 실시간으로 감시할 수 있는 대시보드와 데이터 기반 운영 체계를 구축해, 라인, 설비별 전력, 압축공기 사용을 비교하고 누설, 유휴설비 등 낭비를 조기에 찾아내는 방식을 적용한다[77].

나. 국내 기술 동향

1) 삼표시멘트

삼표시멘트는 산업통상자원부의 AI 자율제조 선도 프로젝트를 통해 시멘트 제조 전 공정에 AI 기술을 접목하여 디지털 전환을 가속화하고 있다. 공장 내 100여 개의 센서와 IoT 장비에서 10초 단위로 수집되는 데이터를 AI가 실시간 분석해 원료 혼합비, 소성로 온도, 연료 투입량 등을 자동 최적화한다. 기존에 숙련자의 경험에 의존하던 클링커 원료 배합과 가마 온도 조정이 AI 기반으로 전환되면서 품질 편차가 줄고, 원료 낭비와 불필요한 에너지 사용이 동시에 감소했다. 특히 소성 과정에서 유연탄 투입을 최소화하고 폐기물·바이오매스 등 순환연료 사용 비중을 2023년 34%에서 2030년 58%까지 확대할 계획이다. 이

75) DigitalDefynd(2025), 5 Ways Toyota is Using AI [Case Study]

76) Rockwell Automation(2025), How AI Boosts Energy Efficiency in Wastewater Treatment

77) Rockwell Automation(2025), Rockwell Automation AI's Power for Sustainable Manufacturing

를 통해 2018년 대비 2030년에는 탄소 배출량을 21%, 2050년에는 54% 감축할 것으로 전망된다[78]. 또한 AI가 설비 진동, 온도 데이터를 분석해 고장을 사전에 예측하는 예측 유지보수를 구현함으로써 가동 중단 시간을 줄이고 에너지 낭비를 방지한다. 이와 함께 폐열회수 발전소를 운영하여 연간 6.6MWh의 전력을 자체 생산, 전체 전력 사용량의 8.3%를 대체하고 약 3만 톤의 탄소 배출량을 감축했다. 아울러 화력발전소 석탄재와 제철소 슬래그를 원료로 재활용해 저탄소 시멘트를 생산, 일반 콘크리트 대비 이산화탄소 배출량을 절반 수준으로 줄였다. 이처럼 삼표시멘트의 AI 기반 공정 최적화는 제조 효율 향상과 에너지 절감, 연료 전환을 결합하여 시멘트 산업의 탄소중립 실현과 친환경 경쟁력 강화를 동시에 달성하고 있다.

2) 효성 중공업

효성중공업은 전력 설비와 제조 공정의 디지털 전환을 강화하기 위해 AI 기반 지능형 자산관리 시스템 ARMOUR를 도입했다. 이 시스템은 10,000건 이상의 고장 데이터를 학습해 상태 예측 성능 95%를 달성했으며[79], 수명 예측 모델의 예측 정확도는 96.5%, 기기 건전도 평가의 예측 정확도는 98.6%를 기록했다[80]. 이러한 높은 예측 정확도는 전력설비의 상태를 실시간으로 평가하고, 이상 징후를 조기 감지해 사전에 유지보수 계획을 수립하게 함으로써, 불필요한 정지 시간을 줄이고 전체 설비 안정성을 높인다. AI는 IoT 센서로부터 수집되는 진동, 온도, 전류 등의 데이터를 실시간 분석하며, 머신러닝 모델이 축적된 데이터를 기반으로 이상 패턴을 식별한다. 이 과정에서 예측 유지보수 체계로의 전환이 가능해지고, 고장이 발생하기 전에 필요한 부품 교체나 정비를 수행할 수 있다. 결과적으로 효성중공업은 설비 가동률 94%를 유지하고 있으며, 대형 팬 제어 방식의 고압 인버터 도입과 AI 기반 운전 최적화를 통해 연간 전력 소비를 8~12% 절감하고, 탄소 배출량을 약 1,800톤 감축하는 효과를 거두고 있다. 환경적 측면에서 효성중공업은 2018년 온실가스 배출량을 기준으로, 2030년까지 배출량을 14.5% 감축하는 중장기 목표를 수립했으며[81], 이는

78) 삼표(2023), 삼표시멘트 ESG 보고서

79) 효성중공업(2021), Power & Motion Magazine Vol.1

80) 효성중공업(2022), Power & Motion Magazine Vol.2

81) 효성중공업(2025), https://www.hyosungheavyindustries.com/en/sustainability/environmental/climate-change

AI 예측분석 기술과 스마트 에너지 관리 플랫폼의 통합 운용을 통해 실현되고 있다. 예측분석 결과는 공장의 에너지관리시스템과 연계되어 설비 운영 스케줄을 최적화하고, 재생에너지의 효율적 통합 운용을 지원한다. 또한 효성중공업은 한국전력공사와 공동 개발한 예방진단 시스템 'SEDA'를 ARMOUR와 결합해 발전소, 데이터센터, 철도, 빌딩 등 다양한 산업군으로 확장 적용하고 있다. 증강현실(AR) 기반 원격 모니터링 기술은 현장 작업자의 안전성을 높이고, 즉각적인 대응이 가능하도록 지원한다. 이러한 성과는 AI 기술과 첨단 전력 설비가 결합될 때, 산업 전반의 에너지 효율성과 탄소 중립 달성 가능성을 크게 높일 수 있음을 보여주는 대표적인 사례다.

3) 현대자동차

현대자동차는 지속가능한 제조를 위한 AI 기반 스마트공장 전략과 재생에너지 도입을 병행해 탄소저감과 생산 효율화를 동시에 추구하고 있다. 먼저, 도장 공정 혁신을 대상으로 기존 140℃에서 20분 동안 경화하던 상도 공정을 90℃ 20분으로 낮추는 저온 경화 기술을 개발했다. 이를 통해 도장 공정에서 약 40%의 에너지 소비와 탄소 배출 저감이 가능하며, 글로벌 모든 공장으로 확장 적용 시 연간 약 16,000톤의 탄소 배출량 감축 효과가 기대된다[82]. 도장 공정은 전체 제조 공정 중 약 43%의 에너지를 소모하는 핵심 구간이라는 점에서, 이 혁신의 효과는 매우 크다. 아울러 현대자동차는 2045년까지 전력 100% 재생에너지 전환인 RE100 달성을 목표로 설정했으며, 이를 위한 전략적 조치로 2024년 국내 최대 규모의 재생에너지 PPA 계약을 체결했다. 해당 계약을 통해 향후 20년간 매년 610 GWh 규모의 재생전력이 공급되며, 이를 통해 연간 약 280,000톤의 탄소 배출량 감축이 예상된다. 또한 울산, 아산, 전주 공장에 총 15 MW 규모의 태양광 발전 설비를 구축하는 등, 자가 발전 기반 확대도 지속적으로 추진 중이다[83]. 이처럼 현대자동차는 AI 기반 기술 혁신(저온 경화 도장 기술)과 대규모 재생에너지 확보(PPA 계약 및 태양광 설비 등)를 동시에 추진함으로써, 제조업 현장의 에너지 소비 구조와 탄소 배출 패러다임 자체를 재설계하고 있다.

---

82) 현대자동차그룹 '룸(2025), https://www.hyundai.co.kr/news/CONT0000000000107747

83) 현대자동차(2025), 2025 현대자동차 지속가능성 보고서 https://www.hyundai.com/kr/ko/sustain-manage/manage-system/sustainability-report

2. 스마트 에너지 관리

스마트 에너지 관리는 전력망과 수요 측 모두에서 IoT와 AI 기반의 지능형 전력망(스마트그리드) 기술을 도입하여 실시간으로 전력 수요를 관리하고, 태양광·풍력 등 재생에너지와 에너지저장장치(ESS), 전기차 충·방전 등 다양한 분산자원의 효율적 활용을 가능하게 한다〈표 3-3〉. 건물과 공장에는 에너지관리시스템(EMS)을 구축해 조명·공조·설비의 운전을 자동 제어하고, 수요예측과 부하최적화 알고리즘을 통해 불필요한 에너지 사용을 줄인다. 이를 통해 피크전력 발생을 억제하고 재생에너지 비중을 높여 전력 생산 부문에서의 탄소배출을 저감하며, 전력요금 절감과 안정적인 전력 수급을 동시에 달성한다. 나아가 이러한 시스템은 정부의 스마트그리드 보급 확대, 에너지효율 혁신, 탄소중립 달성 정책과 연계되어 산업·가정·공공부문 전반의 에너지 효율성과 지속가능성을 강화한다.[84]

〈표 3-3〉 스마트 에너지 관리 주요 특징

구분	주요 특징	내용
1	지능형 전력망 운영	IoT, AI 기반 스마트그리드 기술로 전력망 및 수요를 실시간 관리.
2	에너지 관리 시스템 자동 제어	건물, 공장 내 EMS로 조명, 공조, 설비 제어, 수요예측 및 부하 최적화 적용.
3	부하 최적화, 피크 전력 억제	전력 수요 패턴 분석을 통한 피크 억제 및 전력요금 절감.
4	재생 에너지 통합	재생에너지 확대 및 분산자원 연계 운영으로 탄소배출 저감.
5	에너지 효율 지속 가능성 강화	정부 정책과 연계해 에너지 효율성 및 지속가능성 제고.

자료: 스마트 그리드(에너지) 분야 도입사례 분석집(2017)

가. 국외 기술 동향

1) 미국 Google DeepMind

Google DeepMind는 데이터센터 냉각 최적화를 위해 수천 개 센서에서 수집한 실시간

84) 정보통신산업진흥원(2017), 스마트 그리드(에너지) 분야 도입사례 분석집

온도, 압력, 유량 데이터를 딥러닝 예측 모델과 강화학습 기반 제어 로직에 투입해 냉각 시스템의 세트포인트를 자동 조정한다. 모델은 약 60분 선행의 열부하와 전력 사용 효율 변화를 예측하고, 상한온도, 압력, 장비 보호 규칙의 안전 제약 하에서 제어 방법을 생성해 빌딩 관리 시스템에 반영한다. 해당 자율형 에너지 관리 시스템은 부하 변동과 외기 조건을 따라가며 스스로 최적 운전하여 냉각 에너지 최대 40%와 약 15%의 전력 사용 효율을 증대시켰다. 그 결과 데이터센터 전력, 탄소 집약도가 높은 피크 시간대의 사용 전력을 낮추며 연간 수천 MWh 수준의 전력 절감과 의미 있는 탄소 배출 감축을 달성한다.[85]

2) 영국 National Grid

영국 National Grid의 전력 시스템 운영기관은 겨울철 피크 대응 프로그램인 수요 유연성 서비스를 운영하면서, 가정·사업장의 30분 단위 사용량이 기록, 전송되는 전력계인 스마트미터와 과거 패턴을 바탕으로 AI가 고객군의 참여 가능성과 이동 가능한 전력량을 예측, 세분화하고, 피크 시간대에는 전기차 충전, 히트펌프, 가정용 배터리 같은 설비의 가동 시점, 출력, 기간을 자동으로 재배치해 부하를 옮긴다. 2022년과 2023 겨울 동안 31개 사업자와 160만 가구 및 사업기관이 약 3,300MWh의 피크 전력 사용량 감축을 기록했다. 이는 석탄 등 고탄소 피크 발전기의 기동을 줄여 약 681톤의 탄소 배출 감축을 실현했다. 이러한 구조는 개별 가정, 사업장의 설비를 교체하지 않고도 소프트웨어 중심 운영을 가능케 하여 확장성을 높인다[86].

3) 캐나다 BrainBox AI

BrainBox AI는 단순한 온·습도 유지 제어를 넘어, 예측 기반 자율 최적화(Predictive Autonomous Optimization) 알고리즘을 적용한다. 해당 알고리즘은 외기 조건 변화, 일기예보, 건물 점유 패턴의 변화를 사전에 분석하여 냉난방 부하를 선제적으로 조정함으로써 전력 피크 억제와 불필요한 설비 가동 최소화를 동시에 달성한다. 이를 통해 운영 효율이 향상될 뿐 아니라, 설비의 부분부하 운전 최적화로 인한 기계적 마모 저감과 유지보수 주기 연장 효과도 함께 얻는다. 또한, 클라우드 기반 학습 구조를 통해 다수의 건물 데이터

85) WIRED(2018), Google's DeepMind trains AI to cut its energy bills by 40%

86) National Energy System Operator(NESO)(2023), Demand Flexibility Service delivers electricity to power 10 million households.

를 통합 분석함으로써, 개별 건물에서 발생하는 제어 경험이 전체 네트워크로 확산된다. 이로써 신속한 성능 개선과 지역·기후 특성에 맞춘 제어 모델 정밀화가 가능해지며, 도입 이후에도 지속적인 효율 향상이 가능하다. 실제 45 Broadway 적용 사례에서는 설비 교체 없이 평균 15.8%의 전력 절감과 연간 37.1톤의 탄소 감축이 확인되었으며, 적용 초기부터 단기간 내 성과가 가시화되어 확산 가능성이 높다는 점이 입증되었다.[87]

나. 국내 기술 동향

1) KT AI 빌딩 오퍼레이터

KT AI 빌딩 오퍼레이터는 산업통상자원부 국가기술표준원의 신기술 인증을 받은 상용 솔루션으로, 건물 설비를 교체하지 않고도 에너지 운영을 최적화 → 지시 → 검증의 폐루프로 자동화한다. 해당 설비는 총 3가지의 요소로 구성된다. 첫째, 빌딩설비 제어엔진인 Robo-Operator는 건물별 디지털 트윈 시뮬레이션으로 생성한 가상 운전 데이터와 실운전 데이터를 함께 학습해 냉난방, 공조의 세트포인트와 가동 순서를 산출한다. 둘째, 지능형 컨트롤러인 i-Box는 다양한 표준 프로토콜로 온/습도, 유량, 차압, 밸브 개도 등의 센서 및 계측 데이터를 수집하고, 상·하위 시스템 사이에서 안전 제약을 지키며 제어 명령을 제어한다. 셋째, 온라인 상으로 접속해 1분 단위 환경, 설비 데이터를 수용, 시각화하고, 분석 결과를 운영 화면과 알림으로 실시간 반영할 수 있는 클라우드 기반 서비스 플랫폼으로 구성된다. 이 구조로 인해 중앙의 AI가 실시간 운영조건과 단기 예보를 반영해 냉동기 출수온도, 펌프 회전수, 공조기 외기비, 정압 같은 세트포인트를 선제적으로 조정하고, 현장에서는 i-Box가 다양한 산업 표준 프로토콜을 통해 기존 BMS/제어기에 안전하게 지시를 내린다. 이러한 구성은 기존 BMS 위에서 상위 최적화로 동작하기 때문에 기존 건물 어디서나 빠르게 적용 가능하다. 그 결과 12개 건물 실증에서 10 ~ 20% 절감을 실현한다[88].

2) 현대자동차

현대자동차 아산공장은 공장 전역의 열·전력 흐름을 하나의 관제 체계로 통합하는 에너지관리시스템을 중심에 두고, 설비별 상태, 부하, 소비 데이터를 세분 계측해 운영 알고

87) TIME(2024), How AI Is Making Buildings More Energy-Efficient.

88) KT(2025), KT ESG Report 2024

리즘을 재설계했다. 장거리 스팀 배관에서 발생하던 방열, 블로우다운 손실을 줄이기 위해 중앙 집중형 보일러 의존도를 낮추고 현장 근처에 분산 보일러를 배치했으며, 급탕에는 기동성이 뛰어나 부분부하 효율이 높은 이온히팅 시스템을 적용해 필요한 순간에 필요한 만큼만 열원을 공급하도록 했다. 동시에 도장·압축공기·공조 등에서 버려지던 폐열은 폐열회수 장치로 회수해 예열·난방에 재투입했고, 에너지관리시스템에서는 부하 예측에 따라 보일러 기동/정지와 온도, 압력 셋포인트를 자동 최적화했다. 알람 임계치와 표준운전 절차를 도입해 비정상 운전을 조기에 차단한다. 그 결과 부분부하 비효율과 피크가 완화되어 연간 2,456 톤의 탄소 배출을 줄이고 에너지비 약 14.5억 원을 절감했으며, 가동 안정성과 유지보수 주기의 예측 가능성까지 함께 높인다.[89]

3) LG전자

LG전자 가산 R&D 캠퍼스는 ISO 50001 기반의 에너지경영시스템(EnMS)을 구축하고 이를 건물에너지관리시스템(BEMS)과 결합해 운영 효율을 극대화한 대표적 사례다. 이 캠퍼스는 연구동, 사무동, 실험동 등 상이한 용도와 부하 특성을 가진 건물들을 단일 모니터링 플랫폼에 통합하여, 에너지 흐름과 소비 패턴을 실시간으로 가시화한다. 특히 층별·용도별 서브미터링 체계를 통해 누수 부하를 조기에 식별하고, 부하 특성에 따라 맞춤형 절감 전략을 적용할 수 있도록 했다. 운영 측면에서는 실운영 데이터와 기상, 계절, 점유도 정보를 종합 분석해 냉동기, 냉각탑, 공조기, 조명 등 핵심 설비의 스케줄과 온도, 풍량 셋포인트를 정교하게 조정했다. 주말과 야간에는 비핵심 부하를 자동 차단하고, 외기 조건이 유리할 경우 프리쿨링/프리히팅을 활성화하여 기계 냉, 난방 의존도를 줄였다. 실험동의 경우, 환기 및 배기 수요가 특히 크기 때문에 환기율과 재열부하를 분리 관리하고, 공간 점유 패턴을 기반으로 부분 가동을 확대해 불필요한 에너지 낭비를 줄였다. 또한 주기적인 기준선 갱신과 설비 성능 모니터링을 통해 성능 저하 신호를 조기 감지하여 레트로커미셔닝을 시행함으로써 장기적인 효율 유지와 성능 복원을 동시에 달성했다. 이러한 통합 관리 전략의 결과, 캠퍼스는 연간 14,386 GJ의 에너지를 절감했으며, 이는 약 340.6톤의 탄소 배출량 감축 효과로 이어졌다. 더불어, 피크 수요 억제와 전력요금 구조 최적화로 운영비

89) 현대자동차(2025), 2025 현대자동차 지속가능성 보고서
https://www.hyundai.com/kr/ko/sustain-manage/manage-system/sustainability-report

절감 효과를 거두었고, 장비 과부하 감소로 유지보수 비용 절감과 다운타임 위험 완화라는 부가적 성과도 확보했다.[90]

4) 대한민국 수요자원거래시장

대한민국 수요자원거래시장은 국가 전력계통의 피크 부하를 안정적으로 관리하고, 화석연료 발전기의 가동을 최소화하기 위해 상시 운영되는 전력 수요관리 인프라이다. 산업체와 상업용 건물의 전력 부하는 스마트미터와 원격제어 게이트웨이로 실시간 연결되어, 전력거래소의 지시에 따라 즉시 감축이 가능하다. 평상시에는 각 사업장의 에너지관리시스템 및 건물에너지관리시스템이 설비별 부하, 운영상태, 환경데이터를 지속적으로 모니터링하며, 사전에 설정된 감축 시나리오에 따라 조명, 공조, 압축공기, 비핵심 공정 설비의 가동을 단계적으로 줄인다. 운영 방식은 공조 설비의 운전 시간대 전환, 압축공기 시스템의 압력 단계 하향, 냉, 열 축냉, 축열 설비를 활용한 부하 이동 등이 자동 제어로 이뤄진다. 이를 통해 생산과 서비스 품질을 유지한 채 부하를 줄일 수 있으며, 전력 계통의 안정성과 경제성을 동시에 확보한다. 현재 수요자원거래시장은 전국적으로 약 4 GW의 수요자원을 상시 확보하고 있으며, 연간 감축 전력량은 175,771 MWh에 달한다. 이는 초기 513 MWh 수준에서 약 343배 확대된 수치로, 고탄소 피크 발전기의 기동을 줄여 전력 배출계수에 비례한 상당한 온실가스 감축 효과를 실현하고 있다. 이러한 시스템은 대규모 신규 설비 투자 없이 ICT 기반 운영 최적화를 통해 신속하게 구축 및 확장할 수 있다[91].

3. 스마트 모빌리티 및 물류 최적화

스마트 모빌리티 및 물류 최적화는 AI 기반 경로 최적화 알고리즘과 차량공유 플랫폼을 활용해 운송 차량의 불필요한 이동을 줄이고 공차율을 최소화함으로써 연료 사용과 온실가스 배출을 동시에 감축한다〈표 3－4〉. 지능형 교통체계(ITS)는 교통 신호를 실시간으로 최적 제어하여 차량 정체와 공회전 시간을 줄이고, 도심 내 이동 효율성을 높인다.[92] 또한

90) Clean Energy Ministerial(2022), Global Energy Management System Implementation: Case Study – LG Electronics Gasan R&D Campus,

91) Ko, Wonsuk, et al. “Implementation of a demand-side management solution for South Korea’s demand response program.” applied sciences 10.5(2020): 1751.

물류 창고와 배송망 전반에 걸쳐 데이터를 기반으로 수요를 예측하고 계획을 수립해 운송 회수를 최소화하며, 화물 적재율과 운송 경로 효율을 극대화한다. 이러한 최적화는 연료비 절감과 물류 운영비 절감 효과를 제공할 뿐 아니라 교통 부문 탄소배출 저감에 직접 기여한다. 나아가 정부의 스마트시티 구축, ITS 인프라 확충, 친환경 물류체계 추진전략과 연계해 전국적인 모빌리티 효율화와 탄소중립형 물류 생태계 조성을 촉진한다.[93]

〈표 3－4〉 스마트 모빌리티 및 물류 최적화 주요 특징

구분	주요 특징	내용
1	경로 최적화 및 공차율 최소화	AI 기반 경로 계획 알고리즘을 활용하여 불필요한 운행을 줄이고 차량 공차율을 최소화함.
2	지능형 교통체계	실시간 교통 신호 제어, 교통량 분석, 혼잡 예측 등을 통해 정체 해소 및 공회전 시간 감소.
3	물류 네트워크 효율화	수요 예측 기반 물류 계획 수립으로 운송 회수를 최소화하고 적재율 극대화.
4	에너지 절감 및 배출 저감	연료 사용량 절감과 온실가스 감축 효과를 동시에 실현.
5	정부정책 연계	스마트시티 구축, ITS 인프라 확충, 친환경 물류체계 추진전략 등과 연계하여 전국적 효율화 촉진.

자료: 스마트모빌리티 서비스의 현황 및 발전방안 연구(2020)

가. 국외 기술 동향

1) 미국 SURTRAC

미국 피츠버그시는 도심 교차로 혼잡과 차량 정체로 인한 탄소 배출을 줄이기 위해, 인공지능 기반의 분산형 신호제어 시스템 SURTRAC을 도입했다. 이 시스템은 각 교차로에 소형 컴퓨터를 설치해, 도로 위의 차량 흐름을 감지하는 카메라, 레이더, 루프 검지기 등의 센서로부터 실시간 데이터를 받아 매 초마다 신호를 재계산한다. 교차로는 서로 연결되어 있어, 한 교차로에서 예상되는 차량 흐름 정보를 이웃 교차로로 전달하고, 이를 토대로 초록불이 이어지는 그린웨이브 구간을 최대한 늘려준다. 시스템은 교차로에 도착할 차량 무

92) 경기연구원(2020), 스마트모빌리티 서비스의 현황 및 발전방안 연구

93) 한국행정연구원(2021), 스마트모빌리티 서비스 활성화를 위한 OPL 결과보고서

리를 하나의 단위로 묶어, 대기 시간을 줄일 수 있는 순서로 신호를 바꾸며, 보행 신호나 최소, 최대 초록불 시간, 노란불과 같은 안전 규칙은 반드시 지킨다. 기존의 중앙 집중식 신호 시스템이 수 분 단위로 계획을 바꾸는 것과 달리, SURTRAC은 1초 단위로 대응해 사고나 돌발 상황에도 빠르게 반응할 수 있다. 피츠버그 이스트리버티 지역의 9개 교차로에서 시험한 결과, 평균 통행시간이 약 26% 줄었으며, 차량의 정지, 대기 시간이 40% 이상 감소했다. 결과적으로 하루 약 247갤런의 연료와 2.25톤의 탄소 배출 감축 효과를 거두었으며, 연간으로는 약 588톤의 탄소 절감이 가능하다. 이러한 성과는 교차로 정체와 불필요한 공회전 시간을 줄여 연료 절감과 배출 저감을 동시에 달성한 사례이다[94].

2) 호주 SCATS

호주 시드니에서 개발된 SCATS(Sydney Coordinated Adaptive Traffic System)는 세계적으로 널리 쓰이는 적응형 신호제어 시스템이다. SCATS는 교차로 바닥에 매립된 차량 감지기나 비디오 카메라를 통해 차량 대기 길이와 흐름을 실시간으로 측정한다. SCATS는 보통 1-10개의 교차로를 하나의 서브시스템으로 묶어, 주요 도로를 따라 초록불이 연속되도록 한다. 예를 들어, 특정 도로가 붐비면 해당 도로 쪽 초록불 시간을 늘리고, 반대편은 줄이는 방식으로 동작한다. 이를 통해, 차량의 정지 횟수와 재가속이 줄어들어 연료와 시간을 모두 절약할 수 있다. 시드니와 더불어 여러 해외 도시에서의 운영 결과, SCATS 도입 구간에서 차량의 정지 횟수가 25% 줄었으며 연료 소비와 탄소 배출량은 각각 12%, 15% 감소하는 효과를 보인다. SCATS 시스템은 날씨, 시간대, 교통 수요 변화를 반영해 자동으로 신호를 조정하므로, 교차로 혼잡 해소와 함께 온실가스 저감에도 기여한다[95].

3) 미국 UPS ORION

미국의 글로벌 물류 기업인 UPS는 배송 차량의 위치, 속도, 정차 정보, 소포별 배송 시

94) Smith, Stephen, et al. “Smart urban signal networks: Initial application of the surtrac adaptive traffic signal control system.” Proceedings of the International Conference on Automated Planning and Scheduling. Vol. 23. 2013.

95) SCATS(2022), Sydney Coordinated Adaptive Traffic System(SCATS) https://www.transport.nsw.gov.au/system/files/media/documents/2022/SCATS-Core-brochure-Final-web-spreads_0.pdf

간약속·중량·부피 제약, 실시간 교통 상황을 바탕으로 매일, 매시 경로를 재계산하는 ORION(On-Road Integrated Optimization & Navigation)을 물류 시스템에 적용했다. 복잡한 다양한 거점 경유지가 포함된 최적 경로 찾기 문제를 최적화 알고리즘을 통해 계산하도록 설계했다. ORION 시스템은 복잡한 도로 교통 상황에 따라 즉시 재계산되어 최적의 경로를 제안한다. UPS의 ORION을 통해 운전자 1인당 하루 10~14마일 주행이 줄일 수 있으며 최대 연간 약 1억 마일 주행 감축과 연간 약 1,000만 갤런 연료 절감 달성한다[96].

4) 미국 Amazon

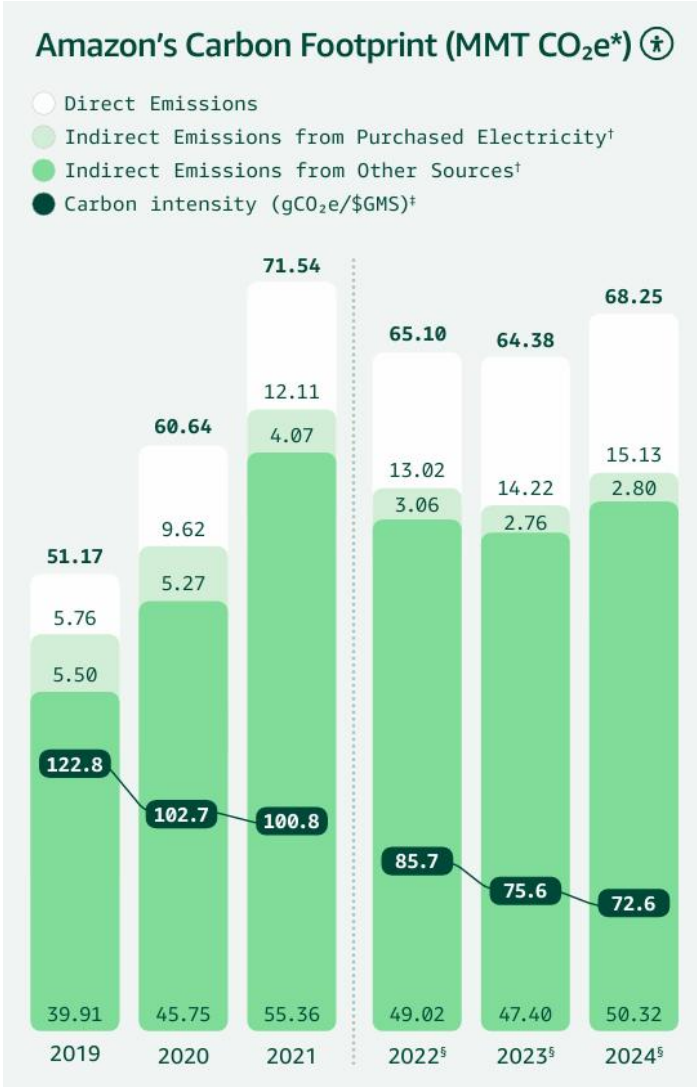
아마존은 글로벌 물류 과정에서 발생하는 탄소 배출을 줄이기 위해, AI 기반 물류 최적화 시스템과 친환경 운송 인프라를 통합한 지속가능 물류 플랫폼을 구축하였다. 해당 시스템은 주문·배송 데이터를 실시간으로 분석하는 AI 예측 모델과, 전기차(EV)·드론·로봇을 포함한 다계층 운송체계를 결합하여, 물류 전 단계에서 효율성과 탄소 절감을 동시에 달성한다. 특히, 2024년 지속가능성 보고서에 따르면 아마존은 2019년 대비 배송 단위당 탄소 배출량을 33% 감축하였다. 〔그림 3－15〕에서 볼 수 있듯이, 전체 탄소 배출량은 2021년 71.54 MMT CO_2 eq에서 2024년 68.25 MMT CO_2 eq로 완만히 감소했으며, 같은 기간 배송 단위당 탄소집약도(Carbon Intensity)는 122.8 → 72.6 g CO_2 eq/$GMS로 크게 낮아졌다. 이는 AI 기반 경로 최적화와 재고 배치 알고리즘이 교통 상황, 날씨, 주문 밀도, 차량 적재율 등의 변수를 실시간 분석하여 최적 운행 경로를 산출하고, 불필요한 공차 이동과 중복 운송을 최소화한 결과이다. 또한, 고객 주문을 통합 처리하는 스마트 배송 통합 프로그램을 통해 2024년 한 해 동안 4억 5,200만 건의 배송을 통합, 4억 9,400만 개의 박스를 절감, 그 결과 33만 5,000톤(335,000 t CO_2 eq) 배출 회피라는 감축 효과를 거두었다. 또한 운송 수단의 전기화를 병행하여 현재까지 6.8백만 건 이상의 배송을 전기차로 수행하였으며, 전기차 제조 기업 리비안(Rivian)으로부터 10만 대 규모의 맞춤형 전기 배송 차량을 도입해 글로벌 친환경 배송망을 확장하고 있다. 동시에, 자율운행 기반의 Prime Air 드론과 Scout 배송 로봇을 실증 운영하여 라스트마일 배송 단계에서 배송 비용 약 30% 절감 및 탄소 배출 저감 효과를 입증하였다.[97]

---

96) BSR(2016), Looking Under the Hood: ORION Technology Adoption at UPS

97) Amazon, 2024 Sustainability Report, “Carbon and Energy – Enhancing Routing and

〔그림 3－15〕 아마존의 연도별 탄소 배출 추이(2019-2024)



자료: Amazon, 2024 Sustainability Report, "Carbon and Energy", pp.9-10.

Fleet Efficiency", pp.8, 11-12.

나. 국내 기술 동향

1) 인천광역시 디지털 물류 실증사업

인천광역시 디지털 물류 실증사업은 화물차 적재함에 스마트 적재공간 관리기기와 자동 상하차 보조기술을 장착하여, 적재 상태를 실시간으로 모니터링하고 물류 전 과정을 디지털화한다. 이를 통해 적재율 제고, 불필요 운행 최소화, 작업 효율 향상이라는 세 가지 핵심 목표를 실현한다. 기술적으로는 적재함 바닥에 장착된 IoT 센서가 화물 하중과 적재 공간 활용도를 지속적으로 측정·전송하며, 자동 상·하차 장치는 인력 의존도를 줄여 상하차 작업 시간을 단축한다. 이러한 기술 결합의 결과, 하차 평균 소요시간이 17.85% 감소하여 작업 효율이 크게 향상되었고, 동일 시간 내 평균 배송 건수가 15.1% 증가하였다. 또한, AI 기반 경로 최적화와 실시간 적재 모니터링으로 대형 트럭의 공차 주행거리가 27.9% 감소하여, 탄소 배출량을 18.7% 실현한다. 특히 도서지역인 백령도에서는 순회 집하 서비스 도입을 통해 차량 통행거리가 57% 줄었으며, 군부대 공급 운행에서는 운행 거리 효율이 26.5% 증가했다. 이러한 실증사업은 스마트 물류 기술이 물류 효율, 작업환경, 탄소 감축 그리고 지역 특성을 고려한 맞춤형 운송 최적화까지 동시에 달성할 수 있음을 보여주는 우수 사례이다[98].

2) 포스코

포스코는 물류 부문에서의 탄소 배출을 줄이기 위해 디지털 기술과 친환경 운송수단을 결합한 ESG 전략을 추진하고 있다. 해상 운송 부문에서는 기존 벙커유 기반 선박 대신 액화천연가스 추진 벌크선을 도입하여 탄소 배출을 대폭 감축하고, 황산화물 배출을 99%, 질소산화물 배출을 85% 이상 저감하였다. 육상 운송에서는 LNG 화물차량을 확대 도입하여 경유 차량 대비 탄소 배출량을 약 19% 감축하였으며, 질소산화물 및 초미세먼지 배출량을 95% 이상 저감하였다. 또한 항만 정박 시 선박 엔진을 가동하지 않고 전력을 공급할 수 있는 육상 전원 공급 설비를 구축하여, 연간 탄소 배출을 추가로 감축하고 미세먼지 약 1.5톤, 탄소산화물 9.9톤, 황산화물 2.5톤을 줄였다. 탄소산화물 9.9톤은 대형 화물차 약 400대가 1년간 배출하는 양에 해당하며, 황산화물 2.5톤은 수천 명의 호흡 건강에 영향을 미칠 수 있는 규모이다. 이러한 복합적인 탄소 감축 노력은 포스코 물류 부문의 탄소 발자

98) 인천광역시(2025), 인천시, 스마트 물류체계 혁신으로 미래기술 선도 나선다 !

국을 획기적으로 줄였으며, 국가 온실가스 감축목표 달성과 국제 해운 탄소중립 규제 대응에 기여하고 있다. 이 모델은 제조·물류 기업 전반에 적용 가능한 대표적인 스마트 물류 저탄소 전환 사례로 평가된다[99].

3) KT-롯데온의 LIS’ FO(리스포)

KT는 롯데그룹의 e커머스 플랫폼인 롯데온(LOTTE ON)과 협력하여 AI 기반 운송 플랫폼 리스포를 전국 물류 현장에 적용함으로써 물류 효과와 탄소 감축을 동시에 달성하고 있다. 리스포는 모빌리티 빅데이터와 AI 기반 최적화 알고리즘을 활용하여, 배송 기사에게 실시간으로 최적 배송 경로와 운행 일정을 자동 제공하는 지능형 운송 관리 시스템이다. 기존에는 운송 담당자가 수작업으로 약 30분 이상 소요되던 경로 수립과 배차 확정 절차를 수행하였으나, 리스포 도입 이후에는 3분 이내로 단축되어 업무 효율이 대폭 향상되었다. 도입 결과 운행 거리는 최대 22% 감소, 운행 시간은 최대 11% 단축, 탄소 배출량은 약 22% 절감되는 정량적 효과가 확인되었다. 특히, 이는 단순한 운행 거리 절감이 아니라, AI가 교통 혼잡, 배송 순서, 차량 적재 효율 등을 종합적으로 고려해 불필요한 공차 이동과 재운송을 최소화함으로써 달성된 성과이다. 결과적으로, 리스포를 통해 배송 기사에게 맞춤형 모바일 애플리케이션을 제공하여 실시간 운행 정보를 직관적으로 확인할 수 있도록 지원함으로써 물류 근로자의 업무 부담을 줄이고, 운송 일정 예측 정확도를 높여 서비스 품질을 향상시켰다.[100]

4) CJ대한통운

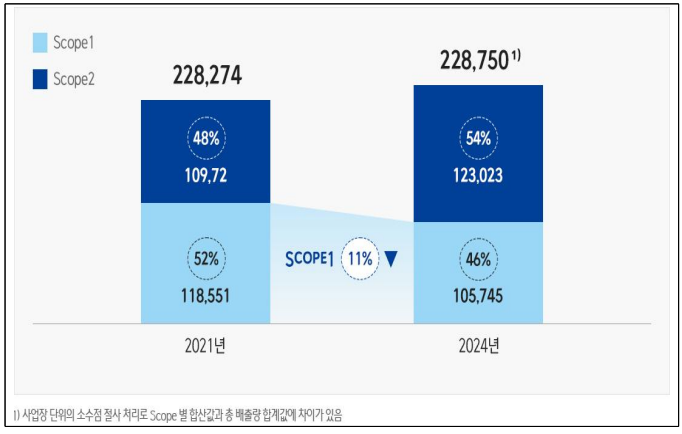
CJ대한통운은 TES 물류기술연구소를 중심으로 개발된 AI 물류 플랫폼은 교통량, 날씨, 도로 상황, 실시간 주문량, 차량 적재율 등을 통합 분석하여 운행 경로 및 배차를 자동 최적화한다. 이 시스템은 기존 인력 의존형 계획 대비 운행 거리를 약 15~20% 단축하고, 공차율을 10% 이상 감소시켜 연료 소모와 온실가스 배출량 절감에 직접적인 효과를 가져왔다. 또한, CJ대한통운은 전국 물류센터와 택배 허브에 AI 기반 에너지 관리 시스템(EMS)과 수송 복화 알고리즘을 적용하여, 차량의 공차 이동을 최소화하고 다중 고객의 물량을 효

99) 포스코(2019), 포스코, ESG, 디지털 물류 발전을 위한 세미나 개최

100) 김동원, 「KT, AI로 롯데마트 물류 탄소 배출 22% 저감」, 조선비즈 디지털조선, 2023. 1. 31.

율적으로 통합 운송하도록 설계하였다. 〔그림 3－16〕에서 볼 수 있듯이, 2024년 기준 Scope 1 배출량은 2021년 대비 11 % 감축(105,745 t CO_2 eq)되었으며, 단위 매출당 온실가스 배출집약도는 기준연도 대비 6 % 감소(2.02 → 1.89 t CO_2 eq/억 원)하였다.

〔그림 3－16〕 CJ대한통운의 Scope 1·2 온실가스 배출량 비교(2021-2024)



자료: CJ대한통운, 「기후변화 대응」

4. 클라우드 · 엣지 · IoT 기반 산업 에너지 효율화

클라우드·엣지·사물인터넷(IoT) 기술은 산업 현장에서 발생하는 데이터를 실시간으로 수집·처리·분석하여 설비 운영 효율을 극대화하고, 에너지 소비를 최소화하는 핵심 디지털 인프라로 부상하고 있다.

이러한 기술들은 AI 기반 최적화 시스템과 달리 데이터의 전송·분석·제어 과정 전반의 구조적 효율화를 통해 탄소 감축 효과를 유도한다. 즉, 클라우드는 대규모 데이터 연산을 고효율 서버로 통합함으로써 불필요한 전력 소모를 줄이고, 엣지 컴퓨팅은 현장 단에서 데이터를 처리하여 지연과 네트워크 전송 전력을 절감하며, IoT는 설비 상태를 정밀하게 감시·제어함으로써 불필요한 가동을 방지한다.

이들 기술의 결합은 산업 공정, 물류, 에너지 관리 등 다양한 분야에서 실시간 제어 기반의 에너지 최적화와 온실가스 감축을 가능하게 하고 있다.

〈표 3－5〉 클라우드 · 엣지 · IoT 기반 산업 에너지 효율화 주요 특징

구분	주요 특징	내용
1	클라우드 기반 자원 효율화	중앙 데이터센터의 고효율 서버 운영으로 개별 기업의 서버 운용 전력 절감. 가상화·자원공유를 통해 IT 인프라 에너지 사용량을 20~30% 절감.
2	엣지 컴퓨팅 기반 실시간 제어	데이터 전송 없이 현장에서 즉각 분석·제어 수행으로 네트워크 전력 손실 및 지연 최소화. 공정별 가동 효율 향상.
3	IoT 기반 설비 모니터링 및 자동 제어	온도·압력·전류 등 설비 데이터를 실시간 감시하고, 이상 상태를 자동 제어해 불필요한 에너지 낭비 방지.
4	산업 전반의 통합 에너지 관리	클라우드-엣지-IoT의 연계를 통해 설비, 공정, 물류, 건물 등 전 주기 에너지 사용 패턴을 통합 관리.

자료: 스마트모빌리티 서비스의 현황 및 발전방안 연구(2020)

가. 국외 기술 동향

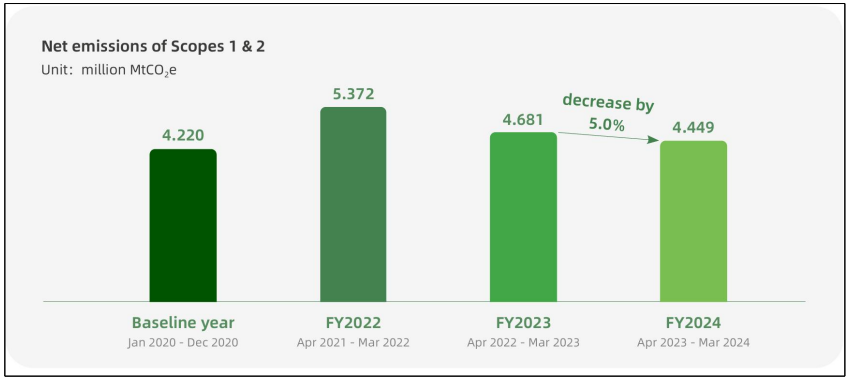
1) 중국 Alibaba Cloud

중국의 Alibaba Cloud는 데이터센터의 전력 효율과 탄소 감축을 동시에 달성하기 위해 지속적인 인프라 고도화와 청정 전력 전환 전략을 추진하고 있으며, 2030년까지 자체 운영의 탄소중립을 실현하고, 동일 시점까지 Alibaba Cloud가 Scope 3 탄소중립을 달성하는 것을 목표로 하고 있다.[101]

알리바바는 데이터센터 효율 개선과 재생에너지 전환을 통해 전사 온실가스 배출량을 실질적으로 감축하였다. 자가 구축형 데이터센터의 전력사용효율(PUE) 은 1.215에서 1.200으로 개선되었으며, 전체 전력소비의 56 %를 청정 전력으로 조달하였다. 이러한 개선 결과 〔그림 3－17〕에서 볼 수 있듯이 운영 부문(Scopes 1 & 2)의 순배출량은 2023년 4.681 $MtCO_2$ eq에서 2024년 4.449 $MtCO_2$ eq로 5.0 % 감소하였고, 〔그림 3－18〕과 같이 Scope 3의 배출강도(Net emission intensity) 역시 8.7에서 8.1 tCO_2 eq/백만 RMB로 7.0 % 감소하였다.

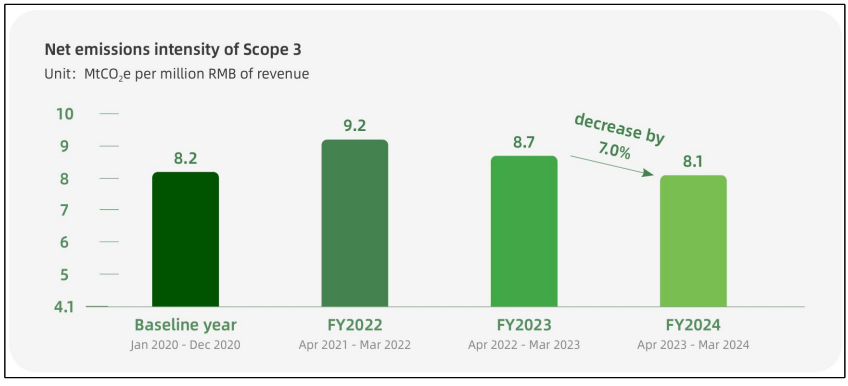
101) Alibaba Group Carbon Neutrality Action Report(2021)

〔그림 3－17〕 Alibaba Group Scope 1 & 2 온실가스 순배출량 추이(2020-2024)



자료: Alibaba Group ESG Report(2024)

〔그림 3－18〕 Alibaba Group Scope 3 배출강도 변화(2020-2024)



자료: Alibaba Group ESG Report(2024)

또한 Alibaba는 데이터센터 전력 공급의 안정성과 지속가능성을 확보하기 위해 중국 허베이성의 500 MW급 source-grid-load-storage 프로젝트와 장쑤성의 20년 장기 전력구매계약(PPA) 을 추진하였다. 이를 통해 클라우드 인프라 운영 전반에서 청정 전력 비중을 안정적으로 확대하고 있다.

특히, 이러한 녹색 인프라 전환과 효율화 결과, Alibaba Cloud의 클라우드 서비스 이용 고객들은 2024 회계연도 기준 약 988만 톤(9.884 Mt) 의 탄소 배출을 감축하였으며, 이는

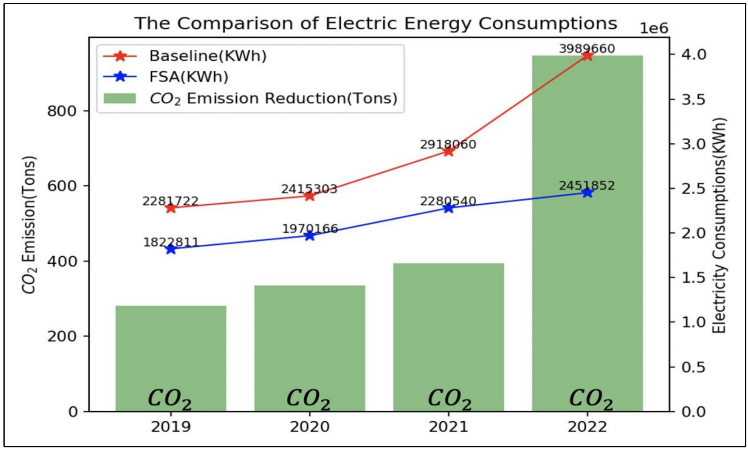
전년 대비 44% 증가한 수치로 보고되었다. Alibaba Group은 이를 "Green Computing Infrastructure" 전략으로 명명하고, AI 서비스·대규모 언어모델 운영 등으로 인한 컴퓨팅 전력 수요 증가에 대응해 청정 컴퓨팅 체계를 강화하고 있다.[102]

2) Ant Group - Full Scaling Automation(FSA)

Ant Group의 FSA 시스템은 AI 기반 클라우드 워크로드 예측 및 자원 자동 스케일링 기술로, 데이터센터의 에너지 효율과 탄소 감축을 동시에 달성한 대표적 사례이다. 이 기술은 서버 부하와 트래픽 패턴을 실시간으로 분석하여, 필요한 만큼의 컴퓨팅 자원을 자동으로 확장·축소함으로써 불필요한 전력 낭비를 방지한다.

논문 "Full Scaling Automation for Sustainable Development of Green Data Centers" (IJCAI 2023)에서는 FSA가 도입된 데이터센터의 실제 운영 결과를 보고하였다. 2019년부터 2022년까지 4년간 누적 CO_2 감축량은 약 1,958톤, 전력 절감량은 약 3,182,000kWh로 나타났으며, 특히 2022년 한 해 동안만 947톤 CO_2, 1,538,000kWh의 절감 효과를 거두었다.

〔그림 3－19〕 Ant Group FSA 적용에 따른 전력 소비 및 CO_2 감축 추이



자료: Wang et al., Full Scaling Automation for Sustainable Development of Green Data Centers, IJCAI 2023

102) Alibaba Group(2024), Alibaba Progresses Towards Carbon Neutrality Goals and Digital Inclusion - 2024 ESG Report, Alibaba Cloud Community, July 23, 2024

〔그림 3－19〕은 베이스라인 대비 FSA 적용 후 전력소비량 감소와 이에 따른 연도별 탄소 배출 저감량을 시각적으로 보여준다. 붉은 선은 기존 전력소비량, 파란 선은 FSA 적용 후 소비량, 녹색 막대는 CO_2 감축량을 나타낸다. 이 시스템은 클라우드 인프라에서 발생하는 비효율적 자원 활용 문제를 최소화하고, 실시간 부하 변화에 따른 최적 전력 운용을 통해 평균 25~30%의 에너지 효율 향상을 달성하였다.

나. 국내 기술 동향

1) 네이버클라우드 데이터센터

네이버클라우드는 디지털 기술 기반의 데이터센터 효율화와 탄소 저감 인프라 구축을 통해, ICT 산업 내 대표적인 친환경 인프라 모델을 제시하고 있다.

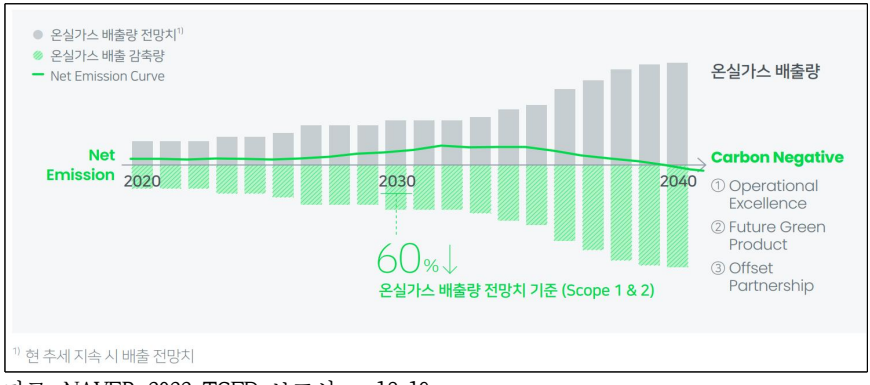
네이버는 전체 온실가스 배출량의 약 99%가 데이터센터와 사옥의 전력 사용에서 기인함에 따라, 2020년 '2040 Carbon Negative' 전략을 수립하고 운영 효율 최적화와 재생에너지 전환을 병행 추진 중이다.

NAVER 2022 TCFD 보고서에 따르면, 회사는 2030년까지 온실가스 배출 전망치 대비 60 % 감축, 2040년까지 Carbon Negative 달성을 목표로 하고 있다.

〔그림 3－20〕에서 볼 수 있듯이, 2020년을 기준으로 2030년까지 배출 전망치(Scope 1 & 2 기준) 대비 60 % 감축을 달성하고, 이후 2040년까지 순배출(Net Emission)을 0으로 만드는 경로를 제시한다.

〔그림 3－20〕 네이버 온실가스 배출 전망 및 감축 경로

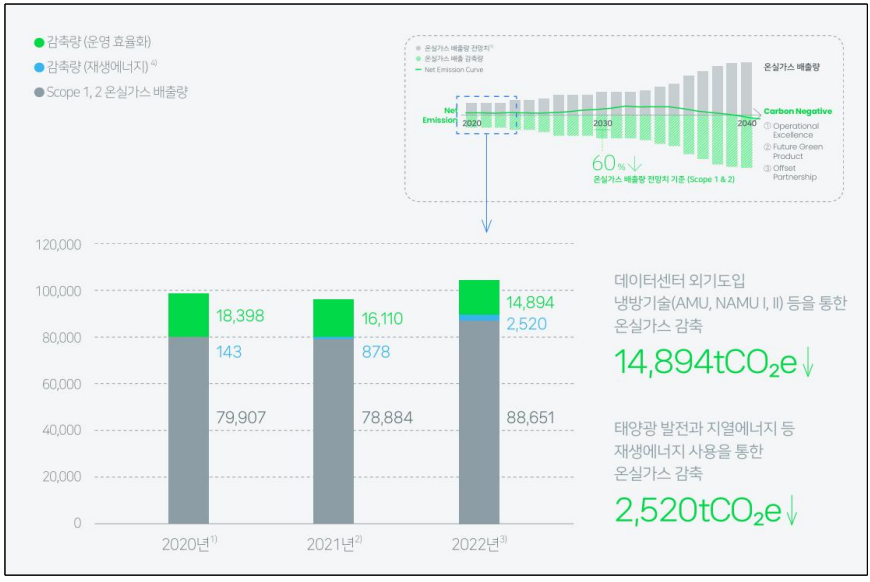
(단위: tCO_2 eq)



자료: NAVER 2022 TCFD 보고서 pp.18-19

〔그림 3－21〕 Scope 1·2 온실가스 배출량 및 감축 실적

(단위: tCO_2 eq)



자료: NAVER 2022 TCFD 보고서 pp.18-19

또한 〔그림 3－21〕는 Scope 1·2 온실가스 배출량의 실제 감축 성과를 보여준다. 2022년 기준 Scope 1·2 온실가스 배출량은 88,651 tCO_2 eq이며, 운영 효율화를 통해 14,894 tCO_2 e, 재생에너지 사용을 통해 2 ,520 tCO_2 eq를 각각 감축하였다. 세종 데이터센터는 춘천 IDC의 운영 데이터를 학습한 AI 제어형 Hybrid 냉방시스템을 도입하였다. 이 시스템은 외기 활용을 극대화하고 냉매·냉각탑 제어를 자동화함으로써, 각 춘천 대비 냉방 에너지 효율을 최대 20 % 개선하도록 설계되었다. 옥상 태양광(연 300 MWh)과 지열 시스템, 폐열 회수 설비를 병행해, 서버실의 폐열을 건물 온수 및 스노우멜팅 시스템에 재활용하는 순환형 구조를 구축하였다. 또한 네이버는 자사 IDC 및 사옥의 에너지 데이터를 통합 관리하기 위해 BEMS 및 AI 기반 예측 모델을 도입하였다. 이를 통해 시간대별 전력 수요와 외기 조건을 분석하여, 공조 시스템의 가동을 자동 최적화하고 연간 5% 이상의 에너지 절감 효과를 지속적으로 달성 중이다. 또한 LEED Platinum 인증과 ISO 14001 환경경영시스템을 운영하며, 탄소배출 Scope 3 관리체계를 2022년부터 구축하고 있다.[103)]

103) NAVER 2021 TCFD 보고서

2) KEPCO·한국에너지공단 에너지 통합관리 플랫폼

한국전력공사(KEPCO)와 한국에너지공단은 산업단지 단위에서의 에너지 효율화를 목표로, AI·IoT 기반의 스마트에너지플랫폼을 구축하고 있다. 이 플랫폼은 개별 기업의 공장에너지관리시스템(FEMS)과 산업단지 단위의 산업단지에너지관리시스템(CEMS)을 통합하여, 전력·열·가스 등의 에너지 데이터를 실시간으로 수집·분석하고 수요 예측 및 제어를 수행한다. 또한 통합운영센터(TOC, Total Operating Center)를 통해 전국 스마트그린산단의 에너지 데이터를 클라우드 기반으로 연계·관리함으로써, 산업단지 전반의 에너지 효율화와 정책 수립에 활용되고 있다.

반월·시화 스마트그린산단에서는 22개 기업에 FEMS를 설치하여 공기압축기·집진기·펌프 등 설비별 동력 제어를 수행한 결과, 평균 25%의 전력 절감 효과를 달성하였으며, 연간 약 5,100만 원의 에너지 비용을 절감하였다. 구미 국가산단에서는 CEMS 분석을 기반으로 주요 부하 설비에 동력제어 솔루션을 적용해 약 30 TOE(ton of oil equivalent)의 에너지를 절감하였으며, 창원 스마트그린산단에서는 AI 기반 공정 스케줄링 조정을 통해 연 120 TOE 절감과 에너지 비용 약 2,000만 원 절감을 실현하였다.

광주 첨단산단에서는 대기전력 차단 기술 적용을 통해 연간 약 27 TOE의 에너지 절감이 보고되었으며, 여수 산단에서는 피크 제어 및 수요반응(DR) 시스템을 도입해 연 85.8 TOE 절감과 누적 800만 원 비용 절감을 달성하였다. 이러한 실증 결과는 AI·IoT 제어 기술을 활용한 통합 에너지 관리체계가 산업단지 단위의 에너지 절감과 탄소 배출 저감에 실질적인 효과를 발휘하고 있음을 보여준다.[104)105)]

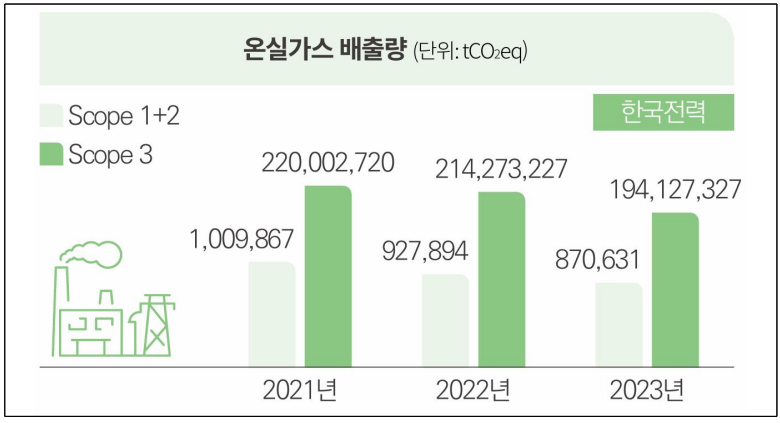
또한 한국전력은 2050년 탄소중립 달성을 목표로 'ZERO for Green' 비전을 공동 선포하고, 재생에너지 및 수소·암모니아 혼소 발전 확대, 전력망 강화, 지능형 전력공급 시스템 구축 등의 전략을 추진하고 있다.

104) 한국산업단지공단, 스마트그린산단 CEMS 우수사례

105) 한국전력 및 전력그룹사 지속가능경영보고서(2024)

〔그림 3－22〕 한국전력 온실가스 배출량 추이

(단위: tCO_2 eq)



자료: 한국전력 및 전력그룹사 지속가능경영보고서(2024), p. 55

〔그림 3－22〕은 한국전력의 온실가스 배출량 추이를 나타낸 것으로, Scope 1 + 2 직접·간접배출량이 2021년 220,002,720 tCO_2 eq에서 2023년 194,127,327 tCO_2 eq로 감소하였으며, Scope 3 배출량도 1,009,867 → 870,631 tCO_2 eq로 감축되었다. 이는 전력공기업의 탈탄소 전환 정책과 탄소중립 이행 노력이 실질적인 성과로 나타나고 있음을 보여준다.

제 4 장 AI 기반 산업의 저탄소화 기술

제 1 절 소프트웨어 기반 솔루션

1. 경량화 모델 개발

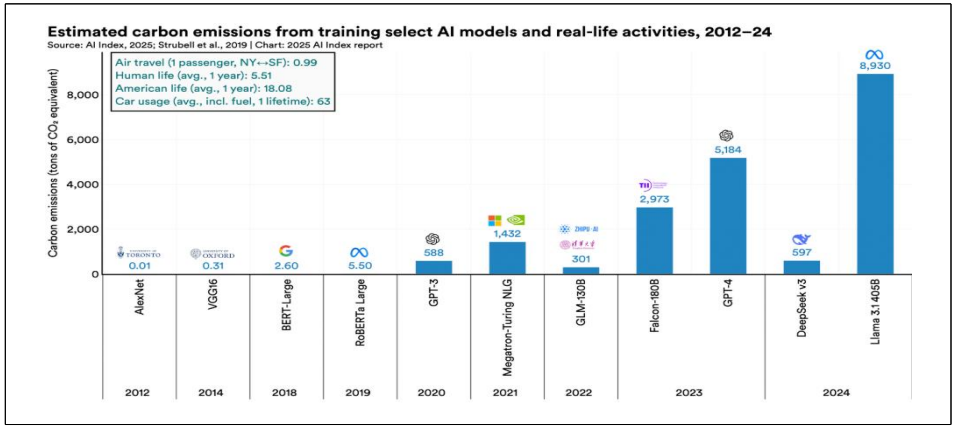
가. 모델 경량화의 필요성

AI 모델은 최근 몇 년간 기계학습의 분야에서 다양한 혁신적인 발전을 이루었으며, 자연어 처리, 컴퓨터 비전 등 다양한 분야에서 우수한 성능을 달성했다. 그러나 이러한 우수한 성능의 이면에는 방대한 학습 데이터를 기반으로 많은 파라미터를 가진 복잡한 모델 구조를 대규모 Graphics Processing Unit(GPU) 서버에서 학습 및 추론을 해야 한다는 한계가 존재한다. 특히, 최근 주목받는 ChatGPT와 같은 초거대 언어 모델은 수억 개 이상의 파라미터를 보유하며, GPU 연산 유닛과 메모리 간의 대규모 데이터 전송과 방대한 연산 수행으로 인해 전력 소모와 탄소 배출이 상당하다. 구체적으로, 〔그림 4-1〕에서 볼 수 있듯이, 2012년 AlexNet(Krizhevsky, et al. 2012) 의 학습 과정에서 발생한 탄소 배출량은 약 0.01톤의 CO_2에 불과했으나, 모델 규모와 복잡성이 증가함에 따라 2020년 GPT-3(Brown, et al. 2020)는 약 588톤, 2023년 GPT-4(Achiam, et al. 2020)는 약 5,184톤, 2024년 LLaMA 3.1 (Touvron, et al. 2023)은 약 8,930톤의 CO_2 를 배출하였다[106]. 이는 세계 평균 1인당 연간 배출량(약 5.51톤)과 비교했을 때 1,620배 이상, 미국 평균 1인당 연간 배출량(약 18.08톤)과 비교했을 때도 약 494배에 해당하는 막대한 수치이다. 결국, 이러한 방대한 에너지 소비와 탄소 배출을 절감하기 위해서는 모델 경량화를 중심으로 한 소프트웨어 기반 최적화 기법의 적용이 필수적이다. 구체적으로는 불필요한 파라미터 수와 연산을 줄이는 프루닝(pruning) 〔그림 4-2〕, 데이터의 표현 정밀도를 낮춰 전력 효율을 높이는 양자화

106) AI Index, Artificial Intelligence Index Report 2025, Stanford Institute for Human-Centered Artificial Intelligence, 2025. [Online]. Available: https://hai.stanford.edu/ai-index/2025-ai-index-report/research-and-development.

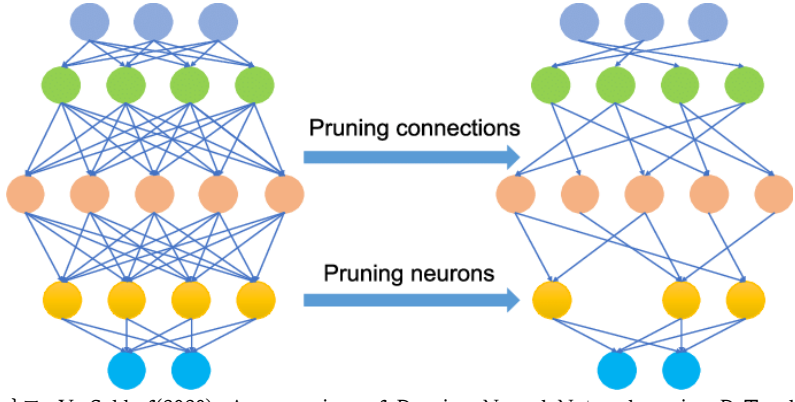
(quantization)〔그림 4-3〕 등이 대표적이다. 이러한 기법들은 학습 및 추론 단계에서 불필요한 연산을 최소화하고, GPU 연산 유닛과 메모리 간 데이터 전송 빈도를 줄여 전력 소모를 효율적으로 억제함하여 동일한 성능을 유지하면서도 탄소 배출량을 크게 줄일 수 있다.

〔그림 4-1〕 AI 모델 발전에 따라 학습 과정에서 발생하는 탄소 배출량 비교



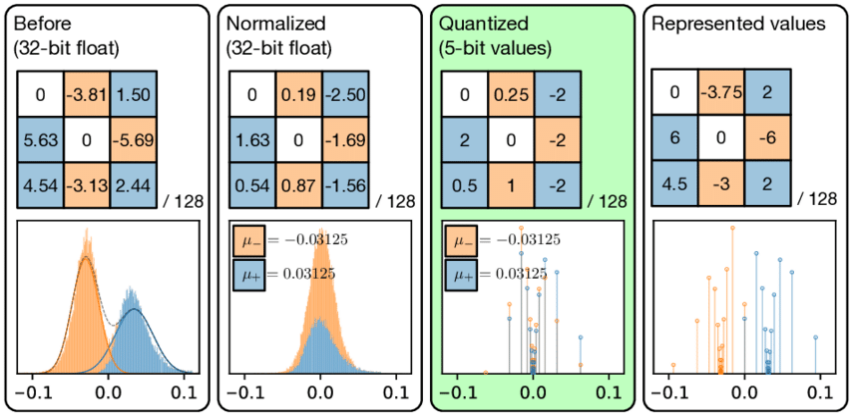
자료: Stanford University, Aritificial Intelligence Index Report 2025, p. 50.

〔그림 4-2〕 딥러닝 네트워크 프루닝 개요



자료: Y. Sakkaf(2020), An overview of Pruning Neural Networks using PyTorch

〔그림 4－3〕 딥러닝 네트워크 양자화 개요



자료: Zhao, et al.(2019), p.3

나. 양자화 개요 및 연구 동향

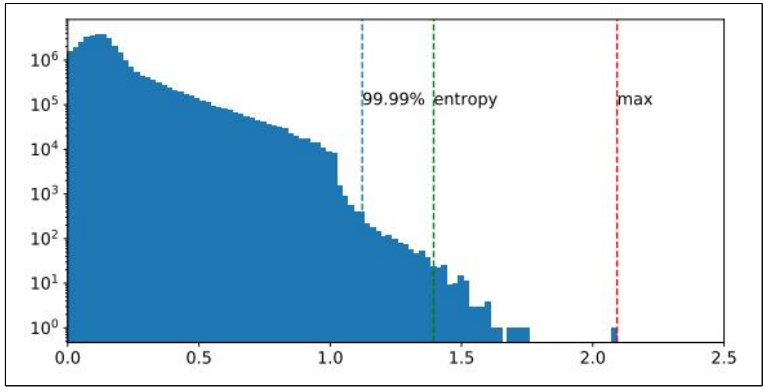
1) 양자화 개요

양자화는 32비트 혹은 16비트 부동 소수점으로 표현된 모델의 가중치, 입력 데이터를 더욱 낮은 비트 정밀도로 변환하여 메모리 사용량을 줄여 궁극적으로 전력 소모 및 탄소 배출을 절감할 수 있는 효과적인 경량화 기법이다. 특히, 가중치 양자화의 경우 AI 모델의 메모리 요구량을 낮추어 탄소 절감에 기여할 수 있으며, 입력 데이터에 양자화를 적용하는 경우, 부동 소수점 행렬 곱셈 연산을 하드웨어 친화적인 정수 연산으로 변환하여 추론 효율성을 향상시킬 수 있다.

2) 양자화 적용 방법

우선, 양자화를 적용하기 위해 양자화가 적용될 범위를 정한다. 양자화 적용 범위는 데이터의 통계적 특성을 기반으로 설정되며, 주로 Max, Entropy, Percentile 등 다양한 클리핑 기준이 사용된다. 예를 들어, Max 기준은 전체 데이터에서 절대값이 가장 큰 값을 상한으로 설정한다. 반면 Percentile은 상위 극소수의 값을 클리핑하여 전체 분포에 대한 양자화 표현력을 높이고, Entropy 기준은 양자화된 값의 정보 엔트로피를 최소화하는 범위를 선택한다. 〔그림 4－4〕에서 볼 수 있듯이, 이러한 방법들은 서로 다른 범위를 제안하며, 선택된 범위는 이후 스케일 팩터 계산과 정수 변환 과정에 직접적으로 영향을 미친다.

〔그림 4-4〕 양자화 적용 범위의 예시



자료: Wu, et al.(2020), p.6

양자화 범위가 정해지면, 범위 내에서 정수 변환을 위한 스케일링 팩터를 구하게 된다. 스케일링 팩터는 임의의 32비트 부동 소수점 입력 x에 대해 선택된 범위 $[\alpha,\beta]$ 내의 데이터를 주어진 양자화 비트 수 b 에 따른 표현 범위에 선형적으로 매핑하기 위해 계산된다. 일반적으로 많이 사용되는 대칭 양자화에서는 스케일링 팩터를 다음과 같이 계산한다:

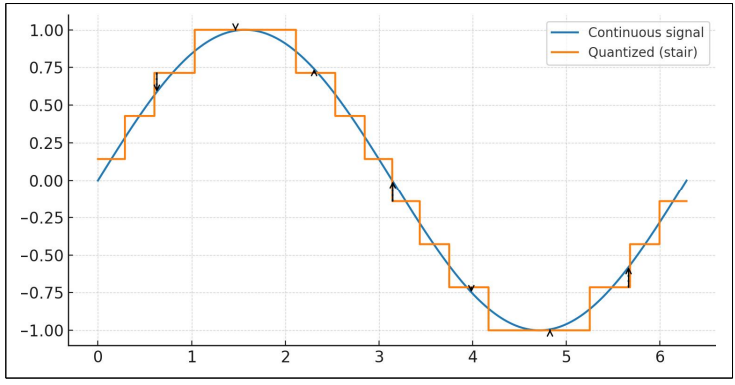
$$s = \frac{\max(|\alpha|,|\beta|}{2^{b-1}-1}$$

결국, 이를 통해 다음과 같이 낮은 비트 표현의 정수 x_q로 변환하게 된다:

$$x_q = clamp(r(x/s)),-2^{b-1},2^{b-1}-1$$

여기서, $clamp(x,\min,\max)$ 함수는 입력 x가 주어진 최솟값, 최댓값 범위를 넘어가지 않도록 하한과 상한을 제한하는 구간 포화 함수를 의미한다. 또한 r은 반올림 함수를 의미하며 이를 통해 데이터를 정수 도메인으로 변환한다. 양자화 과정에서는 데이터를 낮은 비트로 변환하는 과정에서 양자화 오차가 발생하며, 이는 종종 AI 모델의 성능 저하로 이어진다〔그림 4-5〕. 따라서, 모델 크기를 효율적으로 줄이면서도 정확도 저하를 최소화하여, 정확도와 모델 압축 간의 상충 관계(trade-off)를 최적화하는 것이 핵심 과제이다.

〔그림 4－5〕 양자화 에러 예시



자료: 연구진 작성

3) 양자화 종류

양자화는 적용에 있어 크게 사후 훈련 양자화(post-training quantization)와 양자화 인식 훈련(quantization-aware training)으로 나뉜다. 또한, 비트 정밀도 할당 방법에 있어 혼합 정밀도 양자화와 고정 정밀도 양자화로 구분된다.

가) 사후 훈련 양자화

사후 훈련 양자화는 사전 학습된 AI 모델 가중치를 불러와 양자화를 적용하는 방식으로, 스케일링 팩터 산출을 위한 소규모 데이터셋 기반의 보정(calibration) 과정만 수행하면 되기 때문에 실용성이 매우 높다. 그러나 비트 정밀도가 극도로 낮아질 경우 정확도 저하가 심각하게 발생하므로, 최소한의 최적화 및 추론 비용 증가만으로도 성능 저하를 완화하는 것이 핵심 과제로 남는다.

나) 양자화 인식 훈련

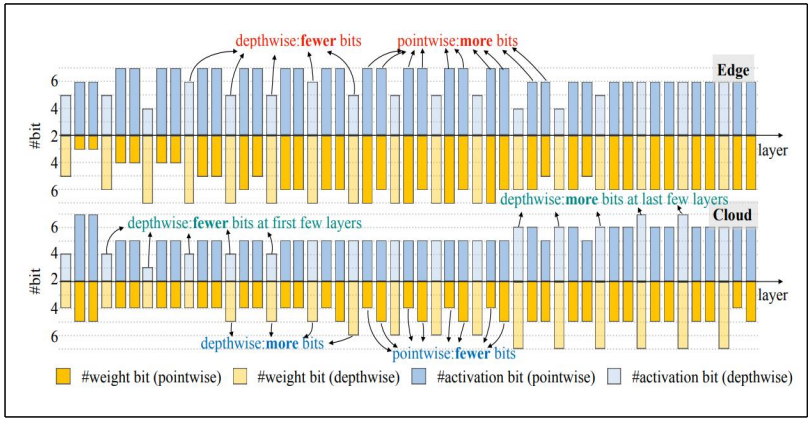
양자화 인식 훈련은 AI 모델 학습 과정에서 양자화 연산을 모사하여 양자화로 인한 오차를 보정할 수 있도록 가중치를 함께 최적화하는 방법이다. 이를 통해 낮은 비트 정밀도 환경에서도 안정적이고 우수한 성능을 확보할 수 있으며, 일반적으로 사후 훈련 양자화보다 높은 정확도를 달성한다. 그러나 양자화 인식 훈련은 양자화를 적용하기 위해 모델을 처음부터 다시 학습해야 하므로 막대한 훈련 비용이 소요되며, 대규모 모델의 경우 하드

웨어 자원과 시간 제약이 커 실시간 또는 자원이 제한된 환경에서의 적용이 제한적이다.

다) 혼합 정밀도 양자화

일반적인 양자화 방법은 AI 모델의 모든 레이어 계층에 동일한 비트 정밀도를 할당한다. 이와 달리, 혼합 정밀도 양자화 방법은 레이어별 양자화 민감도를 측정하여, 민감도에 따라 상이한 비트 정밀도를 할당한다. 구체적으로, 양자화 민감도가 높은 레이어에는 정확도 하락을 방지하기 위해 높은 비트 정밀도를 할당하고, 민감도가 낮은 레이어에는 비트 정밀도를 낮추어 모델의 크기를 공격적으로 줄인다〔그림 4-6〕.

〔그림 4-6〕 혼합 정밀도 양자화 예시



자료: Wang, et al.(2019), p.7

〈표 4-1〉 Convolutional Neural Networks 양자화 기술 보유 현황

기술 보유국	기술 보유 기관	양자화 정밀도	정확도
미국(Dong, et al. 2019)	UC Berkeley	4-bit	75.48%
미국(Nagel, et al. 2020)	Qualcomm	4-bit	75.01%
중국(Li, et al. 2021)	UESTC	4-bit	75.05%
미국(Faraone, et al. 2018)	Xilinx	2-bit	72.30%
미국(Choi, et al. 2018)	IBM	4-bit	76.50%
미국(Zhang, et al. 2018)	Microsoft	4-bit	76.40%

기술 보유국	기술 보유 기관	양자화 정밀도	정확도
호주(Zhuang, et al. 2018)	The University of Adelaide	4-bit	75.70%
한국(Kim, et al. 2019)	Seoul National University	4-bit	77.30%
미국(Wang, et al. 2019)	MIT	4-bit	76.14%
미국(Dong, et al. 2020)	UC Berkeley	2-bit	75.92%

〈표 4－2〉 Vision Transformer 양자화 기술 보유 현황

기술 보유국	기술 보유 기관	양자화 정밀도	정확도
중국(Lin, et al. 2021)	MEGVII	8-bit	81.20%
중국(Yuan, et al. 2022)	Peking University	6-bit	81.55%
중국(Ma, et al. 2024)	ByteDance	6-bit	81.72%
중국(Li, et al. 2023)	CAS	6-bit	81.27%
싱가포르(Zhong, et al. 2023)	Xiamen University	4-bit	79.97%
중국(Yang, et al. 2023)	CAS	4-bit	80.13%
중국(Wei, et al. 2022)	Beihang University	4-bit	79.96%
중국(Wu, et al. 2025)	Beihang University	4-bit	80.21%
중국(Wu, et al. 2025)	Beihang University	4-bit	80.33%

〈표 4－3〉 LLM 양자화 기술 보유 현황

기술 보유국	기술 보유 기관	양자화 정밀도	Perplexity (↓)
미국(Tseng, et al. 2024)	Cornell University	4-bit	3.38
미국(Chee, et al. 2023)	Cornell University	4-bit	3.53
미국(Ashkboos, et al. 2024)	MIT	4-bit	3.79
스위스(Lin, et al. 2024)	ETH Zurich	4-bit	3.41
중국(Zheng, et al. 2025)	Beihang University	4-bit	3.39
미국(Liu, et al. 2024)	Meta	4-bit	3.50
미국(Lin, et al. 2024)	MIT	4-bit	3.46
중국(Shao, et al. 2023)	Shanghai AI Laboratory	4-bit	3.50
미국(Li, et al. 2025)	Meta	2-bit	4.26

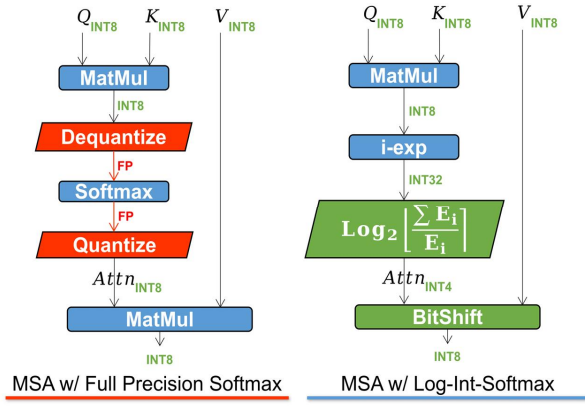
4) 양자화 연구 동향

양자화 연구는 주로 이미지 처리 및 컴퓨터 비전 분야에서부터 활발히 발전해 왔으며, 전통적인 convolutional neural networks(CNNs) 모델은 일반적으로 파라미터 규모가 상대적으로 작고, 입력 데이터(activation)가 연산 병목 현상의 주요 원인이었다. 따라서, 연산 효율성을 높이기 위해 activation 양자화를 통해 부동 소수점 행렬 곱 연산을 정수 연산으로 변환하여 처리 속도를 향상시키는 데 초점이 맞춰졌다. 대표적인 연구로, Dong et al.(2019) 연구에서는 혼합 정밀도 양자화 시 레이어별 비트 수를 효율적으로 할당하기 위해, Hessian 행렬의 최대 고유값을 이용해 각 레이어의 양자화 민감도를 정량적으로 평가하는 Hessian Aware Quantization(HAWQ) 기법을 제안하였다.

Nagel et al.(2020) 연구에서는 사전 학습된 AI 모델에 양자화 적용 시 일반적으로 사용되는 최근접 반올림 방법(rounding-to-nearest)이 최적이 아니라는 점을 이론적으로 분석하였다. 또한 작업 손실을 테일러 전개 기반으로 근사하여, 양자화 반올림 함수를 전체 모델 손실 최소화 관점에서 최적으로 결정하는 adaptive rounding(AdaRound) 기법을 제안하였다. Li et al.(2021) 연구에서는 기존 사후 훈련 양자화 방법이 네트워크 레이어 간 의존성을 고려하지 않아 매우 낮은 비트 정밀도 조건에서 정확도가 급격히 하락하는 문제를 분석하였다. 이를 해결하기 위해 헤시안 행렬 기반의 오차 분석을 통해, 의존성이 강하게 작용하는 네트워크 블록 단위로 모델을 최적화하는 기법을 제안하였다.

최근에는 크게 Vision Transformer(ViT) 모델(Dosovitskiy, et al. 2020)과 LLM 모델의 양자화 연구가 활발히 수행되었다. 대표적인 ViT 양자화 연구로 Lin et al.(2021) 연구는 기존 양자화 방법들이 LayerNorm, Softmax를 정수로 처리하지 않아 완전한 정수 연산이 불가능했던 문제를 지적했으며, Power-of-Two Factor와 Log-Int-Softmax를 도입하여 ViT 모델의 모든 연산을 정수로 구동하는 방법을 제안하였다〔그림 4-7〕.

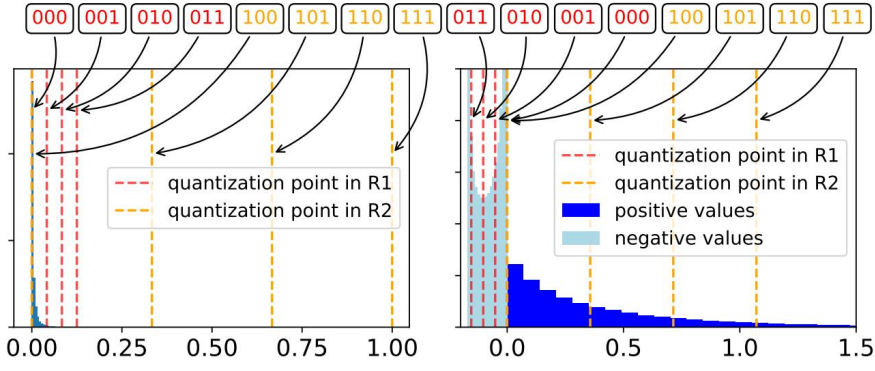
〔그림 4−7〕 완전 정수 연산의 예시



자료: Lin, et al.(2021), p.4

Yuan et al.(2022) 연구에서는 ViT의 활성화 데이터에서 발생하는 비대칭적인 이상치 분포에 의해 양자화 성능이 저하되는 현상을 발견하였으며, 이를 해결하기 위해 activation의 분포를 두 개의 구간으로 나누어 양자화 하는 Twin Uniform Quantization 기법을 제안하였다〔그림 4−8〕.

〔그림 4−8〕 Twin-Uniform Quantization 개요

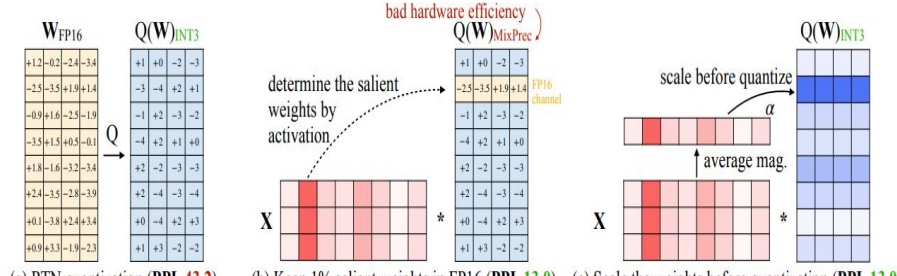


자료: Yuan, et al.(2022), p.7

Ma et al.(2024)는 ViT 양자화 적용 시 정확도 하락의 주요 원인인 이상치 문제의 핵심 해결책으로 기존에 간과되었던 재구성 단위 개념을 제시하였다. 해당 연구에서는 ViT와 같이 구조가 균일한 모델은 여러 블록을 함께 최적화하는 넓은 단위를, Swin Transformer (Liu, et al. 2021)과 같이 계층적 구조를 가진 모델은 더욱 세분화된 단위를 사용하는 것이 효과적임을 증명하였다.

반면, 최근 LLM을 대상으로 한 양자화 연구는 주로 가중치의 비트 정밀도를 낮추어, 모델 파라미터 저장에 필요한 메모리 용량을 줄이는 데 집중되어 왔다. 대표적인 연구로 Frantar et al.(2023)는 기존 사후 훈련 양자화 기법들을 LLM에 적용하기에는 최적화 계산 비용이 과도하게 높아 실용성이 떨어진다는 점을 지적하였다. 이에 따라, OBQ 기법 (Frantar, et al. 2022)을 확장하여, 가중치를 열 단위로 양자화함으로써 최적화 비용을 크게 절감하는 아이디어를 제시하였다. 결과적으로 1,750억개 파라미터를 가진 모델을 단 몇 시간 만에 3-4비트로 압축하며 단일 GPU에서 거대 모델 추론을 가능하게 했다. Lin et al.(2024)는 기존 LLM 양자화 기법이 핵심 가중치를 보호하지 못하는 한계를 지적하였다. 해당 연구에서는 전체 가중치 중 약 1%만이 모델 성능에 결정적인 영향을 미치며, 이러한 중요도는 가중치 값 자체가 아니라 activation 값의 크기에 의해 결정된다는 통찰을 제시하였다. 이를 바탕으로, 활성화 값을 기준으로 중요 채널을 식별하고 스케일링을 통해 보호하는 activation-aware weight quantization(AWQ) 기법을 제안하였다〔그림 4－9〕.

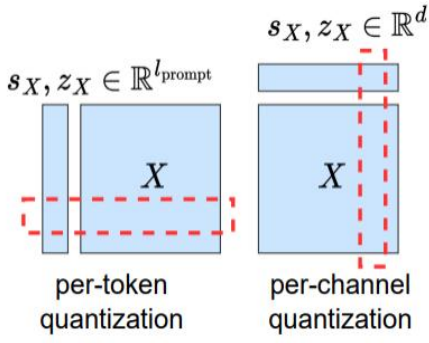
〔그림 4－9〕 Activation-aware Weight Quantization



자료: Lin, et al.(2024), p.3

한편, long-context 데이터를 처리하는 시나리오에서는, 디코딩 단계에서 매 토큰마다 전체 이전 토큰의 KV Cache에 접근해야 하므로, KV Cache의 대규모 메모리 점유가 큰 문제점으로 떠올랐다. 이에 따라, KV Cache에 양자화를 적용하는 시도도 활발하게 이루어지고 있다. 구체적으로, Hooper et al.(2024)은 기존 KV Cache 양자화가 4비트 미만 초저정밀도에서 실패하는 원인이 Key 활성화 값의 특이 분포, RoPE의 영향, 비효율적인 양자화 비트 할당, 동적 이상치 문제 때문이라고 분석하였다. 각 문제를 해결하기 위해 Key는 채널별 RoPE 적용 전에 양자화하고, 민감도에 기반한 비균등 데이터 타입을 생성하며, 이상치를 벡터별로 분리하는 종합적인 접근법을 제안하였다. 한편, Liu et al.(2024)은 LLM의 Key, Value 캐시가 서로 다른 분포 특성을 가져 동일한 양자화 방식을 적용할 수 없다는 문제를 제기하였다. 심층 분석을 통해 Key 캐시는 고정된 이상치 채널을 가지므로 채널별 양자화가 적합하고, Value 캐시는 어텐션 스코어에 의해 혼합되므로 토큰별 양자화로 오차를 각 토큰에 국한시켜야 함을 밝혔다. 이러한 비대칭적 접근에 기반한 튜닝 없는 양자화 알고리즘을 제안하였으며, 메모리 사용량을 2.6배 줄였다〔그림 4－10〕.

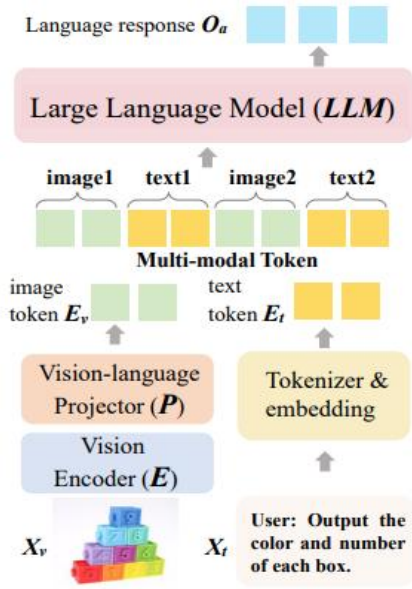
〔그림 4－10〕 토큰, 채널 별 양자화 예시



자료: Liu, et al.(2024), p.2

5) Vision-Language Model(VLM) 양자화 연구 동향

〔그림 4－11〕 VLM 모델 구조



자료: Yu, et al.(2025), p.3

VLM은 시각적 정보와 언어적 정보를 통합적으로 이해하기 위해 설계되었으며, 3개의 멀티모달 신경망 구조(e.g., Vision Encoder, Vision-language Projector, and LLM)로 구성된다〔그림 4－11〕. 첫 번째 구성 요소인 Vision Encoder E는 입력 이미지 X_v로부터 시각적 의미를 표현하는 패치 단위의 벡터를 추출한다. 일반적으로 Vision Transformer(ViT) 기반 구조가 사용되며 이는 아래 수식으로 정의된다:

$$F_v = E(X_v)$$

Vision-language Projector P는 Vision Encoder의 출력 F_v를 입력으로 받아서 언어 모델이 사용하는 텍스트 임베딩 공간으로 투영한다. 이는 두 모달리티 간 표현 공간의 불일치를 해소하기 위한 단계로, 보통 선형 변환 혹은 fully connected layer를 활용하여 layer를 구성하여 학습과 추론을 수행하며, 이 과정은 아래와 같이 수식으로 정의된다:

$$E_v = P(F_v)$$

Projector의 출력 $E_v \in R^{N_v \times d_t}$는 언어 모델 입력 차원 d_t 로 정렬되어, 텍스트 임베딩과 융합을 가능하게 만든다. 한편, 텍스트 입력 X_t는 tokenizer를 통해 단어 단위로 분리되고, 임베딩 레이어를 통해 각 토큰을 벡터 형태로 만든다. 이 과정은 아래와 같이 수식으로 정의된다:

$$E_t = f_{text}(X_t)$$

비전 임베딩 E_v와 텍스트 임베딩 E_t은 시퀀스 단위로 융합되어 각각 VLM의 멀티모달 입력을 구성한다. LLM은 결합된 입력을 기반으로 자동회귀 방식의 언어 생성 과정을 수행하며, 결과적으로 출력 시퀀스 $O_a = [y_1, y_2, ..., y_t]$를 생성한다. 이러한 생성 과정은 다음의 확률적 표현으로 정의된다.

$$LLM(O_a \mid E_v, E_t) = \prod_{i=1}^{l} LLM(y_i \mid E_v, E_t, y < i)$$

여기서 $y < i$는 마지막 토큰 이전까지의 모든 생성 결과를 의미하여, l은 최종 시퀀스의 길이를 나타낸다. 즉, VLM은 비전 토큰과 텍스트 토큰으로 구성된 멀티모달 정보를 바탕으로 새로운 언어 토큰을 반복 순차적으로 예측한다.

이러한 VLM의 구조는 모델의 표현력과 일반화 능력을 극대화할 수 있다. 그러나 수십억 개에 달하는 파라미터와 대규모 멀티모달 입력으로 인해서 연산 복잡도와 메모리 사용량이 폭발적으로 증가하게 된다. 이러한 결과로, 막대한 전력 소모량과 고성능 병렬 처리 장치 구성을 위한 비용도 함께 요구된다. 이는 VLM의 현실적이고 효율적인 배포와 실시간 응답성 및 신뢰성을 높이기 위해서는 정확도 성능을 유지하면서 경량화할 수 있는 모델 압축 기법들(e.g., 양자화, 프루닝, 지식전이) 이 필수적으로 적용되어야만 한다.

LLM을 위한 기존 양자화 기법들은 대부분 시각적 정보 없이 텍스트를 위한 context만 고려하여 설계되었기 때문에 멀티모달 입력 간의 상호작용이 존재하는 VLM 구조에 적합하지 않다는 한계점을 가진다. 따라서, VLM 구조에서는 각 모달리티의 분포적 특성을 반영하고 모달리티 간 상호작용 효과를 유지할 수 있는 VLM 전용 양자화 기법이 필수적이다. 요약하자면, 기존 LLM 양자화는 VLM의 구조적 특성과 입력 분포를 고려하지 않기 때

문에 직접적으로 적용하기 어려우며, 모달리티 간 비대칭적 분포와 상호작용을 반영한 새로운 양자화 기법이 필요하다.

대표적인 VLM 연구인 Li et al.(2025) 는 기존 LLM의 PTQ 방식이 비전 토큰과 텍스트 토큰을 동일하게 처리하기 때문에 두 토큰 사이의 민감도 차이를 간과한다는 점을 지적한다. Li et al.(2025) 는 COCO dataset의 image-caption dataset을 활용하여 supervised fine-tuning(SFT) 손실 함수를 설계하여 gradient 분석을 수행하였으며, 결과적으로 텍스트 토큰의 민감도가 비전 토큰보다 약 10배 이상 크다는 점을 증명한다〔그림 4－12〕.

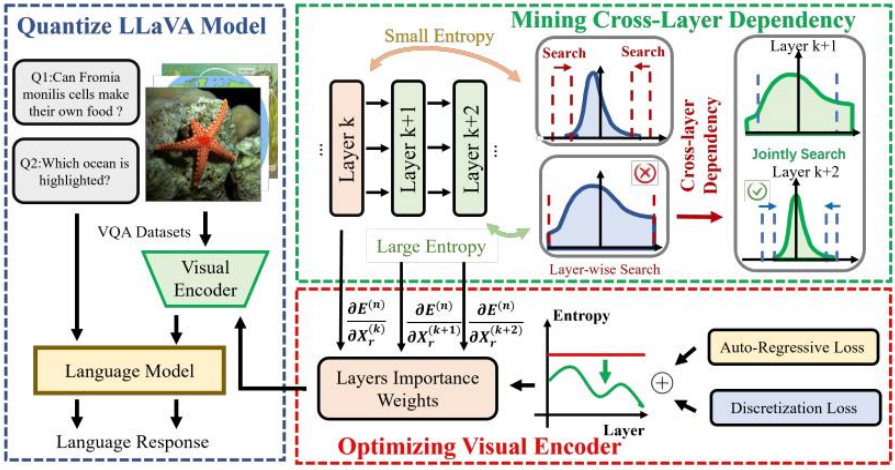
Wang et al.(2024) 연구에서는 기존 양자화 기법인 Frantar et al.(2023), Lin et al.(2024)가 레이어 별 독립적인 rounding function을 적용하는 방식을 채택하여 레이어 간 의존성를 고려하지 않기 때문에, 전체 네트워크 관점에서 최적의 rounding strategy를 찾지 못하고 누적 오차가 발생하여 성능 저하로 이어질 수 있다는 것을 지적한다. 이를 해결하기 위해서, 선행적으로 활성화 값의 엔트로피와 이산화 오차 간의 상관 관계와 레이어 간 의존성을 정량적으로 분석이 필요하며 결과에 따라, block 내에서 이산화 오차를 최소화하는 rounding function을 탐색하고 전체 네트워크 차원에서의 누적 오차를 억제할 수 있도록 레이어 간 의존성을 반영한 block-wise 양자화 프레임워크를 제안한다〔그림 4－13〕.

〔그림 4－12〕 비전, 텍스트 토큰 간 민감도 예시

자료: Li, et al.(2024), p.1

Yu et al.(2025) 연구는 기존 LLM 기반 양자화 기법이 VLM의 멀티모달 구조적 특성과 분포 차이를 반영하지 못한다는 점을 문제로 지적한다. 비전 토큰이 텍스트 토큰보다 비중이 더 높고 분포의 범위가 넓어서 두 모달리티를 동일한 스케일로 양자화하면 텍스트 토큰의 표현이 왜곡된다는 점을 분석하였으며, 이를 해결하기 위해, 해당 연구는 세 가지 기법으로 구성된 정적 양자화 프레임워크를 제안한다.

〔그림 4－13〕 Q-VLM 양자화 프레임워크



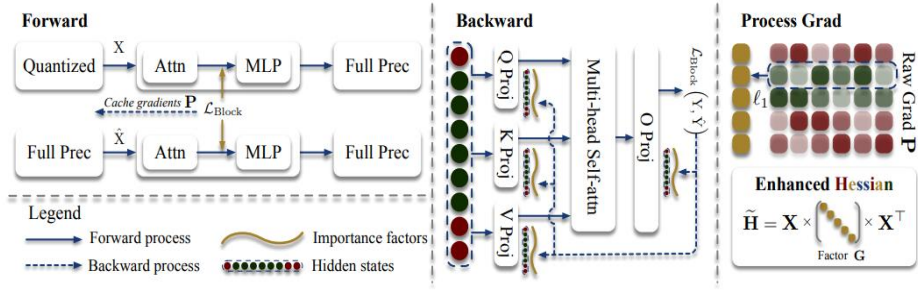
자료: Wang, et al.(2024), p.3

(1) Modality-specific static quantization은 모달리티 별로 서로 다른 스케일을 적용해 비전, 텍스트 토큰 간의 분포 불균형을 완화시키고, (2) Attention invariant flexible switching은 토큰 인덱스와 causal mask를 변형해 토큰 순서가 달라져도 동일한 어텐션 연산을 유지할 수 있도록 만든다, (3) Rotation magnitude suppression은 Hadamard 변환 과정에서 발생하는 가중치의 이상치를 억제하여 양자화 안정성을 높인다. 결과적으로, 이러한 프레임워크를 통해 멀티모달 입력의 분포 특성을 효과적으로 반영하고, 기존 LLM 중심 양자화 방식에서 간과되었던 모달리티 간 상호작용을 정량적으로 고려할 수 있는 기법을 제시한다.

Xue et al.(2025) 연구는 기존 LLM 중심의 Hessian 기반 PTQ 기법들이 모든 토큰을 일정하게 처리함으로써 발생하는 비전-언어 모델의 모달리티 불균형 문제를 중점적으로 다룬

다. 본 연구는 VLM 구조 내에서 텍스트 토큰에 비해 과도하게 많은 비전 토큰이 존재하며, 이러한 Vision over-representation 현상이 양자화 과정에서 비전 정보에 편향된 손실을 유발한다는 점을 지적한다. 이를 해결하기 위해 Xue et al.(2025)는 토큰 별 중요도를 반영한 importance-aware PTQ 프레임워크를 제안한다〔그림 4－14〕. 구체적으로, SFT loss의 gradient를 기반으로 토큰별 민감도를 계산하고, 이를 활용해 중요도가 낮은 비전 토큰의 영향을 줄이고 텍스트 토큰의 표현을 보존하도록 Hessian을 조정하였다. 결과적으로, 해당 연구는 단순한 양자화 정확도 향상에 그치지 않고, 멀티모달 입력 간 불균형을 정량적으로 보상하는 새로운 양자화 패러다임을 제시한다. Xue et al.(2025)는 기존 LLM 기반 PTQ 기법들이 간과했던 모달리티 간 상호작용을 통계적 중요도로 통합 분석함으로써, calibration과 inference 단계에서 효율성과 일관성을 동시에 성취한다.

〔그림 4－14〕 VLMQ 양자화 프레임워크



자료: Xue, et al.(2025), p.4

〈표 4－4〉 VLM 양자화 기술 보유 현황

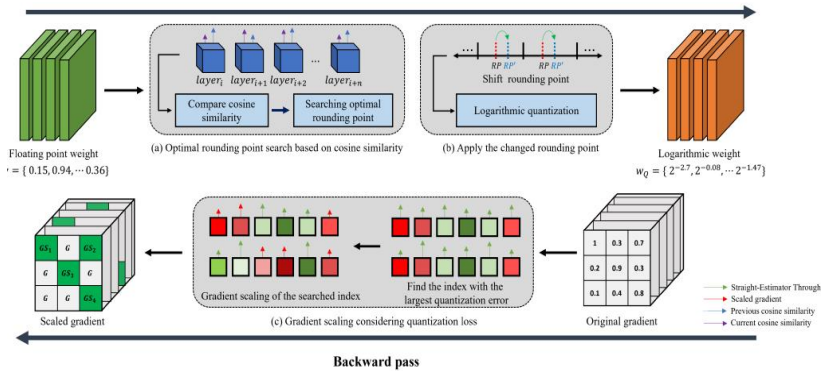
기술 보유국	기술 보유 기관	양자화 정밀도	정확도
중국(Wang, et al. 2024)	Tsinghua University	6-bit	49.86
중국(Li, et al. 2025)	Tsinghua University	3-bit	60.7
중국(Yu, et al. 2025)	HOUMO AI	8-bit	84.32
중국(Xue, et al. 2025)	TeleAI	3-bit	79.28

다. 양자화 연구 성과

1) HLQ: Hardware-Friendly Logarithmic Quantization Aware Training for Power-Efficient Low-Precision CNN Models(IEEE ACCESS, 2024)

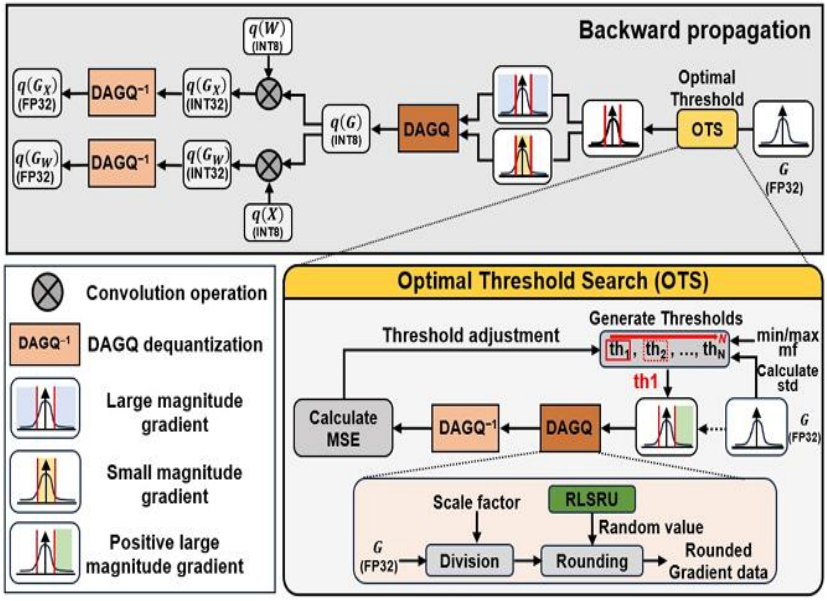
해당 연구에서는 기존 로그 기반 양자화 방식이 rounding point(RP) 편향과 부정확한 기울기 전파로 인해 저비트 환경에서 정확도가 크게 저하되는 문제를 발견하였다. 이에 따라, 본 연구는 양자화 손실과 코사인 유사도에 기반해 RP를 학습 과정에서 동적으로 조정하는 적응형 RP 기법과, 양자화 손실이 큰 파라미터의 기울기 크기와 방향을 조정하는 기울기 스케일링 기법을 결합한 방법을 제안하였다. 결과적으로 다양한 CNN 모델에서 가중치, 입력 데이터에 4-bit 양자화 적용 시 기존 대비 정확도 향상과 함께, RTL 구현 결과 전력 소모 82.3%를 절감하여 전력 효율적 추론을 가능하게 하였다〔그림 4－15〕.

〔그림 4－15〕 Hardware-Friendly Logarithmic Quantization Aware Training



자료: Choi, et al.(2024), p.3

〔그림 4－16〕 Gradient distribution-aware INT8 quantization



자료: Jeong, et al.(2025), p.2

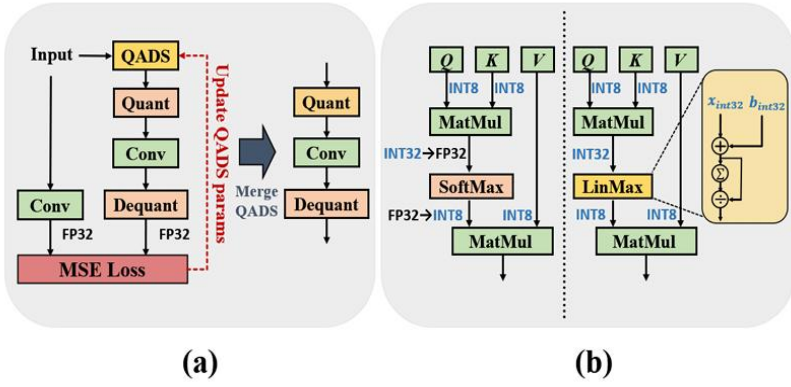
2) LowGradQ: Adaptive Gradient Quantization for Low-Bit CNN Training via Kernel Density Estimation-Guided Thresholding and Hardware-Efficient Stochastic Rounding Unit(DATE, 2025)

해당 연구에서는 저비트 CNN 학습에서 기울기 이상치로 인한 정확도 저하와 기존 채널 단위 양자화의 높은 연산, 메모리 부담 문제를 해결했다. 이를 위해 커널 밀도 추정 기반으로 기울기를 소규모와 대규모로 구분해 각각 다른 스케일 팩터를 적용하는 이중 스케일 적응형 기울기 양자화와 난수 재사용이 가능한 linear feedback shift register(LFSR) 기반 확률적 반올림 유닛을 제안하였다. 해당 기법은 ResNet-50 모델에서 부동 소수점 32비트 대비 +0.51% 정확도 향상과 함께, 기존 모델 대비 LUT와 FF 사용량을 최대 97% 이상 절감하였다〔그림 4－16〕.

3) HyQ: Hardware-Friendly Post-Training Quantization for CNN-Transformer Hybrid Networks(IJCAI, 2024)

해당 연구에서는, 기존 양자화 기법들이 CNN이나 ViT 단일 구조에만 특화되어 있어, 두 구조가 결합된 CNN+ViT 하이브리드 모델에서는 정확도가 크게 하락하는 문제를 발견하였다. 이를 해결하기 위해 하이브리드 모델의 각 구성 요소에 맞는 별도의 하드웨어 친화적 양자화 전략을 적용하는 아이디어를 제안하였다. 우선, CNN 블록의 이상치 문제를 해결하기 위한 quantization-aware distribution scaling(QADS) 방법과, Transformer 블록의 Softmax 연산을 정수 전용 선형 함수로 근사하는 Linear Softmax를 제안하여, 하드웨어 자원 사용량을 크게 줄이면서도 우수한 정확도를 달성하였다〔그림 4－17〕.

〔그림 4－17〕 (a): Quantization-aware Distribution scaling.(b): Linear Softmax



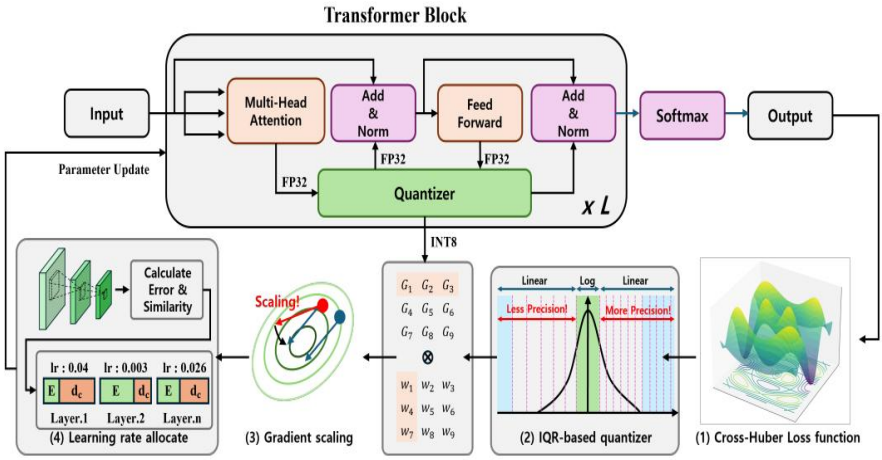
자료: Kim, et al.(2024), p.3

4) GradQ-ViT: Robust and Effcient Gradient Quantization for Vision Transformers (AAAI, 2025)

해당 연구에서는 ViT의 gradient 양자화 과정에서 outlier와 복잡한 loss landscape으로 인해 학습이 불안정해지는 문제를 지적하였다. 이를 해결하기 위해, gradient 분포의 사분위 범위를 기반으로 구간별 양자화 정밀도를 적응적으로 할당하는 기법과, cross-entropy와 huber Loss를 결합한 CH-Loss를 도입하여 outlier 영향을 완화하였다. 또한 양자화 후 변화된 gradient 크기와 방향을 원래 gradient에 가깝게 복원하는 gradient scaling 알고리

증과, gradient-양자화 오차 유사도를 반영하여 학습 안정성을 높였다〔그림 4－18〕.

〔그림 4－18〕 GradQ-ViT 개요



자료: Choi, et al.(2025), p.4

5) Zero-Centered Fixed-Point Quantization with Iterative Retraining for Deep Convolutional Neural Network-Based Object Detectors(IEEE Access, 2021)

해당 연구에서는 고정소수점 기반 양자화 적용 시, 0을 기준으로 하는 대칭형 범위 설정이 불가능하여 발생하는 양자화 오차 문제를 지적하였다. 이를 해결하기 위해, 0을 중심으로 한 대칭형 범위를 가지는 zero-centered fixed-point quantization 기법을 제안하였으며, 이를 통해 객체 검출 모델에서의 불필요한 양자화 편향을 줄였다. 또한 zero-centered fixed-point quantization 적용 후 성능 저하를 보완하기 위해 반복적 재학습 전략을 도입하여 정확도를 회복하였다. 실험 결과, 제안 기법은 다양한 객체 검출 모델에서 양자화에 따른 mean average precision(mAP) 감소를 최소화하고 임베디드 환경에서 효율적인 추론을 가능하게 하였다.

라. 프루닝 개요 및 연구 동향

1) 프루닝 개요

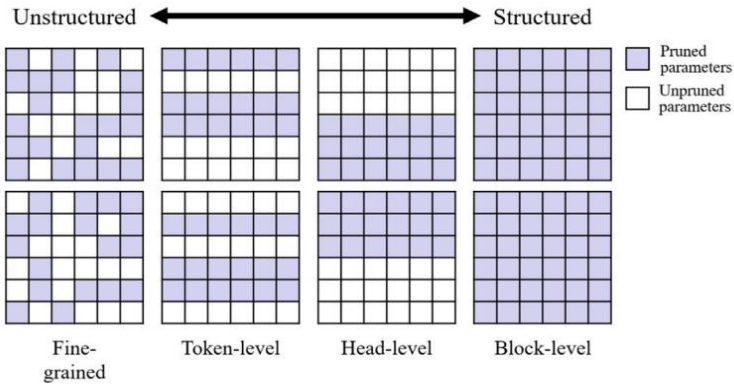
프루닝(Pruning)은 학습된 모델의 가중치 또는 뉴런 중 중요도가 낮은 요소를 제거하여

모델의 크기와 메모리 요구량을 줄이는 효과적인 경량화 기법이다. 불필요한 연결이나 파라미터를 제거함으로써 메모리 요구량과 데이터 전송량을 절감할 수 있으며, 이는 전력 소모 감소와 함께 탄소 배출 절감에 기여한다. 특히, 가중치 프루닝은 희소성을 활용해 메모리 효율을 높일 수 있고, 구조적 프루닝은 필터, 채널, 레이어 단위의 가중치 자체를 줄여 데이터 저장 및 전송 과정에서의 에너지 사용을 최소화함으로써 친환경적인 AI 모델 운용을 가능하게 한다.

2) 프루닝의 세밀성

프루닝은 제거하는 파라미터의 세밀성(granularity)에 따라 비구조적, 구조적 프루닝으로 분류된다. 〔그림 4－19〕에서 볼 수 있듯이, 비구조적 프루닝은 파라미터를 개별 단위로 제거하여 구조적 패턴에 비해 더 높은 압축률을 달성할 수 있다. 그러나 제거 패턴이 불규칙하므로, 실질적인 연산 가속 및 메모리 사용량 절감 효과를 얻기 위해서는 희소성을 효율적으로 처리할 수 있는 전용 하드웨어 지원이 필요하다. 반면 구조적 프루닝은 일정한 규칙을 따라 파라미터를 제거함으로써 모델 구조 자체를 간소화하므로, 전용 하드웨어 없이도 연산 가속이 가능하다. 또한 파라미터 크기를 직접적으로 축소하여 메모리 사용량 절감 효과를 제공한다. 그러나, 비구조적 프루닝에 비해 상대적으로 압축률이 낮아지는 단점이 존재한다.

〔그림 4－19〕 프루닝 세밀성에 따른 파라미터 제거 패턴 비교

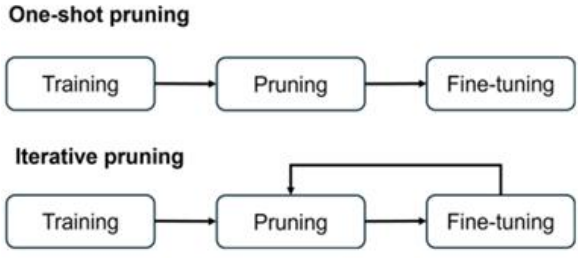


자료: Kang, et al.(2024), p.6

3) 프루닝의 스케줄링

프루닝은 적용 방법과 시기에 따라 one-shot, iterative 프루닝으로 나뉜다. one-shot 프루닝은 목표하는 프루닝 비율을 한 번에 네트워크에 적용하는 방식으로, 프루닝 이후 미세조정을 통해 성능을 회복하며, 상대적으로 간단하고 빠르지만, 성능 저하가 크게 발생할 수 있다. Iterative 프루닝은 목표하는 프루닝 비율에 도달하기 위해 네트워크에 점진적으로 프루닝을 적용하는 방식으로, 일반적으로 프루닝 적용 후 미세 조정을 반복하면서 점진적으로 목표한 프루닝 비율에 도달한다. 네트워크에 변화가 조금씩 적용되기 때문에 성능 손실은 적지만, 과정이 복잡하고 최적화 시간이 오래 걸릴 수 있다〔그림 4-20〕.

〔그림 4-20〕 프루닝 스케줄링에 따른 프로세싱 과정 비교



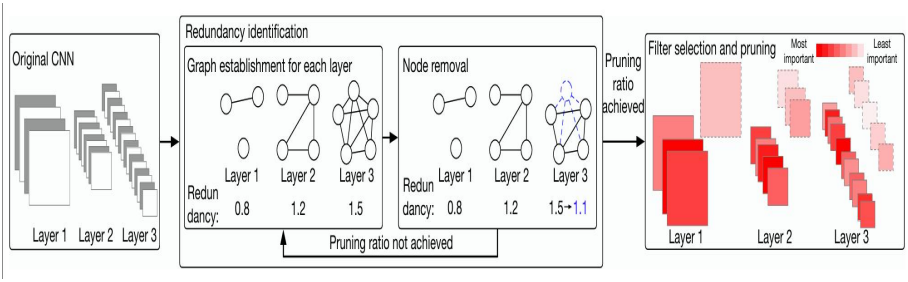
자료: 연구진 작성

4) 프루닝 연구 동향

프루닝 연구는 전통적인 CNNs 기반의 컴퓨터 비전 분야에서부터 활발하게 연구되었다. 대표적으로 Molchanov et al.(2019) 연구에서는 필터의 기여도를 손실 변화량으로 정의하고, 이를 1차 및 2차 테일러 전개를 통해 근사하여 필터 중요도를 계산하는 기법을 제안하였다. 이를 통해 전역적으로 비교 가능한 중요도 척도를 제공하며, 중요도가 낮은 필터를 반복적으로 제거하는 구조적 프루닝을 수행한다.

Wang et al.(2021) 연구에서는 구조적 중복도를 기반으로 한 채널 프루닝 기법을 제안하였다. 이를 위해, 각 합성곱 레이어를 그래프로 표현하고, L-커버링 수와 묶 공간 크기를 통해 레이어 내 필터 간 중복도를 측정한다. 이후 중복도가 가장 큰 레이어에서 필터를 제거하는 레이어 적응형 프루닝 방법을 제안하였다〔그림 4-21〕.

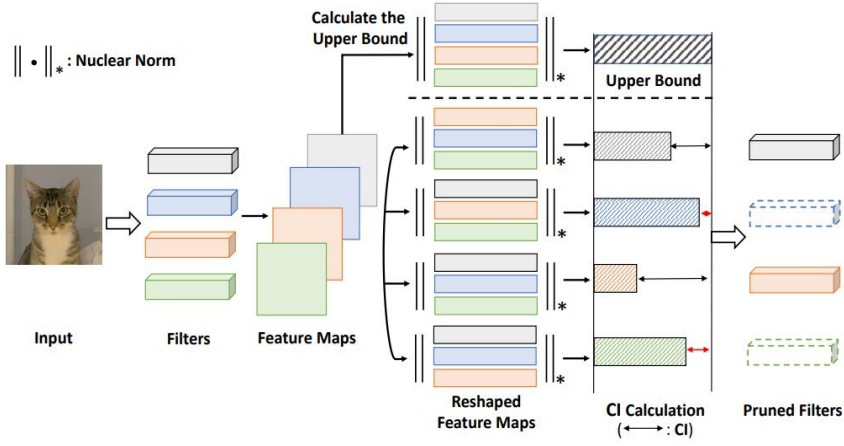
〔그림 4−21〕 구조적 중복도 기반의 채널 프루닝 기법



자료: Wang, et al.(2021), p.4

Sui et al.(2021) 연구에서는 채널 간 독립성을 기반으로 필터 중요도를 산정하는 channel independence-based pruning(CHIP) 기법을 제안하였다. 각 채널의 feature map을 행렬로 변환한 뒤, 해당 채널을 제거했을 때의 norm 변화량을 측정하여 독립성을 계산한다. 독립성이 낮은 채널은 다른 채널에 의해 대체 가능성이 높다고 보고 제거 대상 필터로 선정한다〔그림 4−22〕.

〔그림 4−22〕 채널 독립성 기반의 필터 중요도 산정 및 프루닝 기법

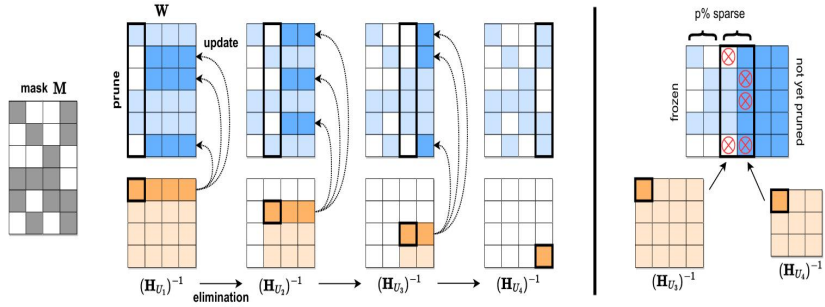


자료: Sui, et al.(2021), p.6

최근 LLM을 대상으로 다양한 프루닝 연구가 수행되었다. 대표적으로 Frantar et al. (2022) 연구에서는 사전 학습된 모델에 대해 재학습 없이 소량의 캘리브레이션 데이터만으로 가중치 프루닝과 양자화를 동시에 수행할 수 있는 optimal brain compressor(OBC) 프레임워크를 제안하였다. 이는 고전적인 optimal brain surgeon(OBS) 알고리즘을 레이어 단위로 정확하게 구현하여, 가중치의 손실 증가 영향을 기반으로 파라미터를 하나씩 제거하고, 그때마다 나머지 가중치를 closed-form으로 보정하는 방식이다. 이를 통해 NP-hard인 계층별 압축 문제를 효율적으로 풀어내고, 프루닝과 양자화를 결합한 복합 압축 환경에서도 높은 정확도를 유지하였다.

Frantar et al.(2023) 연구에서는 수십억에서 수백억 파라미터 규모의 GPT 계열 대규모 언어 모델을 재학습 없이 one-shot로 60%까지 희소화할 수 있는 SparseGPT 알고리즘을 제안하였다. 이는 프루닝 문제를 매우 대규모 희소 회귀 문제로 변환한 뒤, 효율적인 근사 해법을 사용하여 각 계층의 입력-출력 관계를 최대한 보존하면서 가중치를 제거하는 방법을 제안하였다〔그림 4－23〕.

〔그림 4－23〕 SparseGPT 알고리즘 개요



자료: Frantar, et al.(2023), p.4

Xia et al.(2024) 연구에서는 대규모 사전 학습 LLaMA2-7B 모델을 대상으로, 사전에 지정된 목표 아키텍처에 맞춰 레이어, 어텐션 헤더, 완전연결층 차원을 선택적으로 제거하는 targeted structured pruning 기법을 제안하였다. 이후 도메인별 손실 감소율에 따라 학습 배치의 데이터 비율을 동적으로 조정하는 dynamic batch loading 기법을 도입하여, 구조적

프루닝 후 추가 사전학습에서 성능 회복을 가속화하였다.

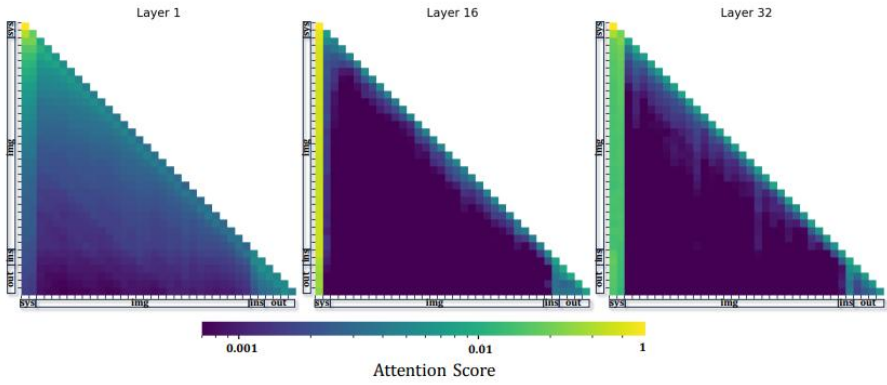
5) VLM 프루닝 연구 동향

VLM은 이미지와 텍스트를 동시에 처리하기 위해 수천 개의 비전 토큰과 수백 개의 텍스트 토큰을 결합하여 LLM의 입력으로 사용한다. 그러나 이러한 구조는 특히 고해상도 이미지를 입력으로 받을 때 시퀀스 길이가 급격히 증가하며, 이는 연산량과 메모리 요구량을 급격히 증가시킨다. 이러한 구조적 특성은 prefill 단계의 지연 시간 증가, 추론 시 메모리 병목 및 실시간 배포의 한계로 이어진다.

하지만 이미지에 포함된 정보는 텍스트보다도 더 높은 중복성을 가지며, 언어 응답 생성에서 모든 비전 토큰이 효과적으로 기여되지 않는다. 즉, 이미지의 일부 토큰은 중요한 의미적 정보를 포함하지만, 상당수의 토큰은 중복적이고 불필요한 정보를 담고 있다. 결과적으로 이러한 정보 분포는 VLM 구조에서 불필요한 계산과 메모리 낭비를 초래한다.

이러한 경향을 반영하여 최근 연구에서는 VLM을 위한 전용 프루닝 기법이 제안되고 있다. 대표적으로 Chen et al.(2024)는 기존 VLM 모델에서 나타나는 비효율적인 어텐션 구조를 지적하고 이를 개선하기 위한 동적 비전 토큰 프루닝 기법을 제안한다. 해당 연구에서 LLaVA, QwenVL-chat 등의 VLM을 분석한 결과, 더 깊은 레이어들에서 비전 토큰이 거의 어텐션 되지 않고 대부분의 시각 정보가 초반 레이어들에서만 집중되는 현상을 발견한다〔그림 4-24〕. 이는 VLM이 시각 정보를 과도하게 중복 처리함을 의미하고 이를 예방하지 않으면 연산량과 메모리 사용량의 비효율적 증가를 유발하게 된다. 이를 위해서, 해당 연구는 모델의 초반부에서 어텐션 패턴을 바탕으로, 중요도가 낮은 비전 토큰을 후반부 레이어에서 선택적으로 제거하는 기법을 제안한다. 구체적으로, 각 토큰의 평균 어텐션 스코어를 계산하고 임계값 이하의 토큰을 제거함으로써, self-attention뿐 아니라 feed-forward에서도 불필요한 연산을 줄이도록 만든다. 해당 연구는 추가적인 재학습 없이 기존 VLM에 바로 적용할 수 있는 plug-and-play 구조로 설계되었으며, 성능 하락 없이 45%의 FLOPs 감소를 성취한다.

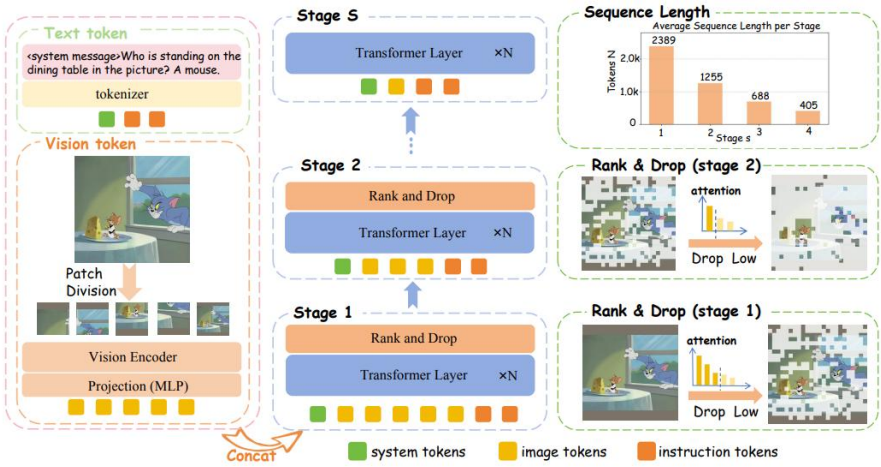
〔그림 4－24〕 레이어에 따른 어텐션 분포도



자료: Chen, et al.(2024), p.6

Xing et al.(2025)는 VLM에서 발생하는 비전 토큰의 과도한 중복문제를 지적하고, 이를 효율적으로 완화하기 위한 계층적 토큰 축소 기법을 제안하였다. 해당 연구는 LLaVA-1.5 및 LLaVA-NeXT와 같은 VLM을 실험적으로 분석한 결과, 모든 비전 토큰이 모든 레이어에서 동등하게 중요하지 않으며, 특히 후반 레이어로 갈수록 비전 토큰의 정보 기여도가 점진적으로 감소한다는 사실을 발견한다. 이는 초반 레이어에서는 전역적 시각 정보를 통합하고, 후반 레이어에서는 특정 정보와 관련된 국소적 영역에 집중한다는 것을 의미한다. 이러한 분석을 바탕으로 제안된 기법은 모델의 레이어를 여러 단계로 분할하고, 각 단계의 마지막에서 시각 토큰의 중요도에 따라 일부 토큰을 점진적으로 제거하는 구조를 갖는다. 토큰 중요도는 텍스트 토큰과 시각 토큰 간의 어텐션 스코어를 이용해 계산되며, 연산량이 매우 작은 유사도 기반 계산으로 수행된다. 이를 통해 초기 레이어에서는 전체 이미지를 충분히 해석하기 위해 모든 비전 토큰을 유지하되, 후반 레이어에서는 점차 불필요한 토큰을 줄여 효율성을 극대화한다〔그림 4－25〕. 해당 연구는 별도의 학습이나 구조 변경 없이 기존 VLM에 plug-and-play 방식으로 적용 가능하며, 학습 및 추론 효율을 동시에 개선한다. 특히 LLaVA-NeXT 모델에 적용한 실험 결과는 학습 시간과 FLOPs를 각각 40%, 55% 줄이면서도 기존 성능을 유지할 수 있음을 보인다.

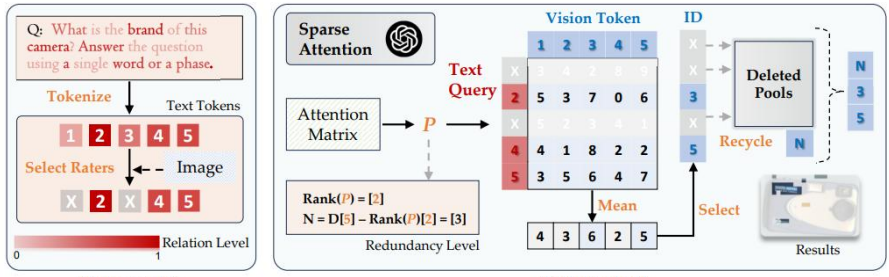
〔그림 4－25〕 PyramidDrop 프레임워크 개요



자료: Xing, et al.(2025), p.6

Zhang et al.(2025)는 텍스트, 비전 간 상호작용을 분석하여, 언어적으로 중요한 비전 토큰만을 남기고 불필요한 토큰을 제거하는 방식의 training-free sparsification 프레임워크를 제시한다〔그림 4－26〕. 이 연구는 먼저 텍스트 토큰과 비전 토큰 간의 correlation strength를 분석하여, 언어적 의미 형성에 실제로 기여할 수 있는 텍스트 토큰을 중요 토큰으로 정의한다. 이후 sparsification priority matrix을 구성하고, 중요 토큰과 밀접한 연관을 가지는 비전 토큰만을 선택적으로 유지한다. 중요도가 낮은 토큰은 rank-based pruning strategy에 따라 제거되며, 프루닝으로 인해 발생할 수 있는 정보 손실은 token reconstruction 과정을 통해 보완된다. 이와 같은 접근은 단순히 토큰 수를 줄이는 압축 기법이 아니라, 텍스트와 비전 간 의미의 일관성을 유지한 채 비효율적인 토큰을 제거하는 구조적 최적화로 볼 수 있다. 실험 결과에서는 VLM baseline 대비 약 37%의 추론 지연 감소, 77.8%의 토큰 압축률, 그리고 97% 이상의 정확도 유지율을 달성한다.

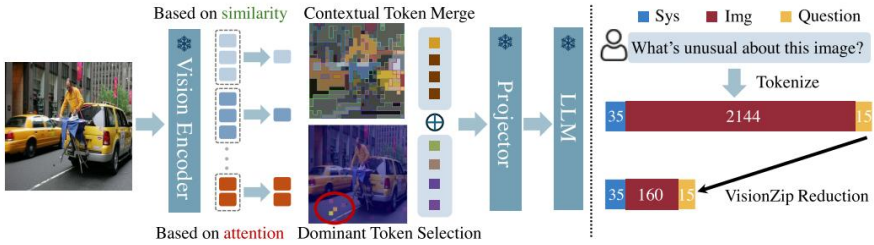
〔그림 4－26〕 SparseVLM 프레임워크 개요



자료: Zhang, et al.(2025), p.4

Yang et al.(2025) VLM에서 이미지 입력이 지나치게 많은 비전 토큰으로 표현됨으로써 발생하는 계산 복잡도와 메모리 비효율 문제를 지적하고, 이를 완화하기 위한 informative visual token selection 기반의 토큰 압축 기법을 제안한다. 해당 연구에서는 CLIP, SigLIP 등 범용 비전 인코더에서 생성되는 비전 토큰의 중복성과 불균등한 정보 분포를 분석한다. 결과적으로, 소수의 비전 토큰이 높은 어텐션 스코어를 가지며 대부분의 이미지 정보를 담당하는 반면, 나머지 다수의 토큰은 낮은 중요도를 보이며 중복된 정보를 반복적으로 포함한다는 점을 발견한다. 이러한 분석을 바탕으로, Dominant visual token selection과 contextual token merging이라는 두 단계로 구성된 training-free redundancy reduction 전략을 제안한다. 우선 비전 인코더의 어텐션 맵을 기반으로 CLS 토큰을 이용해 가장 중요한 비전 토큰들을 선별하고, 그 외의 토큰들은 유사도 기반으로 병합하여 정보 손실을 최소화한다. 이로 인해서 전체 비전 토큰의 수를 크게 줄이면서도 의미적 표현의 풍부함을 유지할 수 있게 된다〔그림 4－27〕. 해당 기법은 별도의 재학습 없이 plug-and-play 형태로 기존 모델에 적용 가능하며, LLaVA-NeXT에서 8배 빠른 prefill 속도와 2배 빠른 추론 속도를 지원하며 동시에 기존 정확도 성능을 유지한다.

〔그림 4－27〕 VisionZip 프레임워크 개요



자료: Yang. et al.(2025), p.3

〈표 4－5〉 VLM 프루닝 기술 보유 현황

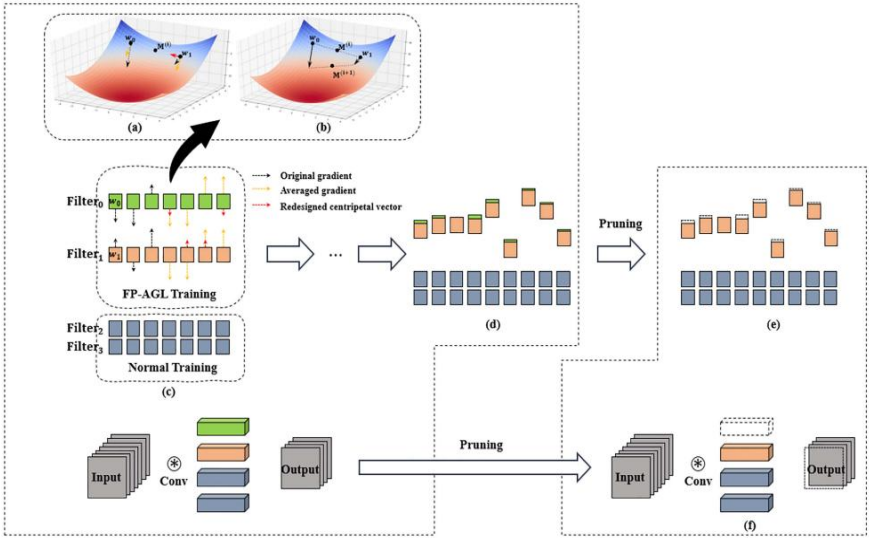
기술 보유국	기술 보유 기관	비전 토큰 압축률	정확도
중국(Chen, et al. 2024)	Peking University	66.7	52.5
중국(Xing, et al. 2025)	University of Science and Technology of China	66.7	56.5
중국(Yang, et al. 2025)	Chinese University of Hong Kong	66.7	57.3
중국(Zang, et al. 2025)	Peking University	66.7	57.7

마. 프루닝 연구 성과

1) FP-AGL: Filter pruning with adaptive gradient learning for accelerating deep convolutional neural networks(IEEE Transactions on Multimedia, 2022)

해당 연구에서는 기존 centripetal stochastic gradient descent(C-SGD) 기반 필터 프루닝 기법이 손실 변화에 대한 고려 없이 모든 파라미터에 구심 벡터를 적용함으로써 성능 저하와 느린 수렴 문제를 야기한다는 점을 지적하였다. 이를 해결하기 위해, 손실 변화를 반영한 1차 테일러 전개 기반의 재설계 구심 벡터(redesigned centripetal vector, RCV)와 원래의 그래디언트 방향을 최대한 유지하며 가중치를 적응적으로 업데이트하는 adaptive gradient learning(AGL) 기법을 제안하였다. 제안된 FP-AGL은 두 필터를 빠르고 정확하게 동일한 값으로 수렴시켜 성능 저하 없이 효율적인 필터 제거가 가능하다〔그림 4－28〕.

〔그림 4－28〕 Filter pruning with adaptive gradient learning

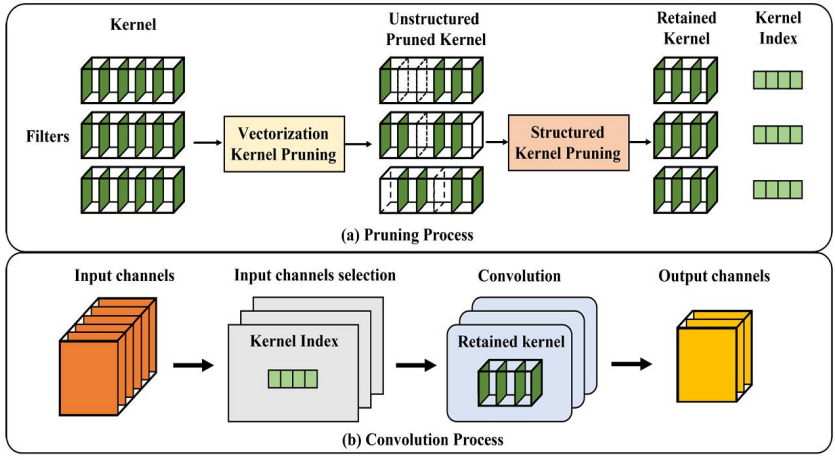


자료: Kim, et al.(2022), p.4

2) V-SKP: Vectorized Kernel-based Structured Kernel Pruning for Accelerating Deep Convolutional Neural Networks(IEEE Access, 2023)

해당 연구에서는 기존 커널 프루닝이 GPU에서 실제 가속 효과를 얻기 어려운 한계를 지적하고, 커널을 벡터로 취급해 크기(L1 norm)와 방향성(cosine similarity)을 함께 고려한 중요도 기반의 vectorized structured kernel pruning(V-SKP) 기법을 제안하였다. 제안 기법은 레이어별 4D 가중치 구조를 유지한 채 동일 비율로 커널을 제거하고, 커널 인덱스 세트를 활용한 kernel index convolution으로 구조화된 프루닝 가중치의 연산을 지원함으로써, 파라미터와 FLOPs 감소 효과를 보였다〔그림 4－29〕.

〔그림 4－29〕 Vectorized Kernel-based Structured Kernel Pruning

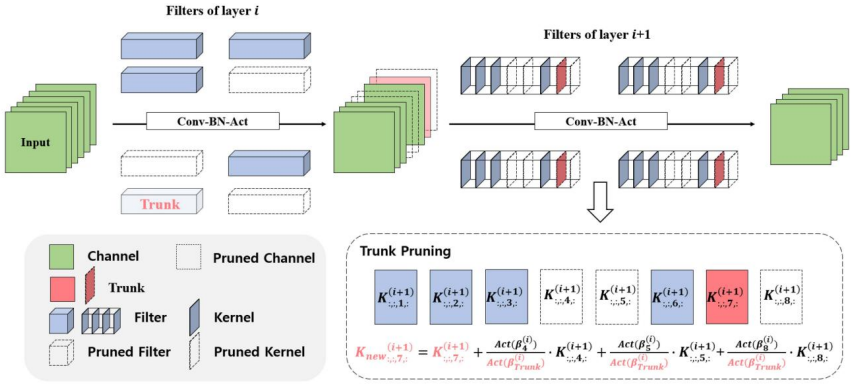


자료: Koo, et al.(2023), p.4

3) Trunk Pruning: Highly Compatible and Versatile Channel Pruning Without Fine-Tuning(IEEE Transactions on Multimedia, 2024)

해당 연구에서는 기존 sparsity training 기반 채널 프루닝에서 batch normalization 스케일링 팩터가 0에 수렴하더라도 shift parameter가 남아 있어 추론 시 왜곡이 발생하고, 이를 보정하기 위해 반드시 미세조정이 필요한 문제를 지적하였다. 이를 해결하기 위해, pruning 대상 채널의 정보를 다음 레이어의 하나의 커널에 흡수하여 unpruned 네트워크의 출력을 그대로 재현하는 Trunk Pruning 기법을 제안하였으며, 이를 통해 미세조정 없이도 성능 저하를 최소화하고, 음수 영역을 갖는 활성화 함수와 높은 호환성을 달성하였다〔그림 4－30〕.

〔그림 4－30〕 Trunk Pruning 개요



자료: Kim, et al.(2024), p.7

바. Merging 개요 및 연구 동향

1) Merging 개요

Merging은 학습된 모델 내에서 방대하고 유사한 성분을 통합하여 효율성을 개선하는 모델 경량화 기법이다. 중복된 표현이나 기능적으로 유사한 파라미터들을 병합함으로써 계산 복잡도와 메모리 사용량을 줄이고, 추론 속도를 개선하며 전력 소모를 절감할 수 있다.

Merging에는 여러 가지 방식이 있는데, 가장 대표적인 방식으로 요소들의 평균으로 병합하는 평균 merging이 있다. 이보다 조금 더 진보한 방식으로 각 요소의 중요도를 반영하여 병합하는 가중 merging이 있다.

또한, 요소들의 특성을 분석하여 공통된 표현 공간으로써 Subspace를 탐색하고 이를 기반으로 각 특성을 반영하여 병합하는 subspace-based 병합도 활발하게 연구되고 있다.

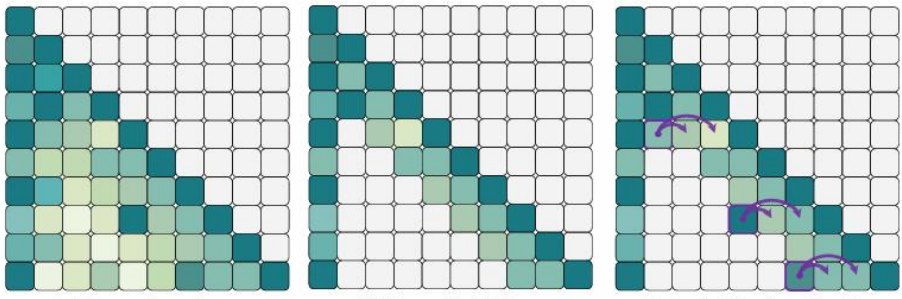
Merging은 일반적으로 적용 대상에 따라 토큰 merging, 가중치 merging, 모델 merging으로 구분한다. 토큰 merging은 입력 토큰 간 중복 정보를 줄여 입력 길이를 단축하고, 가중치 merging은 유사한 가중치를 결합하여 구조를 단순화한다. 모델 merging은 여러 모델의 학습된 지식들을 하나로 통합하여 별도의 재학습 없이 새로운 모델을 생성할 수 있는 효율적인 접근 방식이다. 이처럼 merging은 파라미터 감소뿐 아니라, 모델의 지식 재활용과 범용성 향상에 도움이 되는 대표적인 최신 효율화 기법으로 주목받고 있다.

2) 토큰 merging 기법

토큰 merging은 여러 입력 토큰 중 유사하거나 중복된 정보를 갖는 토큰들을 하나로 병합하여 입력 시퀀스의 길이를 줄이는 경량화 기법이다. LLM은 self-attention 연산의 복잡도가 입력 시퀀스 길이에 따라 지수적으로 증가한다. Decode-only 구조로 인하여 토큰을 순차적으로 하나씩 생성하는 auto-regressive 특성은 각 생성 단계마다 모든 이전 토큰에 대한 어텐션 연산을 반복 수행해야 하므로 전체 연산량을 증가시킨다. 이러한 이유로 최근 연구에서는 불필요하게 반복되거나 정보 중복성이 큰 토큰을 병합하여 연산 부담을 줄이는 KV cache token merging 기법이 제안되고 있다. 이는 저장이 필요한 key-value cache의 크기를 줄여 메모리 사용량을 절감할 뿐 아니라, 매 스텝에서 수행되는 어텐션 연산 자체를 경감시켜 추론 속도를 개선한다.

대표적인 연구로 Zhang, et al.(2024)이 있다. 해당 논문은 LLM 추론 시 급격히 증가하는 KV 캐시 메모리 사용량을 줄이기 위해 기존 단순 캐시 삭제 방식인 eviction 대신, 중요도가 낮은 KV들을 남은 캐시와 가중 병합하여 정보 손실을 최소화한다. 병합 과정에서는 어텐션 스코어나 유사도 기반 가중치를 사용해 출력 변화가 최소화되도록 설계한다. 이 방법은 LLaMA, OPT 등 다양한 모델에서 범용적으로 적용 가능하며, 성능 저하 없이 최대 60%의 메모리 절감 효과를 달성한다. 결과적으로 Zhang, et al.(2024)은 정보 보존형 캐시 압축 기법으로서, 효율적인 LLM 추론의 새로운 방향을 제시한다〔그림 4-31〕.

〔그림 4-31〕 KV-Cache Merging

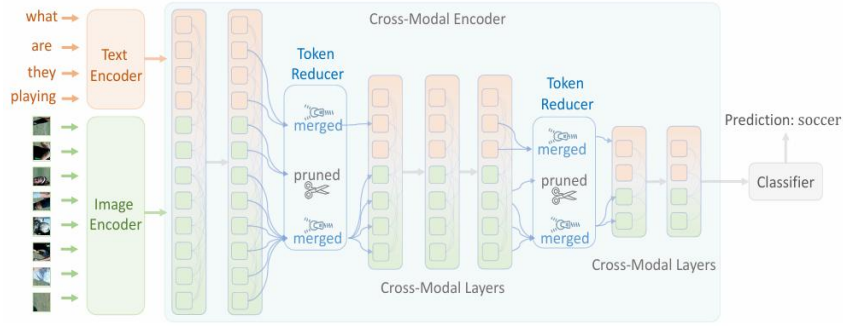


자료: Zhang, et al.(2024), p.2

한편, VLM 에서는 이미지 입력을 패치 단위로 토큰화할 때 발생하는 중복된 시각 정보를 효율적으로 처리하기 위해 비전 토큰 merging이 활용된다. 이미지의 인접 영역은 종종 유사한 시각적 특징을 공유하므로, 이러한 토큰들을 병합하여 더 압축된 형태의 시각 표현을 구성할 수 있다. 이 방식은 불필요한 비전 토큰을 제거하면서도 중요한 시각 정보를 보존하므로, 연산량을 줄이면서도 성능 저하를 최소화할 수 있다.

대표적인 연구로 Cao, et al.(2023)이 있다. 해당 논문은 VLM에서 비전 및 텍스트 토큰을 효율적으로 줄이기 위한 토큰 프루닝과 merging을 합친 기법을 제안한다. 이 기법은 text-informed pruning과 Modality-aware Merging을 결합하여, 중요도가 낮은 이미지 토큰을 제거하고 유사한 시각 및 텍스트 토큰을 병합한다. Cao, et al.(2023)은 경량 token reducer 모듈을 여러 크로스-모달 레이어들에 삽입하여, 별도의 대규모 재학습 없이도 토큰 수를 줄이면서 모델의 정확도 저하를 최소화한다. 실험 결과에서 토큰 병합 및 제거를 통해 추론 처리량을 최대 2배까지 높이고, 메모리 사용량은 절반 이상 줄이면서도 정확도 하락은 1% 미만 수준으로 유지함을 보인다〔그림 4-32〕.

〔그림 4-32〕 PuMer 개요



자료: Cao, et al.(2024), p.4

토큰 merging은 LLM에서는 KV Cache 효율화를, VLM에서는 비전 토큰 압축을 통해 추론 효율성과 메모리 사용 효율을 동시에 개선할 핵심 기법으로 주목받고 있다. 결과적으로, 토큰 merging은 다양한 멀티모달 모델에서도 모델 크기와 연산량의 균형을 유지하면서 성능 저하 없이 효율성을 극대화할 수 있는 전략적 접근으로 확장되고 있다.

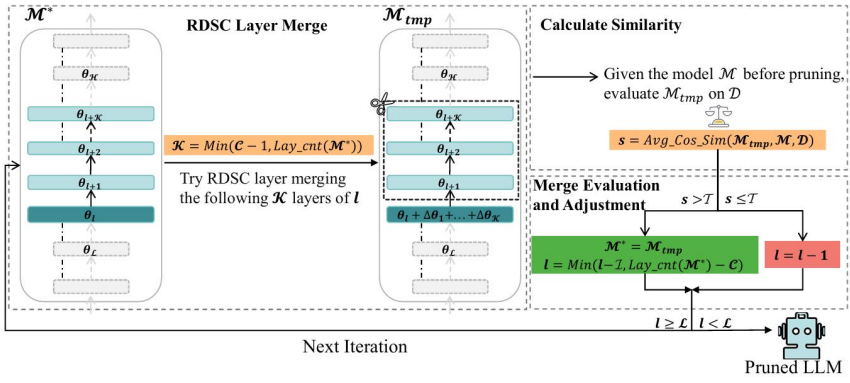
3) 가중치 merging 기법

가중치 merging은 단일 모델 내부의 가중치를 병합하여 모델의 크기를 줄이고, 연산 효율을 높이는 모델 압축 기법이다. 이 방식은 단순히 파라미터를 제거하는 것이 아니라, 유사하거나 중복된 가중치 성분을 통합함으로써 모델의 표현력을 유지하면서도 경량화를 달성한다. 특히 LLM에서는 구조적으로 중복된 정보를 갖는 가중치들이 다수 존재하기 때문에 이들을 효과적으로 병합할 수 있는 알고리즘이 필요하다.

가중치 merging은 크게 inter-layer merging과 intra-layer merging으로 구분할 수 있다. Inter-layer 방식은 서로 인접하거나 유사한 역할을 갖는 여러 레이어들의 가중치를 평균 또는 선형 결합하여 병합하는 방법이다. 이 접근은 모델의 깊이를 줄이면서도, 모델의 후반부의 상위 레이어가 학습한 표현 정보를 보존할 수 있다는 장점이 있다.

대표적인 연구로 Yang, et al.(2024)이 있다. 해당 연구는 transformer 기반의 LLM에서 후반부 레이어들을 이전 레이어로 병합하는 방식의 구조적 레이어 프루닝 기법을 제안한다. 여기서 제안된 핵심 방법은 RDSC reserving-differences-while-Seeking-common layer merge(RSDC)으로, 병합 대상 레이어의 차이를 보존하면서 공통 정보를 이전 레이어에서 처리할 수 있도록 만드는 것이다. 이 방식은 모델의 구조를 크게 변경하지 않으면서도 25-30% 정도의 레이어 제거 비율에도 불구하고, 평균 정확도 성능을 80% 이상 유지하며, 기

〔그림 4－33〕 LaCo 개요

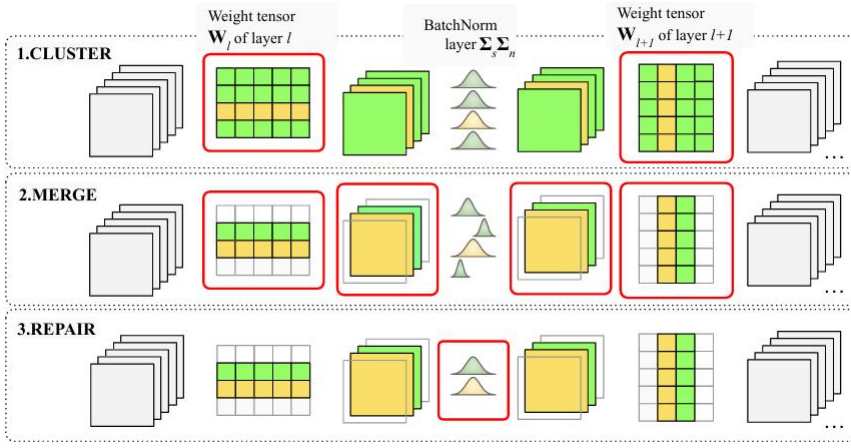


자료: Yang, et al.(2024), p.3

존의 구조적 프루닝 기법들보다 성능 저하가 비교적 없는 것으로 보고되었다. 더욱이, post-training 실험을 통해, 병합된 모델이 원본 모델의 파라미터 특성을 잘 계승됨을 보여주며, 레이어 별 유사성 분석을 통해 병합 가능성의 동기를 논의한다〔그림 4-33〕.

Intra-layer 방식은 하나의 레이어 내부에서 채널 단위나 뉴런 단위의 가중치를 병합하는 방법이다. 이는 과도하게 세분화한 채널 간의 중복 표현을 제거하고, 각 채널의 중요도나 유사도를 기준으로 효율적인 통합을 수행한다. 이러한 병합 과정은 모델의 구조를 유지하면서도 파라미터 수를 줄일 수 있어, 성능 저하 없이 계산량과 메모리 사용을 감소시킨다.

〔그림 4-34〕 Model Folding pipeline



자료: Wang, et al.(2025), p.2

대표적인 연구로 Wang, et al.(2025)이 있다. 해당 연구는 데이터나 fine-tuning 없이도 모델을 효율적으로 압축할 수 있는 model folding 기법을 제안한다. 이 방법은 여러 레이어에 분포된 유사한 뉴런들을 중첩시켜 네트워크의 구조를 단순화하면서도 모델의 통계적 특성을 유지한다. 저자들은 이를 위해 k-means 클러스터링 기반의 뉴런 병합과 통계 손실을 최소화하는 variance repair 절차를 적용한다. 또한 데이터가 전혀 없는 상황에서도 작동할 수 있도록 Fold-Approximate REPAIR(AR) 및 Fold-Deep Inversion 기반 REPAIR(DIR)

두 가지 변형을 제시한다. 다양한 벤치마크에서 제안된 방법론을 실험한 결과, ResNet18이나 LLaMA-7B 같은 모델에서 fine-tuning 없이도 기존 압축 기법과 유사한 성능을 유지하며, 데이터 접근이 제한된 환경이나 edge device에서의 모델 경량화에도 효과적인 것으로 나타났다〔그림 4-34〕.

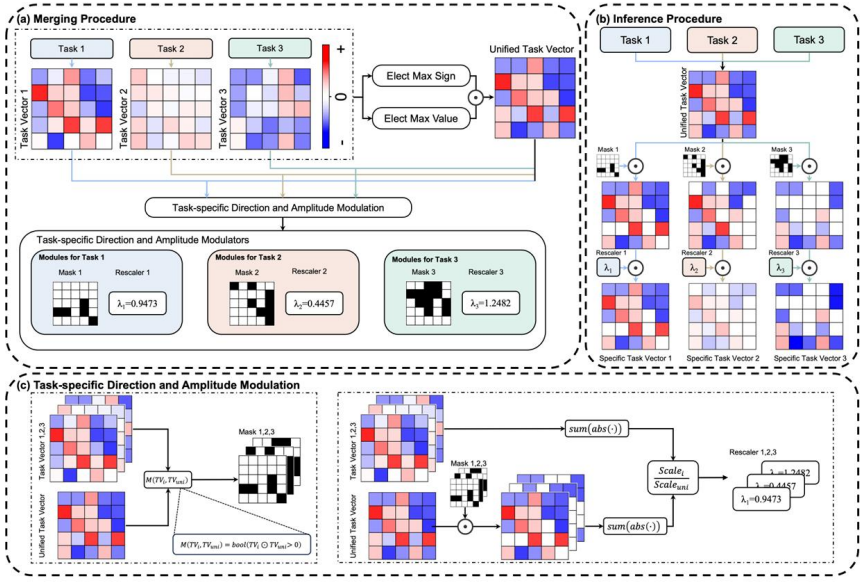
4) 모델 merging 기법

모델 merging은 여러 개의 개별 모델을 하나의 통합된 모델로 결합하여, 각 모델이 가진 지식과 기능을 동시에 활용할 수 있도록 하는 기법이다. 이러한 접근은 여러 태스크나 도메인에 특화된 전문가 Experts을 병합함으로써, 별도의 재학습이나 추가 데이터 없이도 멀티태스크 처리 능력을 확보할 수 있다는 장점이 있다. 특히 최근에는 LLM 간의 가중치 merging 연구가 활발히 이루어지고 있다. 이를 통해 여러 모델의 지식을 하나의 네트워크로 통합하려는 시도가 늘고 있다. 단순한 평균 merging에서부터, 각 모델의 분포적 차이를 최소화하는 정렬 기반 merging, 혹은 low-rank 변환을 이용한 LoRA 기반 merging까지 다양한 방법이 제안되고 있다.

이러한 모델 merging 전략들은 기존의 모델 ensemble과 달리, 여러 모델의 출력을 결합하는 것이 아니라, 실제 파라미터 공간에서 모델들을 통합한다는 점에서 계산 효율성이 높다. 그 결과, 모델 크기는 증가하지 않으면서도 여러 전문가 모델이 가진 지식을 하나의 네트워크 안에서 재사용할 수 있다. 또한 mixture of experts(MoE) 구조와 결합할 경우, 병합된 모델이 상황에 따라 특정 전문가의 파라미터를 동적으로 활성화함으로써 더 효율적인 추론이 가능해진다. 이러한 특징 덕분에 모델 merging은 멀티태스크 학습, 지식 전이, 도메인 적응 등 다양한 영역에서 핵심적인 경량화 및 범용화 기법으로 주목받고 있다.

대표적인 연구로 Huang, et al.(2024)이 있다. 해당 논문은 서로 다른 태스크로 미세조정한 여러 모델을 추가 학습이나 데이터 없이 하나의 통합 모델로 병합하는 방법을 제안한다. 먼저 여러 모델의 가중치를 바탕으로 통합된 모델을 elect하고, 이후 task-specific mask와 rescaler로 불리는 경량 조정자를 생성해 방향과 크기를 조절함으로써 병합된 모델이 각 태스크 모델을 잘 근사하도록 만든다. 이 방식은 tuning-free, 즉 추가 데이터나 미세조정 없이 병합이 가능하다는 특징을 갖고 있다. 실험 결과, 다양한 비전 모델, NLP 모델, PEFT 모델, 멀티모달 모델 병합에서도 기존 병합 방법 대비 성능 저하가 거의 없고, 높은 병합 효과를 보인다〔그림 4-35〕.

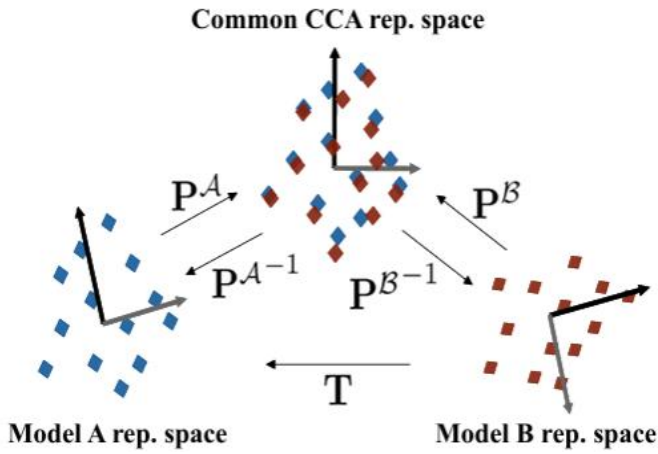
〔그림 4－35〕 EMR-Merging 개요



자료: Huang, et al.(2024), p.3

다른 대표적인 연구로 Horoi, et al.(2024)이 있다. 해당 연구는 여러 신경망 모델을 단일 모델로 병합할 때, 기존의 뉴런 permutation 방식이 갖는 제약을 극복하기 위해 canonical Correlation analysis(CCA) 기반 merging 기법을 제안한다. 구체적으로, 각 모델의 내부 표현을 CCA로 permutation 할 수 있는 공통 표현 공간을 찾고, 이를 이용해 뉴런 간 1:N 대응을 허용하는 병합 변환을 계산한다. 이 방식은 단순 평균 방식보다 더 유연하게 모델 간 표현 차이를 줄이며, 여러 모델(2개 이상)을 동시에 병합하는 상황에서도 우수한 성능을 보인다. 실험 결과, 동일 데이터나 서로 다른 데이터 분할 모델을 병합할 때 기존 방식 대비 낮은 손실로 성능을 유지하면서 병합 효과를 개선함을 보인다〔그림 4－36〕.

〔그림 4－36〕 CCA Merge 시각화



자료: Horoi, et al.(2024), p.1

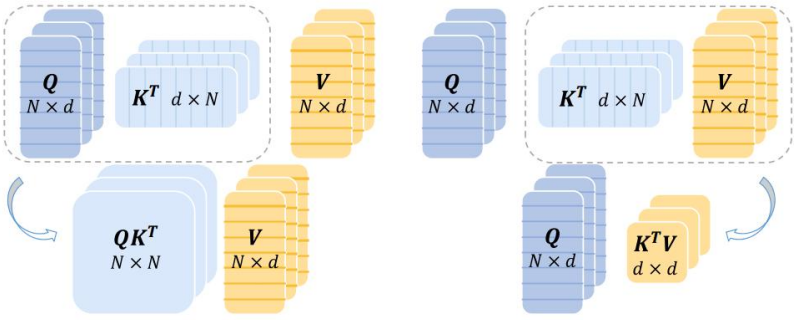
2. 에너지 지향적 알고리즘 최적화

가. 에너지 지향적 알고리즘 최적화의 필요성

최근 다양한 분야에서 활용되는 Transformer 기반의 대규모 AI 모델은 여러 연산 모듈로 구성되며, 그중 핵심 구조인 attention 연산은 높은 연산 복잡도를 지녀, 모델 규모가 커질수록 막대한 연산 자원과 메모리 대역폭을 요구하게 된다. 이러한 구조적 특성은 학습과 추론 전반에서 높은 전력 소모를 유발하며, 이는 대규모 데이터센터의 운영 비용 증가뿐 아니라 탄소 배출량 증가와 같은 환경적 부담으로 이어진다. 이러한 문제를 해결하기 위해 최근에는 단순한 모델 경량화뿐만 아니라, 연산 복잡도가 높은 attention 연산을 단순화하고 효율적으로 재구성하는 다양한 방법이 연구되고 있다. 그중에서도 linear attention은 기존 attention 메커니즘을 수학적 변환과 연산 재배치를 통해 효율적으로 구현하는 방법으로 주목받고 있다〔그림 4－37〕. Linear attention 기법은 입력 토큰 간 관계 계산 과정에서 불필요한 연산과 중복 메모리 접근을 줄여, 동일 하드웨어 환경에서도 자원 소모를 완화하고 전력 소비 및 탄소 배출을 감소시킬 수 있다. Linear attention의 가장 큰 장점은 최소한의 성능 저하로 연산량과 메모리 요구량을 동시에 줄일 수 있다는 점이다. 이는 단순한 처리 속도 향상을 넘어, 데이터센터와 같은 대규모 인프라 환경에서는 attention 연산

과정에서 발생하는 데이터 전송을 최소화하여 에너지 효율성을 향상한다.

〔그림 4－37〕 Attention 연산 패턴에 따른 연산 복잡도 비교



자료: Han, et al.(2024), p.1

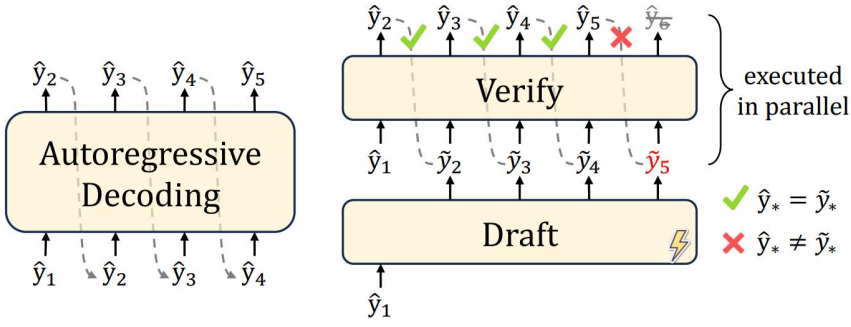
한편, AI 모델 추론의 에너지 효율성을 극대화하기 위한 방법으로 Neural Architecture Search(NAS) 기법이 주목받고 있다. NAS는 사람이 직접 설계한 고정된 네트워크 구조 대신, 주어진 자원 제약과 목표를 고려하여 최적의 모델 구조를 자동으로 탐색하는 기법이다. 특히, 에너지 지향적 NAS는 연산량과 메모리 대역폭 사용을 최소화하면서도 성능을 유지하거나 향상시킬 수 있는 구조를 찾는 데 중점을 둔다. 이러한 과정에서, 탐색 알고리즘은 다양한 연산 블록, 레이어 구성, 데이터 플로우 방식을 조합해 수많은 후보 아키텍처를 평가하며, 그중 에너지 효율이 가장 높은 설계를 선택한다. 이렇게 도출된 구조는 하드웨어 환경에서 더 적은 전력과 메모리로 운용이 가능하며 데이터 전송 및 연산 과정에서의 탄소 배출을 절감한다. 결과적으로, NAS는 단순히 모델의 정확도 향상뿐 아니라, 에너지 효율성을 최우선으로 하는 차세대 AI 모델 설계에 있어 필수적인 도구로 자리잡고 있다.

Linear attention, NAS가 아키텍처 차원의 에너지 절감을 이끈다면, 추론 시점의 동적 비효율을 다루기 위해서는 런타임 정책이 필요하다. 실제 서비스 트래픽에서는 입력 길이·도메인이 모든 요청마다 달라지며, 고정된 경로로 모든 레이어들을 끝까지 통과시키는 방식은 불필요한 연산과 메모리 이동을 초래한다. 그러므로 동일한 모델 구조라도, 토큰 생성 과정에서 얼마나 자주 모델을 부를 것인가와 한 번 부를 때 얼마나 깊게 레이어들을 통

과시킬 것인가를 상황에 맞게 조절하는 adaptive decoding이 에너지 지향 최적화의 대안으로 주목받고 있으며, 크게 두 가지로 분류할 수 있다.

Speculative decoding은 소형 draft 모델이 후보 토큰을 초안으로 미리 제시하고, verify 모델이 한 번의 검증으로 여러 토큰을 확정하여 검증 모델 호출 수를 줄이는 기법이다〔그림 4-38〕. 이를 통해 파라미터 수가 많은 검증 모델을 매번 토큰을 생성할 때마다 호출하지 않아도 되어, 기존 LLM 대비 원본 모델 호출 횟수가 크게 줄어든다.

〔그림 4-38〕 Speculative decoding 개요



자료: Xia, et al.(2022), p.1

Speculative decoding을 적용했을 때, 검증 모델 한번 호출에서 확정되는 토큰 개수의 기댓값은 다음과 같이 정의한다:

$$E(tokens) = \frac{1-a^{\gamma+1}}{1-a}$$

여기서 초안 길이 γ는 draft 모델이 한 번에 미리 제안하는 토큰의 개수를 뜻하며, 수락률 α는 제안된 초안 가운데 verify 모델이 실제로 받아들이는 비율로, draft 품질이 높을수록 수락률도 높아진다. 수락률이 높을수록, 그리고 draft 길이가 길수록 검증마다 확정 토큰 수가 증가해 verify 모델 호출 수가 빠르게 감소한다. 또한 draft 모델의 토큰당 연산량을 검증 모델의 토큰당 연산량으로 나눈 비율을 c라고 정의할 때, 이론적인 연산량 변화량은 다음과 같이 정의한다:

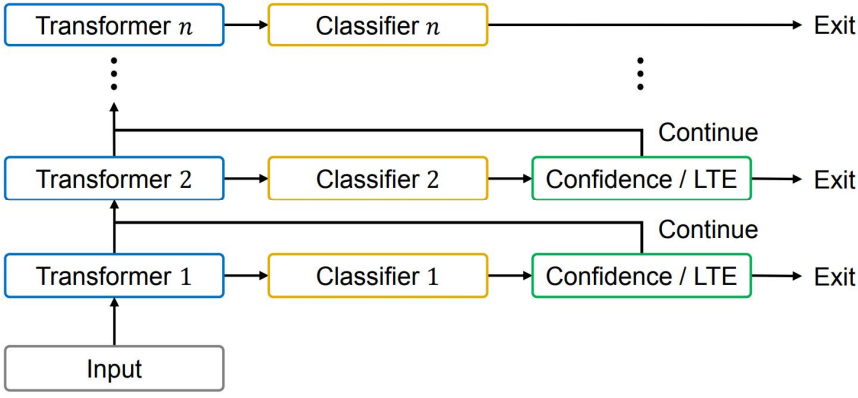
$$OpsFactor = \frac{(1-\alpha)(\gamma c+\gamma+1)}{1-\alpha^{\gamma+1}}$$

수락률 α에 따라 총 연산량은 증가할 수 있지만, 파라미터 수가 훨씬 많은 verify 모델 호출 수가 감소한다. 호출 수가 줄면 LLM 추론 과정에서 필요한 데이터(e.g., 가중치, K/V 캐시)의 메모리 이동 횟수가 선형적으로 감소해 메모리 대역폭 병목이 완화되며, 전체 전력에선 이득된다. 또한 draft 모델의 연산은 상대적으로 적어서, verify 모델 호출 감소 이득이 성립하는 영역이 넓다.

한편, 최근 LLM의 디코딩 과정은 통상적으로 매 토큰 생성 시 모든 디코더 레이어들을 순차 통과하는 고정 경로를 따른다. 그러나 실제로는 입력의 난이도와 맥락 정보가 토큰마다 일정하지 않아서 일부 토큰들은 중간 레이어에 도달할 때 이미 충분한 목표 분포, 즉 정상적인 출력을 위해 수렴한 상태에 도달한다. 여기서 말하는 난이도는 모델이 해당 토큰들에 얼마나 확신을 가지고 예측할 수 있는지를 가리키는 개념이다. 만약 문맥이 충분하고 다음 단어가 거의 정해져 있을 때 모델의 출력 분포는 한 후보에 확률이 몰리고, 반대로 의미가 모호하거나 장거리 의존성이 필요한 경우에는 여러 후보로 확률이 분산된다. 여기서, 전자를 난이도가 낮다고 표현하고 반대로 후자를 난이도가 높다고 표현한다. Early-exit은 이 점에 착안하여, 모델이 충분히 확신하고 있다고 판단되는 지점에서 남은 레이어들의 계산을 생략하고 다음 토큰 예측으로 즉시 넘어가는 동적 깊이 제어 기법이다〔그림 4-39〕. 여기서 확신은 가장 높은 후보의 확률인 top-1, top-1과 top-2의 확률 간 차이, entropy, 또는 예측기(e.g., lightweight multi-layer perceptron)가 산출한 정지 확률 등의 간단한 신호로 근사한다. 중요한 점은 이 판단이 레이어 내부의 복잡한 표현을 다시 학습하지 않고도 빠르게 계산되어야 한다는 것이며, 판단이 내려지면 해당 토큰에 대한 추론의 잔여 블록은 전부 건너뛰어 해당 블록에서 수행될 연산뿐 아니라 KV 캐시의 생성과 저장도 발생하지 않는다. 즉, early-exit은 FLOPs 절감과 메모리 이동 감소를 동시에 유도하는 메커니즘이다. Early-exit은 특히 자기회귀 생성의 구조적 특성과 잘 맞는다. 확신이 빠르게 형성되는 토큰들이 연속해서 나타날 경우, 평균 통과 레이어 수가 눈에 띄게 낮아져 토큰당 지연이 줄어든다. 문서가 길수록, 그리고 맥락이 충분히 누적될수록 초반 대비 후반 토큰에서 확신이 더 빨리 커지는 경향이 나타나는데, 이는 긴 컨텍스트 디코딩에서 early-exit의 누적 이득이 커지는 이유를 설명한다. 반대로 난이도가 높은 토큰이나 정보가

모호한 위치에서는 깊이 절감폭이 줄어든다. 이 때는 품질을 우선시해 마지막 레이어까지 계산을 수행한다. 따라서 early-exit은 입력 난이도에 따라 경로 길이를 조절하는 기법에 가깝고, 전 구간에서 동일한 계산을 강제하는 기존의 구조보다 효율적이다.

〔그림 4－39〕 Early-exit 개요



자료: Xin, et al.(2021), p.1

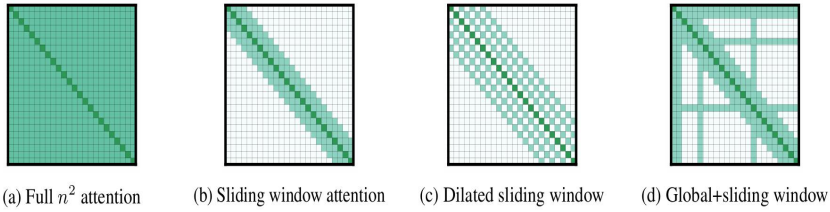
이러한 런타임 경로 단축 기법들은 대표적인 경량화 기법인 양자화와도 호환된다. 양자화가 정밀도를 낮춰 모델 자체의 연산·메모리 비용을 낮춘다면, speculative decoding과 early-exit은 동일 모델 위에서 모델 호출 횟수와 호출당 깊이를 줄여 실행 경로 길이를 단축한다. 결과적으로 동일 품질 조건에서 양자화를 함께 적용하면 지연과 에너지를 크게 절감할 수 있고, 앞서 설명한 기법들인 linear attention, NAS와 함께 결합이 가능해져서 운영 비용과 탄소 발자국을 크게 줄일 수 있는 기법으로 각광받는 추세이다.

나. Linear attention 기법의 연구 동향

Beltagy et al.(2020) 연구에서는 기존 Transformer의 attention 연산이 시퀀스 길이에 대해 $O(n^2)$ 으로 확장되는 구조적 한계로 인해 긴 입력 시퀀스를 처리하기 어렵다는 문제를 지적하고, 이를 해결하기 위해 Longformer를 제안하였다. Longformer는 〔그림 4－40〕의(d)에서 볼 수 있듯이, 로컬 윈도우 기반의 슬라이딩 어텐션과 작업 특화 global attention을 결합하여 메모리 및 연산 복잡도를 시퀀스 길이에 선형적으로 증가시키는 구

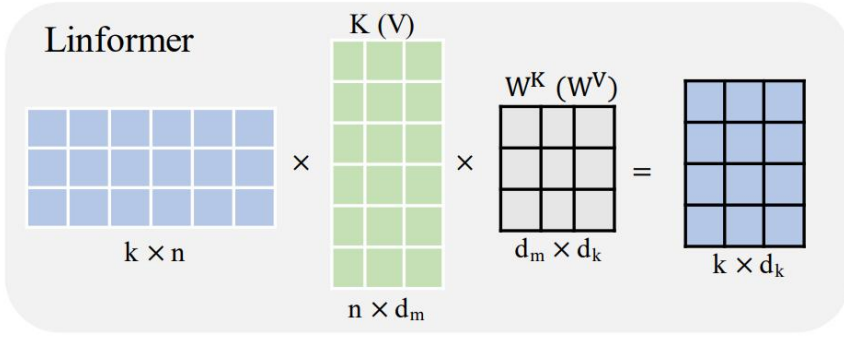
조를 제안하였다. 결과적으로, 사전 학습된 RoBERTa 모델에 적용했을 때 긴 입력 시퀀스 처리 작업에서 우수한 성능 향상을 보였다. 결국, GPU 메모리 사용량이 크게 줄어 동일 하드웨어에서 더욱 긴 시퀀스를 처리할 수 있고, 메모리 대역폭 병목을 완화해 추론 에너지 효율성을 높였다.

〔그림 4－40〕 Attention 별 연산 패턴 비교



자료: Betagy, et al.(2024), p.3

〔그림 4－41〕 Linformer attention 연산 구조



자료: Wang, et al.(2020), p.5

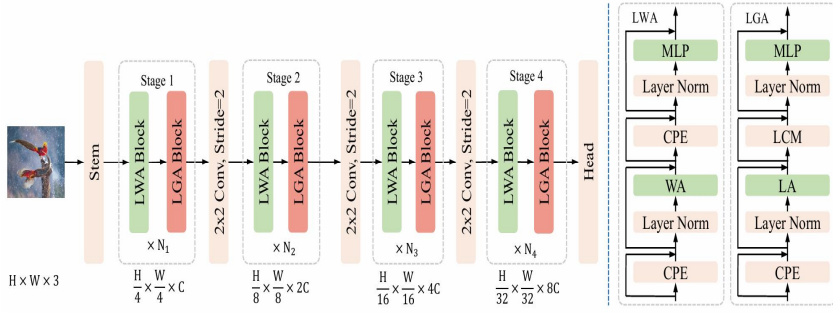
Wang et al.(2020) 연구에서는 attention 행렬이 low-rank 구조를 가진다는 관찰에 기반해, $O(n^2)$복잡도의 attention 연산을 $O(n)$으로 줄이는 Linformer를 제안하였다. Linformer는 〔그림 4－41〕에서 볼 수 있듯이 Query-Key 내적 행렬을 저차원으로 행렬로 projection하여 전체 attention을 low-rank 구조로 근사함으로써, 메모리 사용량과 연산량을 동시에 절감하였다. 결과적으로 사전 학습된 모델과 미세 조정 과정에서 모두 기존 Transformer

와 유사한 성능을 유지하면서 학습 및 추론 속도 및 에너지 효율성을 향상했다.

Choromanski et al.(2021) 연구에서는 소프트맥스 기반 attention을 양의 직교 랜덤 피처를 활용한 FAVOR+ 메커니즘으로 근사하는 Performer를 제안하였다. 이는 희소성이나 low-rank 가정 없이도 선형 시간, 공간 복잡도로 attention을 근사할 수 있도록 하며, 편향 없는 근사와 낮은 분산을 이론적으로 보장한다. 결과적으로 모든 attention 연산은 커널 매핑 후 단순 행렬곱으로 변환되어, 병렬화 효율이 높고 GPU/TPU의 SIMD 연산 유닛 활용률을 극대화할 수 있다.

Zheng et al.(2025) 연구에서는 ViT 모델에 linear attention 적용 시 발생하는 이미지 토큰 집중성 부족 문제를 개선하기 위해, ReLU 기반 특징 맵핑과 로컬 집중 모듈을 결합한 L2ViT를 제안하였다. 〔그림 4-42〕에서 볼 수 있듯이, L2ViT 모델은 Global window attention과 local window attention을 교차 배치하여 정보 표현력을 확보하면서도 선형 연산 복잡도를 유지하였다. 결국, local window attention 단위의 연산 분할을 통해 캐시 효율이 개선되고, 메모리 접근 범위가 줄어 GPU 연산 시 데이터 이동 병목 및 전력 소모를 최소화할 수 있는 방법이다.

〔그림 4-42〕 L2ViT Attention 연산 및 모델 구조



자료: Zheng, et al.(2025), p.5

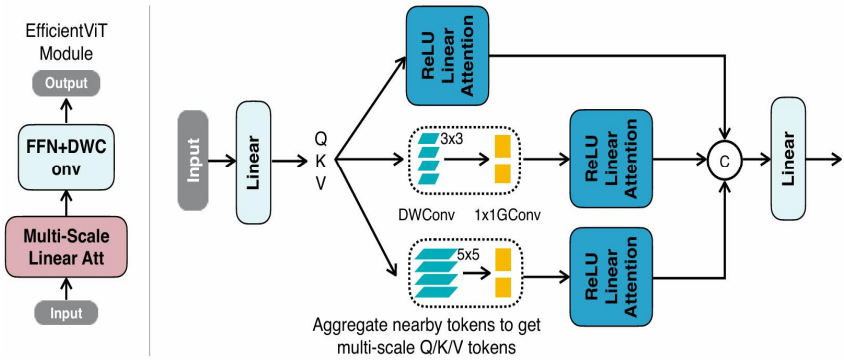
Han et al.(2025) 연구에서는 linear attention의 성능 저하 원인을 토큰 간 집중 능력 부족과 특징 다양성 감소로 분석하고, 이를 해결하기 위해 단순화된 맵핑 함수와 랭크 복원

모듈을 도입한 FLatten Transformer를 제안하였다. 이는, Query-Key 특성을 조정하고 depth별 합성곱을 통해 표현 다양성을 복원함으로써, 낮은 복잡도와 높은 성능을 동시에 달성하였다. 결과적으로, 연산 파이프라인에서 불필요한 데이터 종속성이 줄어들어 병렬 처리 효율이 높아지고, 메모리 접근 패턴이 단순화되어 GPU/TPU 캐시 히트율이 개선된다.

Katharopoulos et al.(2020)에서는 기존 Transformer의 self-attention이 시퀀스 길이에 따라 제곱으로 증가하는 연산 복잡도를 가지는 구조적 한계를 지적하며, 이를 해결하기 위해 attention 메커니즘을 커널 기반의 선형 연산 구조로 재해석하는 접근을 제안하였다. 해당 연구에서는, 토큰 간 유사도 계산을 확률 분포로 직접 구성하는 대신, 이를 특정한 형태의 양의 정의 커널로 근사함으로써, 연산 순서를 재배치할 수 있는 수학적 토대를 마련하였다. 이 접근은 토큰 쌍 간의 개별 연산을 모두 수행하는 방식이 아니라, 토큰 집합의 특징 표현을 먼저 구성한 뒤 이를 효율적으로 결합하는 방식으로 동작하여, 전체 연산량과 메모리 사용량을 시퀀스 길이에 비례하는 수준으로 낮춘다. 이러한 선형화 기법은 모델의 동작 원리를 유지하면서도, 긴 입력 시퀀스를 처리할 때 발생하는 연산 및 메모리 병목을 효과적으로 완화할 수 있는 이론적 근거를 제공하며, 결과적으로 더 적은 연산 자원과 전력으로도 대규모 입력 처리가 가능하도록 하여 에너지 효율성을 향상한다.

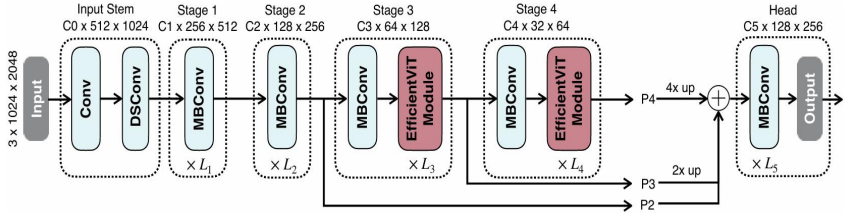
Cai et al.(2024) 에서는 기존 소프트맥스 attention 연산에서 발생하는 메모리, 연산 병목 현상을 해결하기 위해, ReLU 기반 linear attention 구조를 설계했다 〔그림 4－43〕. 추가적으로, 〔그림 4－44〕에서 볼 수 있듯이, 단순한 depthwise, pointwise convolution 연산을 결합하여 토큰 정보를 여러 스케일에서 집계하여, 멀티스케일의 이미지 특징들을 효과적으로 학습할 수 있는 transformer + convolution 형태의 하이브리드 구조인 EfficientViT 모델을 제안하였다. EfficientViT 모델은 전역적인 문맥 수용영역과 다중 스케일 표현 학습을 유지하면서도, 연산 복잡도를 시퀀스 길이에 비례하는 선형 수준으로 낮추고, 하드웨어 병렬화 효율을 높였다. 결국, 메모리 사용량을 크게 줄이고, 기존 ViT 대비 에너지 효율성을 크게 향상시켰다.

〔그림 4－43〕 EfficientViT의 ReLU 기반 linear attention 연산 모듈



자료: Cai, et al.(2024), p.3

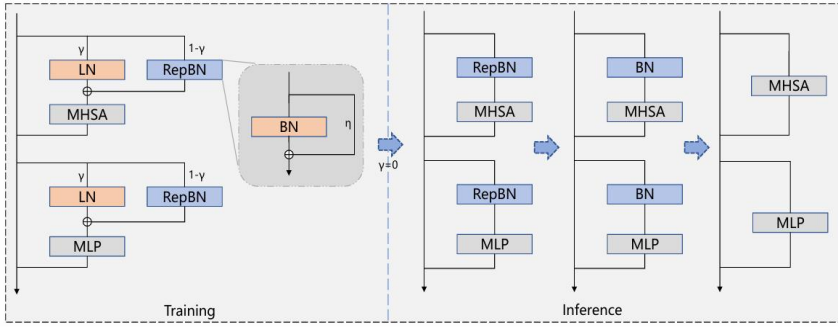
〔그림 4－44〕 EfficientViT 모델 구조



자료: Cai, et al.(2024), p.4

Guo et al.(2024) 에서는 트랜스포머 모델의 학습, 추론 효율을 높이기 위해, 학습 과정에서 layer norm을 점진적으로 batch norm 형태로 치환하는 progressive reparameterized batch norm(PRepBN) 기법을 제안했다. 이를 통해 학습 안정성을 유지하면서도, 〔그림 4－45〕에서 볼 수 있듯이 추론 시 불필요한 통계 계산을 제거하였다. 또한 attention 연산에서 소프트맥스와 복잡한 행렬곱 대신 ReLU 커널과 Depthwise Convolution 기반의 최적화구조를 사용했다. 결국, 메모리 접근 비용과 FLOPs를 크게 줄이면서도, 기존 방법 대비 정확도 손실을 최소화하였다.

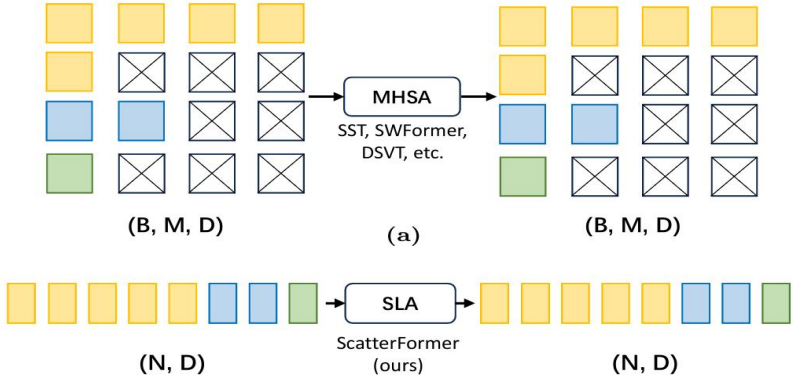
〔그림 4－45〕 Progressive reparameterized batchnorm 동작 구조



자료: Guo, et al.(2024), p.3

He et al.(2024)에서는 기존 window attention 구조의 비효율성을 극복하기 위해 scattered linear attention(SLA)을 제안했다〔그림 4－46〕. SLA 방법은 패딩이나 윈도우 정렬 없이 입력 장면 전체를 하나의 시퀀스로 처리할 수 있게 하였다. 또한 대규모 행렬곱을 작은 청크 단위로 분해하는 커널 최적화 기법을 도입해, 가변 길이 시퀀스를 효율적으로 처리했다. 결국, 불규칙한 토큰 배치에서도 성능이 안정적으로 유지되며, 실제 하드웨어에서의 효율성이 향상되었다.

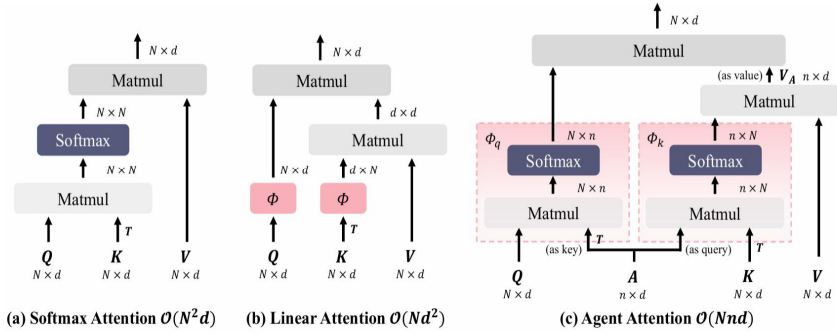
〔그림 4－46〕 일반적인 attention 연산과 scattered linear attention의 비교



자료: He, et al.(2024), p.2

Han et al.(2024)는 대규모 시퀀스 모델의 어텐션 연산 효율을 극대화하기 위해 sliding window attention과 chunk-wise recurrence를 결합한 agent attention 방법을 제안했다〔그림 4-47〕. Agent attention은 입력 시퀀스를 고정 크기 청크로 나눈 뒤, 각 청크에서 local attention을 수행하고, 인접 청크 간의 상태를 순환적으로 전달함으로써 global 문맥 정보를 유지한다. 이를 통해 기존 attention 대비 메모리 사용량이 입력 시퀀스 길이에 무관하게 일정하며, 연산 복잡도도 시퀀스 길이에 선형적으로 변화하게 되어, 매우 긴 시퀀스 에서도 메모리 사용량을 최소화하며 빠른 추론이 가능하다.

〔그림 4-47〕 Attention 패턴 별 연산 과정 및 복잡도 비교



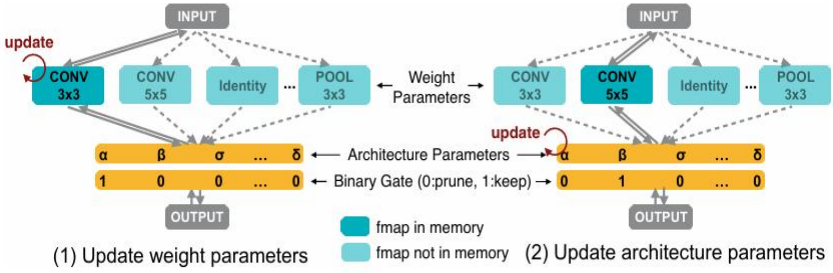
자료: Han, et al.(2024), p.3

다. Neural Architecture Search 기법의 연구 동향

Tan et al.(2019) 연구에서는 모바일 환경에서의 CNN 설계 시 정확도와 지연 시간 간의 균형을 자동으로 맞추기 위해, 실제 기기에서의 추론 속도를 직접 측정하는 다중 목적 최적화 기반 automated mobile neural architecture search(MNAS) 기법을 제안하였다. 해당 연구에서는, 기존 FLOPs 기반 연산 복잡도 측정 방법이 기기별 지연 시간과 불일치하는 문제를 해결하기 위해, 네트워크 탐색 목표 함수에 실측 지연 시간을 포함시켰다. 또한 동일한 레이어 구조를 반복 사용하는 대신, 네트워크 전반에서 다양한 레이어 구성을 허용하는 계층적 분해 검색 공간을 설계하여 연산 효율을 높였다. 결과적으로, 제안된 MnasNet은 모바일 플랫폼에서 정확도-지연 시간 최적화를 동시에 달성할 수 있는 효과적인 방법임을 보였다.

Cai et al.(2019) 연구에서는 기존 NAS가 소규모 데이터셋 또는 프록시 태스크로만 탐색하는 한계를 극복하고, 대규모 데이터셋과 실제 하드웨어 환경에서 직접 탐색하는 ProxylessNAS를 제안하였다. 모든 후보 경로를 포함하는 과대매개망을 구성하되, GPU 메모리 폭증 문제를 해결하기 위해 경로를 이진화하여 한 번에 하나의 연산만 활성화하도록 설계하여 탐색 과정에서의 전력 소모를 최소화 하였다. 또한 이를 통해 메모리 사용량과 탐색 시간을 일반 학습 수준으로 낮추면서도 대규모 탐색 공간을 유지할 수 있었다. 추가적으로, 미분이 불가능한 하드웨어 지표를 최적화하기 위해 지연 시간 모델링을 정규화 항으로 포함하거나 강화학습 기반 최적화를 적용하였다. 결과적으로, GPU, CPU, 모바일 환경별 맞춤형 구조 설계가 가능함을 입증하였다〔그림 4-48〕.

〔그림 4-48〕 Binarized Architecture Parameter Search

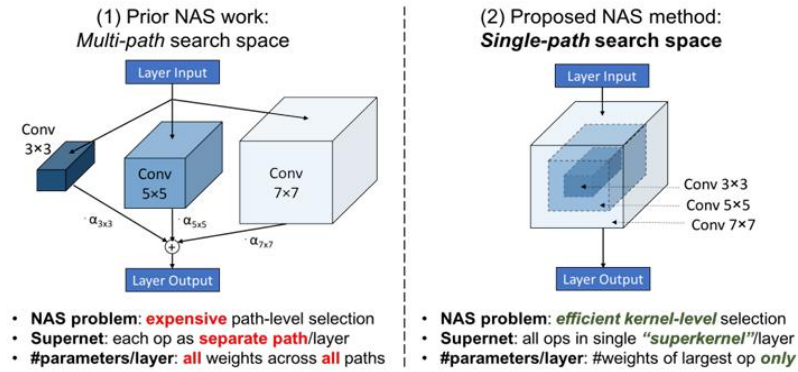


자료: Cai, et al.(2019), p.4

Stamoulis et al.(2019) 연구에서는 기존 다중 경로 기반 NAS의 메모리 및 연산 비효율을 개선하기 위해, 모든 후보 연산을 단일 커널 경로로 통합한 single-path NAS를 제안하였다〔그림 4-49〕. 제안하는 방법은 각 레이어에서 후보 연산별 별도 파라미터를 유지하는 대신, 커널 가중치의 부분집합 선택만으로 구조를 탐색하여 학습 파라미터 수와 탐색 비용을 획기적으로 줄였다. 이를 통해 탐색 시간을 4시간 미만으로 단축하여 탐색 과정에서 발생할 수 있는 탄소 배출 및 전력 소모를 획기적으로 줄이는 방법을 제안하였다. Wu et al.(2019) 연구에서는 모바일 기기 환경에서 정확도와 지연 시간을 동시에 최적화하기 위해, 기기 지연 시간을 직접 반영하는 differentiable NAS(DNAS) 기반의 FBNet을 제안하였

다. 기존 강화학습을 기반으로 한 NAS 방법은 수천 개 아키텍처를 반복 학습해야 하는 비효율 문제를 가지고 있으며, 이를 해결하기 위해 Gumbel-Softmax를 사용해 아키텍처 선택을 확률 분포로 연속화하여 경사 하강법 기반의 학습이 가능하도록 했다. 이를 통해 탐색 비용을 기존 방법 대비 약 420배 절감할 수 있었다.

〔그림 4－49〕 Multi-path NAS와 single-path NAS의 비교



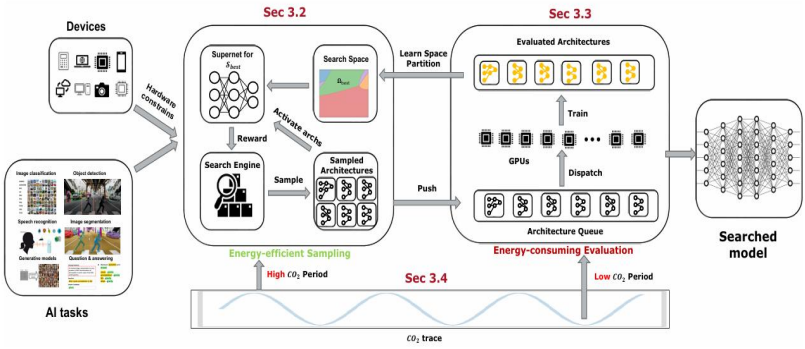
자료: Stamoulis, et al.(2019), p.2

Bakhtiarifard et al.(2024) 연구에서는 NAS 연구에서 간과되기 쉬운 에너지 소비 문제를 해결하기 위해, 에너지 사용량을 포함한 탭형 벤치마크 EC-NAS를 제안하였다. NAS-Bench-101 구조를 기반으로 GPU, TPU 학습 시의 전력 소비를 측정하여, 후보 아키텍처별 에너지 프로파일을 데이터베이스화하였다. 또한 모든 아키텍처를 직접 측정하는 비효율을 줄이기 위해, 일부 샘플만 측정한 후 multi-layer perceptron(MLP) 기반 대리모델로 나머지 구조의 에너지 소비를 예측하도록 하여, 정확도와 에너지 소비를 동시에 고려하는 다목적 최적화 NAS 를 제안하였다. 결과적으로, EC-NAS를 활용한 탐색은 성능 저하 없이 에너지 효율이 높은 아키텍처를 발굴할 수 있었으며, 환경 지속 가능성을 고려한 모델 설계 연구를 촉진할 수 있음을 입증하였다.

Zhao et al.(2023)에서는 NAS 과정에서 발생하는 높은 탄소 배출 문제를 해결하기 위해, 탄소 강도 변화를 고려한 carbon-efficient NAS(CE-NAS) 프레임워크를 제안하였다. 제안하는 방법은 고탄소 시기에는 에너지 효율적인 one-shot NAS 평가를, 저탄소 시기에는 에

너지 소모가 크지만 정확도가 높은 Vanilla NAS 평가를 수행하는 이중 전략을 사용한다. 또한 LaMOO 기반의 다목적 최적화 기법을 적용하여 탐색 공간을 분할하고, 유망한 서브 공간 내에서 효율적인 샘플링과 평가를 진행한다. GPU 자원 배분은 탄소 강도에 따라 동적으로 조정되며, 이를 통해 불필요한 고탄소 연산을 줄이고 검색 효율성을 유지한다. HW-NASBench 데이터셋과 전력망 탄소 배출 추적 데이터를 활용한 시뮬레이션 결과, CE-NAS는 기존 NAS 방법 대비 가장 낮은 탄소 배출량과 우수한 네트워크 탐색 성능을 달성하였다. 이를 통해 NAS 설계 단계에서의 환경 영향을 줄이면서도 효율적인 모델 탐색이 가능함을 입증하였다〔그림 4－50〕.

〔그림 4－50〕 Carbon-efficient NAS 프레임워크



자료: Zhao, et al.(2023), p.2

Zhao et al.(2024) 연구에서는 CE-NAS를 한층 발전시켜, 강화학습 기반 GPU 자원 배분과 시계열 transformer를 활용한 탄소 강도 예측 모델을 결합한 탄소 효율적인 NAS 프레임워크를 제안하였다. 이 방식은 고탄소 시기에는 one/few-shot NAS를 통해 샘플링 중심으로, 저탄소 시기에는 일반적인 NAS를 통해 평가 중심으로 GPU 자원을 배분한다. 또한 LaMOO 기반 다목적 최적화와 qNEHVI 알고리즘을 적용하여 탐색 공간을 효율적으로 축소하고 Pareto 최적화에 초점을 맞춘다. 추가적으로, 지배도 수를 활용해 평가 우선순위를 정하고, 실제 학습이 필요한 후보를 저탄소 구간에서 집중적으로 처리한다. 실험 결과, 매우 우수한 정확도를 유지하면서, 기존 방법 대비 최대 7.22배의 탄소 절감을 달성하였다.

Xu et al.(2021) 연구에서는 기존 NAS 기법들이 아키텍처 평가를 위해 대규모 학습을 수행해야 하며, 이 과정에서 막대한 연산량과 탄소 배출이 발생한다는 점을 지적하였다. 이를 해결하기 위해, 학습 없이도 아키텍처 성능을 예측할 수 있는 gradient kernel hypothesis를 제안하였다. 이는 무작위 초기화된 네트워크의 gradient gram matrix(F-norm)를 이용해 수렴 속도와 일반화 성능을 간접적으로 추정하는 방식이다. 특히, gram matrix의 평균값(MGM)이 학습 손실 감소 및 검증 정확도와 높은 상관관계를 가짐을 실험적으로 검증하였다. 이를 바탕으로, 상위 k개의 MGM 값을 가진 아키텍처만 선별하여 전체 학습을 진행하는 KNAS 기법을 설계하였으며, NAS-Bench-201과 텍스트 분류 과제에서 기존 대비 수십 배 빠른 검색 속도와 경쟁력 있는 정확도를 달성하였다. 결과적으로, KNAS는 GPU 사용량을 크게 줄여 에너지 소비 및 탄소 배출을 완화하는 친환경 NAS 방법론을 실현하였다.

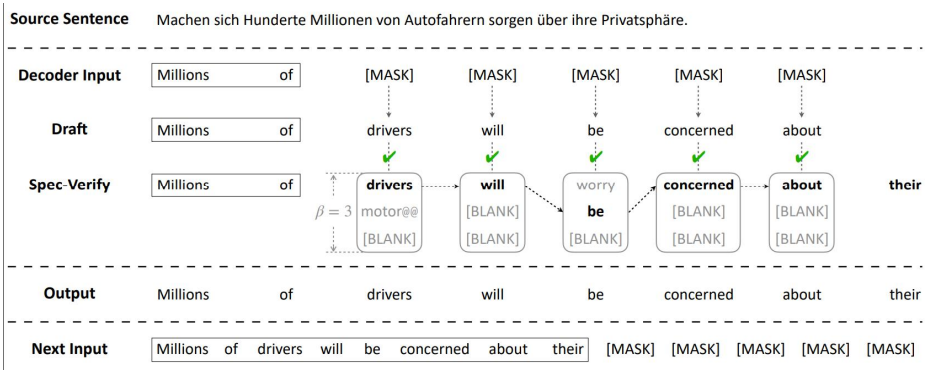
라. Adaptive decoding의 연구 동향

Xia et al.(2022) 연구에서는 mask를 사용하는 인코더-디코더 기반 계열에서 토큰마다 대형 모델을 순차적으로 호출하면서 발생하는 지연과 자원 소모가 커지는 구조적 한계를 지적하고, 이를 해소하기 위해 speculative decoding을 제안하였다. 핵심은 〔그림 4−51〕처럼 draft-then-verify 패턴을 도입하는 것으로, draft가 다음 토큰들의 초안을 빠르게 생성하고, 큰 verify 모델이 한 번의 Spec-Verification 단계에서 최장 접두어, 즉 draft가 생성한 토큰 열의 맨 앞에서부터 verify의 예측과 연속으로 일치하는 가장 긴 구간을 수락함으로써 verify 모델 호출 수를 줄인다. 이 검증 절차가 분포 보존, 다시 말해 품질 유지를 만족하도록 설계될 수 있음을 보였고, 번역·요약 등 다양한 seq2seq 작업에서 품질 저하 없이 추론 속도를 크게 향상될 수 있음을 실험을 통해 보인다. 결과적으로, speculative decoding을 적용해 모델 자체를 수정하지 않고 실행 경로를 단축하여 메모리 이동으로 인한 병목을 완화한다.

Schuster et al.(2022) 연구에서는 LLM 추론이 토큰에 접근할 때 마다 모든 디코더 레이어들을 끝까지 계산하는 탓에 불필요한 연산·메모리 이동이 발생한다는 점을 지적한다. 이를 위해 제안된 confident adaptive language modeling(CALM)는 토큰 별로 확신이 충분해지는 시점에서 early-exit을 허용하고 전체 시퀀스 수준에서 텍스트 일치의 확률적 보장

을 위해서 learn-then-test 기반 캘리브레이션을 적용한다. CALM은 로컬 지표인 소프트맥스 분포의 top-1, 마진, 은닉 상태 포화도, 예측기 출력들을 활용하여 레이어별 종료 여부를 판단하고, 감소형 임계치와 state propagation을 결합해 효율-품질 간 trade-off를 개선한다. 결과적으로 동일 모델/가중치 조건에서 평균 통과 깊이를 줄이면서 품질 제약을 만족해, FLOPs와 KV 캐시 생성/저장에 따른 메모리 이동을 함께 낮출 수 있음을 보인다.

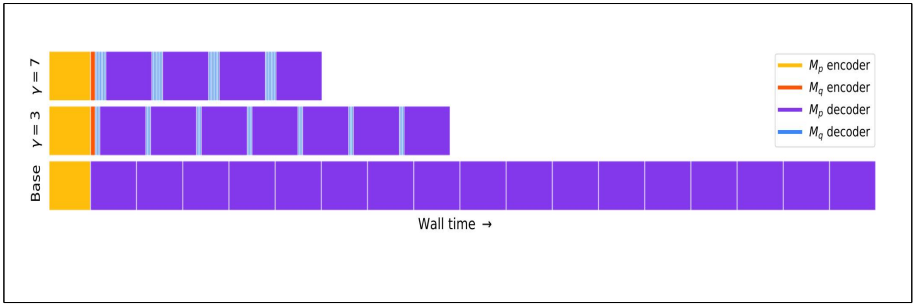
〔그림 4－51〕 Speculative decoding 방식의 토큰 생성 과정



자료: Xia, et al.(2022), p.4

Leviathan et al.(2023) 연구에서는 자기회귀 구조에서 토큰마다 원본 모델을 호출해야 하는 구조적 병목을 지적하고, 이를 해소하기 위해 speculative decoding을 체계화하고 기존 draft-then-verify 절차를 포함시켜 draft 모델이 다음 토큰들을 먼저 초안으로 생성하도록 만들고 verify 모델이 한 번의 병렬 검증으로 최장 접두어를 수락하는 방식이다. 여기서, 검증 단계는 분포 보존을 만족하도록 설계되며, 최종 출력이 verify 모델 단독으로 생성했을 때의 분포와 일치성을 보장될 수 있도록 만든다. 또한 트리 구조의 초안 검증, 캐시 재사용, 다양한 draft 모델 선택과의 호환성을 논의하며, 품질 손실 없이 대형 모델 호출 수를 줄여 메모리 이동 병목을 완화하는 런타임 가속 프레임워크임을 제안한다. 결과적으로, 수락률, 초안 길이, 상대 비용으로 가속·연산·메모리 효과를 체계적으로 분석하여 후속 연구의 기반이 된다〔그림 4－52〕.

〔그림 4－52〕 Speculative decoding의 초안 길이에 따른 wall time 분석



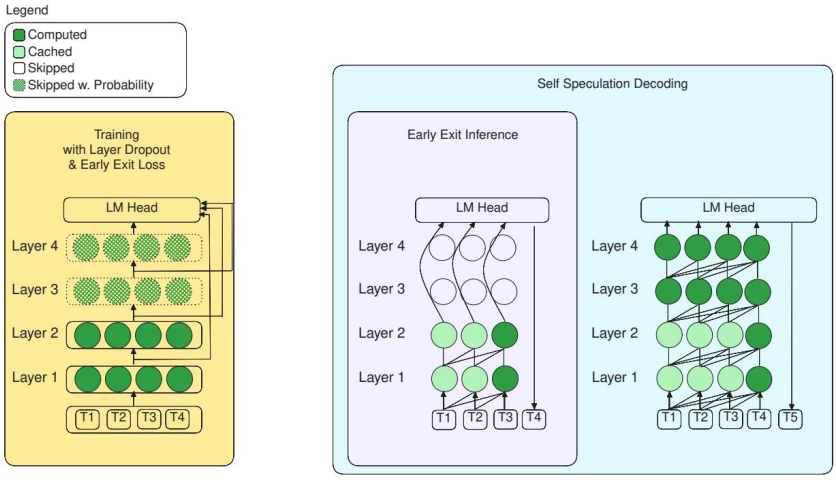
자료: Leviathan, et al.(2023), p.6

Chen et al.(2024) 연구에서는 대형 모델을 여러 GPU에 나눠 돌리는 환경에서도 early-exit 기법을 적용할 수 있는 방법을 제안한다. 학습 단계에서는 파이프라인으로 나뉜 각 구간이 추가 통신 없이도 중간 출력에 대해 학습 신호를 받을 수 있게 연결해서 기존 학습 흐름을 크게 바꾸지 않고 early-exit을 익히게 만든다. 추론 단계에서는 KV 캐시를 쓰는 일반 LLM 실행과 서로 충돌하지 않도록, 중간에서 멈춰도 캐시가 어긋나지 않게 정리하고 다음 토큰 생성으로 곧바로 이어지게 실행 흐름을 조정한다. 이렇게 하면 기존 분산 프레임워크인 Megatron-LM과 수정없이 호환되면서, 평균적으로 통과하는 레이어 수를 줄여 지연과 메모리 이동을 낮출 수 있고 동시에 출력 품질은 유지할 수 있음을 보인다.

Elhoushi et al.(2024) 연구에서는 LLM 추론에서 모든 레이어들을 끝까지 통과시키는 고정 경로가 불필요한 연산과 메모리 이동을 유발한다고 보고, 학습-추론-검증을 한 흐름으로 설계한 Layer Skip을 제안한다〔그림 4－53〕. 학습 단계에서 레이어 드롭아웃을 전방은 낮게·후방은 높게 부여하고, 공유 LM head로 중간 레이어에도 직접 언어 모델 목적을 걸어 각 층이 독자적으로 추론 능력을 갖도록 만든다. 이렇게 학습된 모델은 추론 시 로짓 기반의 기준에 따라 early-exit을 수행해 평균 통과 깊이를 낮춘다. 이어서 남은 후반 레이어를 활용한 self-speculative를 수행해서 조기 종료로 인한 잠재 오류를 빠르게 확인하고 수정한다. draft와 verify가 동일 모델의 서로 다른 구간을 사용하므로 KV 캐시와 exit query를 재사용할 수 있고, 이로써 기존의 speculative decoding 대비 메모리 발자국을 줄이면서 연산량을 개선한다. 결과적으로 Layer Skip은 별도 모듈이나 추가 헤드 없이 깊이 단축과 병렬 검증을 결합해, 동일 품질에서 지연과 에너지를 함께 낮추는 end-to-end 학

습/추론이 가능한 프레임워크를 제시한다.

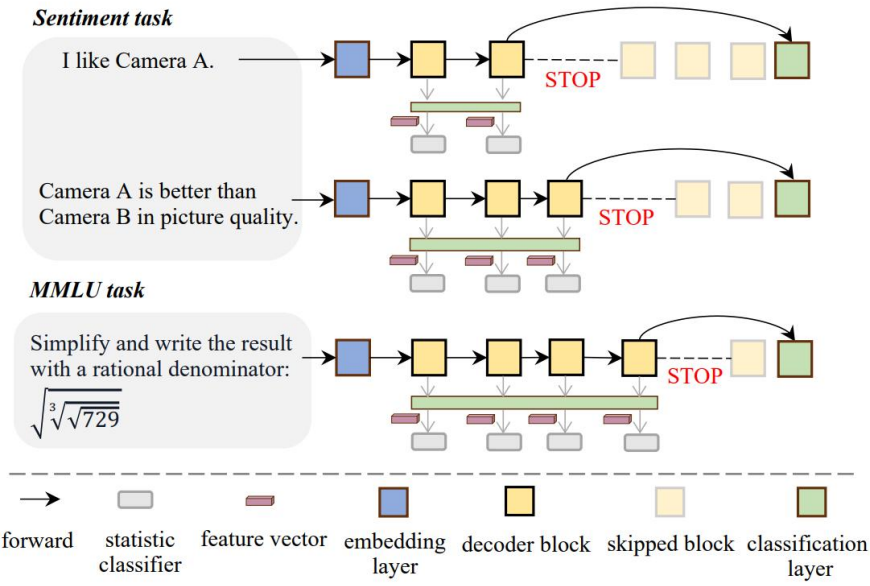
〔그림 4－53〕 LayerSkip의 end-to-end 개요



자료: Elhoushi, et al.(2024), p.2

Fan et al.(2024) 연구에서는 LLM 추론이 모든 토큰들을 마지막 레이어까지 통과시키는 고정 경로 때문에 입력 길이, 난이도와 무관하게 과잉 연산과 불필요한 메모리 이동이 발생한다는 한계를 지적한다. 이를 해결하기 위해 해당 연구에서 제안된 AdaInfer는〔그림 4－54〕와 같이 각 디코더 블록에서 logit 기반 신뢰도 신호(e.g., Top-1 확률, Top-1/2 마진 확률)를 이용해 모델이 충분한 확신에 도달하면 남은 레이어들에 대한 계산을 생략하는 early-exit 방식을 취한다. 이 접근은 모델 가중치를 변경하거나 재학습할 필요가 없고, 생략된 블록에서 연산량뿐 아니라 KV 캐시 생성·저장까지 발생하지 않게 하여 연산량과 메모리 이동을 동시에 줄인다. 결과적으로 동일 하드웨어에서 더 긴 입력과 다양한 도메인에 대해 평균 통과 깊이를 낮추어 토큰당 지연과 에너지 사용을 완화하고, 정적 구조 최적화와 병행 시 추론 에너지 효율성을 한층 높일 수 있음을 보인다.

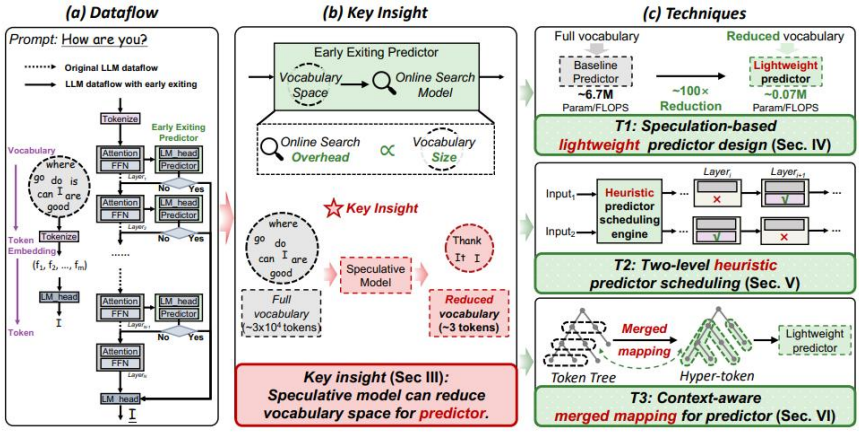
〔그림 4－54〕 AdaInfer 개요



자료: Fan, et al.(2024), p.2

Xu et al.(2025) 연구에서는 LLM 추론에서 early-exit의 예측기가 모든 레이어들에서 전 어휘 분포를 확인하면서 오버헤드가 커진다는 한계를 지적하고, 이를 줄이기 위해 speculative 모델을 이용해 예측기의 검색 공간을 대폭 축소하는 SpecEE를 제안한다〔그림 4－55〕. Draft 모델이 소수의 speculative 토큰을 제시하면, 예측기는 전체 어휘 대신 이 축소된 후보 집합만 보며 종료 여부를 판단한다. 이때 probability shift을 활용한 경량 MLP 예측기, 종료가 자주 일어나는 레이어들에만 예측기를 두는 오프라인·온라인 2단계 휴리스틱 스케줄링, speculative 트리에서 경로를 하나의 하이퍼-토큰으로 묶어 지수적 매핑 복잡도를 선형으로 낮추는 컨텍스트 인지 병합 매핑을 제시한다. 모델 파라미터를 바꾸거나 대규모 재학습 없이 적용 가능하며, 양자화·vLLM·sparsity 등과 직교적으로 결합되는 런타임 가속 프레임워크로서의 유용성을 보인다.

〔그림 4－55〕 SpecEE 개요



자료: Xu, et al.(2025), p.2

제 2 절 하드웨어 기반 솔루션

1. 맞춤형 AI 가속기 설계

가. 맞춤형 AI 가속 시스템 필요성

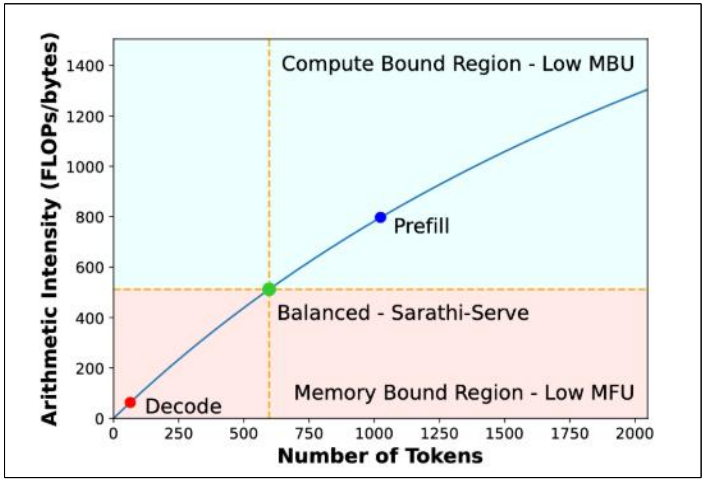
대부분의 AI 모델의 학습, 추론은 GPU 상에서 수행되며, GPU는 범용성을 극대화하기 위해 실수형 데이터와 정수형 데이터를 모두 포함한 다양한 데이터 유형과 광범위한 데이터 처리를 지원하도록 설계되어 있다. 이러한 GPU는 다양한 AI 모델을 높은 병렬성으로 기반으로 처리할 수 있다는 장점이 있지만, 특정 응용 분야나 모델 구조에서는 불필요한 연산 유닛과 복잡한 제어 로직이 함께 동작하게 되어 전력 소모와 발열이 과도하게 발생하는 문제가 있다. 예를 들어, 대표적인 GPU인 NVIDIA사의 H100의 경우, LLM 모델의 학습 과정에서 3.7MWh(메가와트시)의 연간 전력 소비량을 가지고, A100 GPU의 경우도 마찬가지로 약 3.5MWh의 연간 전력 소비량을 가져 데이터 센터의 GPU 활용은 과도한 전력을 소모한다[107]. 특히 데이터센터 환경에서는 제한된 전력 공급량으로 인해 고성능을 유지하면서도 전력 효율을 높여야 하지만, GPU는 높은 전력 소모를 야기하며, 탄소 배출 저감 목표와 상충하는 경우가 많다.

데이터 센터의 과도한 GPU 사용과 이로 인한 전력 소모량 문제를 해결하기 위하여, 저전력으로 동작 가능하며 AI 모델의 특성과 목표 응용 분야에 최적화된 맞춤형 AI 가속기를 설계하는 것이 필요하다. 이러한 맞춤형 가속기의 필요성은 LLM에 앞서, 딥러닝의 초기 핵심 모델인 CNN 연구에서도 이미 제기된 바 있다. CNN은 이미지 인식, 분류 등 다양한 비전 응용 분야에서 탁월한 성능을 보였지만, 연산량이 매우 크고 데이터 이동이 빈번하여 에너지 효율이 낮은 문제가 있었다. 이를 해결하기 위해 등장한 초기 맞춤형 CNN 가속기들은 데이터 재사용을 극대화하고 데이터 이동 거리를 최소화하기 위해 row-stationary, weight-stationary와 같은 spatial 데이터플로우 구조를 채택하였다. 이러한 구조에서는 연산에 필요한 가중치, 입력 특성맵, 부분합 데이터를 하드웨어 내부에서 저장하여 반복적으로 재사용함으로써, 외부 메모리 접근에 따른 전력 소모를 크게 줄였다. 또한, sparsity-

107) https://www.epnc.co.kr/news/articleView.html?idxno=303904

aware 처리 기법을 도입하여 0이 아닌 데이터만 연산에 참여하도록 하여 불필요한 연산을 줄이고 연산기 및 메모리 자원의 효율성을 높였다. 이러한 데이터 이동 최소화와 재사용 중심의 설계 철학은 이후 Transformer 및 LLM 가속기로 확장되었으며, 현재의 맞춤형 AI 가속 시스템의 근간을 이루는 핵심 개념으로 발전했다. 특히, 최신 LLM인 decoder-only LLM은 입력 프롬프트를 모두 입력받아 분석하는 prefill 단계와 매 스텝마다의 입력 토큰을 기반으로 출력을 생성하는 decode 단계가 있다. 두 단계는 서로 다른 연산량, 메모리 요구치를 갖기 때문에 추론 과정에 있어 서로 다른 특성(compute bound, memory bound)을 가진다. 상반되는 특성의 단계를 모두 고려하지 않은 AI 가속기 설계는 각 단계 모두 최적화할 수 없을 뿐 만 아니라 두 단계에서 모두 가속기 내의 연산기를 충분히 활용할 수 없다. 즉, 두 단계의 연산 특성, 메모리 요구 특성을 충분히 고려한 맞춤형 AI 가속기가 절실하다〔그림 4-56〕.

〔그림 4-56〕 각 단계별 입력 토큰당 연산 강도 그래프



자료: A. Agrawai et al.(2024)

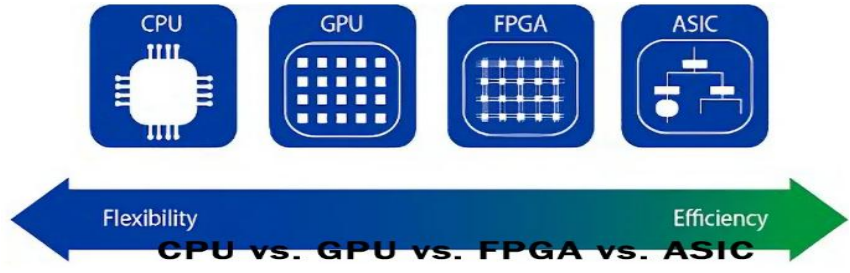
먼저, prefill 단계에서는 입력 토큰의 개수인 입력 시퀀스 길이만큼 병렬적으로 연산이 수행된다. 즉, prefill 단계에서 활용되는 가중치는 입력 토큰 간 공유되어 General Matrix Multiplication(GEMM) 연산을 수행하게 된다. 이에 따라, 재사용되는 가중치의 비중이 커짐

과 동시에 높은 연산 강도를 갖는다. 결과적으로, prefill 단계에서는 연산을 위한 하드웨어 플랫폼의 처리량에 따라 전체 가속 시스템의 성능이 제한되는 compute bound 영역에 해당된다. 또한 입력 시퀀스 길이가 커질수록 prefill 단계의 연산 강도는 선형적으로 증가하므로, 하드웨어 상의 구현에 있어 더 많은 연산기와 우수한 병렬처리 능력을 요구한다.

이와 달리 decode 단계에서는 이전 스텝에서 생성된 하나의 토큰을 입력받아 처리한다. 이때, 가중치 뿐만 아니라 KV 캐시에 대한 접근을 요구하게 되므로, 상당한 메모리 트래픽을 야기한다. 또한 prefill 단계와 달리, 단일 토큰에 대한 연산으로 인해 General Matrix Vector multiplication(GEMV) 연산을 수행한다. 즉, decode 단계는 연산 강도 대비 상당한 메모리 트래픽을 요구하기 때문에, 전체 시스템의 성능이 시스템으로 전달되는 메모리의 양으로 결정되는 memory-bound 영역에 해당된다. 또한 입력 시퀀스 길이가 커질수록 decode 단계로 전달되는 KV 캐시의 양이 증가하므로, 시스템에 전달해야 하는 메모리양이 더욱 늘어나는 경향을 가지게 된다.

이러한 AI 모델의 연산, 메모리 특성에 맞도록 설계되는 맞춤형 가속기는 앞서 설명한 LLM의 두 단계와 같은 AI 모델의 병목 구간을 분석하여 불필요한 연산을 제거하고 데이터 이동을 최소화하는 구조로 설계할 수 있으며, 특정 연산 전용 연산 유닛과 경량 데이터 처리를 지원하는 연산기를 구현함으로써 최소한의 하드웨어 자원으로 성능 대비 전력 소비를 크게 절감할 수 있다. 이러한 맞춤형 가속기는 저전력/고성능 동작을 가능하게 하는 FPGA, ASIC과 같은 플랫폼에 구현되며, 전력 소모와 발열을 낮추고, 외부 메모리 접근이라는 과도한 전력 소모의 주요 원인을 최소화할 수 있다.

〔그림 4－57〕 프로세싱 유닛 플랫폼별 호환성, 효율 비교



자료: PCB hero(2024)

범용적인 프로세서인 CPU, GPU와는 달리 Field-Programmable Gate Array(FPGA)는 프로그래밍 가능한 집적 회로로, 다양한 알고리즘을 저전력으로 동작시키면서도 원하는대로 구현할 수 있는 특징을 가지고 있다. 이러한 재구성 가능한 특징으로 인해 딥러닝 모델의 학습, 추론 가속을 위하여 활용되어왔다. 또 다른 플랫폼인 Application Specified Integrated Circuits(ASIC)은 특정한 용도나 목적에 특화된 집적 회로로, 범용 GPU와 달리 딥러닝과 같이 특정 작업을 우수한 처리량으로 수행한다는 특징을 가지고 있다. 이러한 특정 작업에 최적화된 집적 회로 구성으로 인하여 저전력으로 동작하기 때문에, 전력 효율 측면에서 매우 우수한 성능을 보인다. 즉, FPGA, ASIC 플랫폼에서 구현된 맞춤형 AI 가속기는 저전력/고성능 설계에서 필수적이다. 맞춤형 AI 가속기의 대표적인 예시인 low-bit 데이터 형식 기반으로 경량화된 모델에 최적화된 설계는 칩 면적과 제조 비용 절감에도 기여하여 하드웨어 제조 단계에서도 환경적 부담을 줄일 수 있다〔그림 4-57〕.

또한 파이프라인 기법과 데이터 재사용률을 극대화할 수 있는 연산기 구현, 메모리 대역폭과 연산기의 처리량의 통일과 같은 다양한 최적화 기법을 적용한 맞춤형 가속기를 적절한 플랫폼 상에 구현할 경우 전력 소모량 대비 연산 처리량을 극대화할 수 있으며, 이는 디지털 저탄소화 대전환을 실현하기 위한 핵심적인 기술적 기반이 된다.

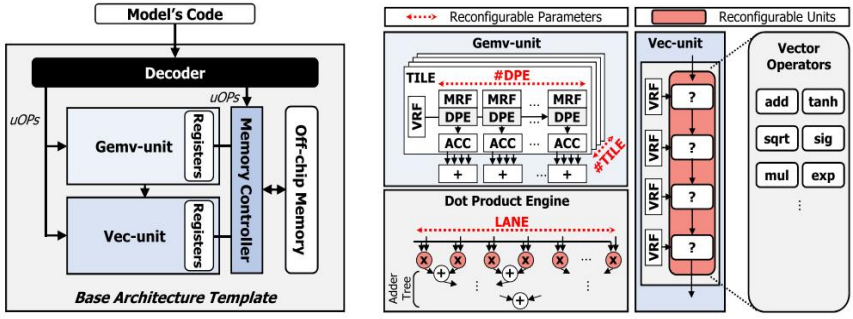
나. 맞춤형 AI 가속 시스템 연구 동향

1) A Fast and Flexible FPGA-based Accelerator for Natural Language Processing Neural Networks(Hur, et al.(2023))(TACO'23)

해당 연구는 FPGA 상에서 구현된 효율적이면서 재구성 가능한 모듈 기반 아키텍처인 Flexrun을 제안하여 다양한 자연어 처리 모델에 범용적으로 동작 가능한 AI 가속 시스템을 제안한다. 구체적으로, FlexRun의 아키텍처 템플릿은 사전 정의된 메모리 컨트롤러 및 디코더, 파라미터화 되어 옵션에 맞게 변경 가능한 Gemv-unit, Vec-unit으로 구성되어 있다. 이를 통해, 다양한 모델을 효율적으로 가속, 포팅하기 위해 주어진 모델과 FPGA 사양에 최적화된 GEMV-unit 구조를 찾고, 해당 GEMV-unit 구조를 활용한 파이프라인을 고려함과 동시에 최소 지연 시간을 갖는 Vec-unit 구조를 탐색한다. 즉, 다양한 모델과 상황에 최적화된 구조의 파라미터를 탐색하여 최적의 아키텍처를 구성한다〔그림 4-58, 59〕. Intel의 Stratix 10 MX FPGA에서 구현된 FlexRun은 GPT2 모델의 다양한 크기에 적용하여 확인

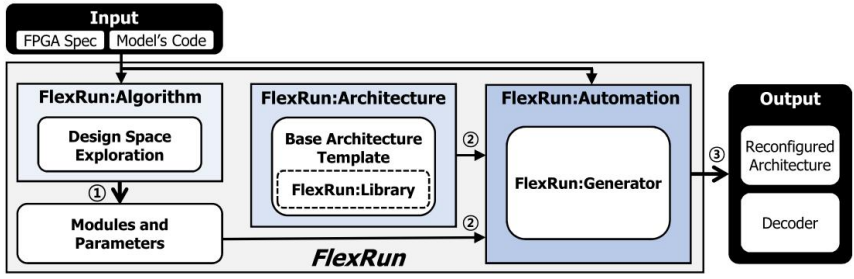
한 결과, NVIDIA사의 V100 GPU 대비 평균 2.59배 우수한 지연 시간 개선을 보인다. 이러한 LLM 맞춤형 AI 가속 시스템을 통해 FPGA 기반 설계로 인한 저전력 및 고성능 동작을 가능하게 했으며, 최종적으로 탄소 배출량 저하에 기여한다.

〔그림 4－58〕 Flexrun의 블록 다이어그램 및 연산 유닛 구조도



자료: Hur, et al.(2023). p.11

〔그림 4－59〕 Flexrun 가속 시스템의 개요도 및 모델 포팅 과정



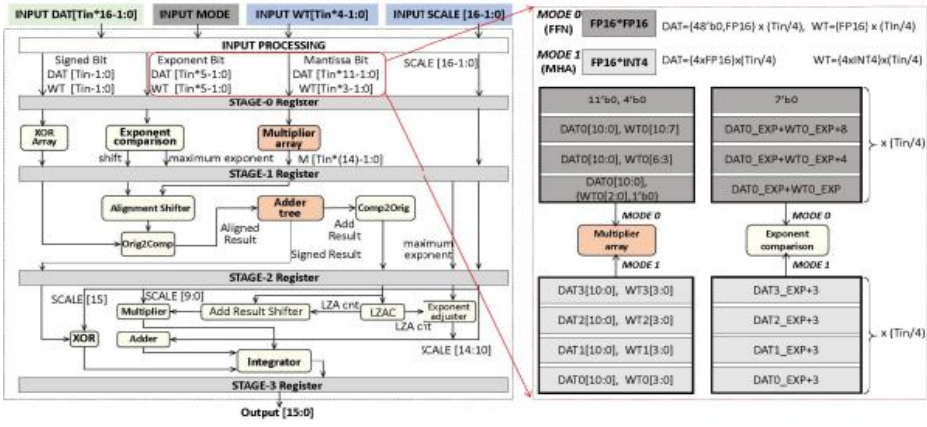
자료: Hur, et al.(2023). p.10

2) EdgeLLM: A highly efficient cpu-fpga heterogeneous edge accelerator for large language models(Huang, et al.(2025))(TCAS-1'25)

해당 연구는 LLM의 배포 과정에서 가장 큰 문제인 과도한 연산량 및 메모리 오버헤드 문제를 해결하기 위하여 로그 스케일 기반의 구조적 가중치 프루닝과 그룹 단위 양자화를 동시에 적용하였다. 이 과정에서 연산 처리를 위한 연산 유닛의 처리량과 High Bandwidth

Memory(HBM)의 메모리 대역폭을 일치시켜 연산기가 아무런 연산을 수행하지 않는 시간을 최소화하고 면적, 전력 효율을 극대화한다. 또한 입력에 따라 생성된 데이터인 KV 캐시 데이터(FP16)와 양자화된 가중치(INT4)의 다양한 데이터 형식에 대처하기 위하여 혼합 정밀도 벡터 곱셈기를 제안한다〔그림 4-60〕. 구체적으로, 제안하는 연산기는 LLM의 Feed forward network(FFN)의 연산은 FP16 간의 곱셈을 수행하며, multi-head attention(MHA) 연산은 병합된 INT4와 FP16 간의 곱셈을 처리하게 설계되어, 면적 효율 및 다양한 모델의 호환성을 높였다. 또한 가속기 내 모든 입력과 출력 텐서의 행/열 차원을 일치시킴으로써, LLM 연산 내 다양한 특징맵 행렬의 구현을 위해 필요한 데이터 재배열에서 발생하는 오버헤드를 모두 제거한다.

〔그림 4-60〕 Edgellm의 혼합 정밀도 벡터 곱셈기 구조도

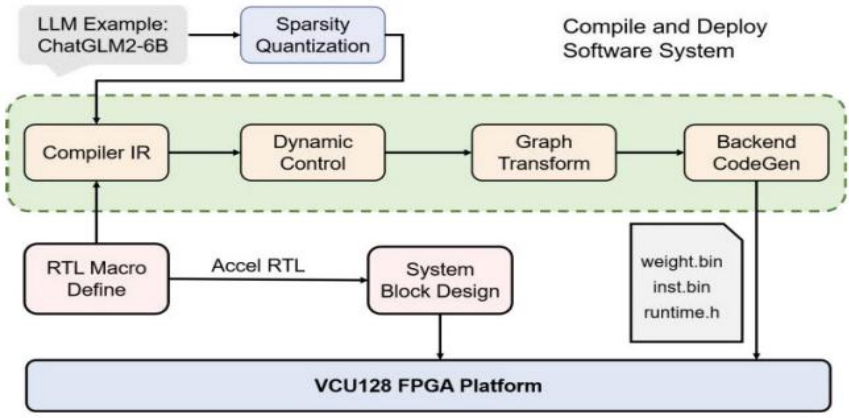


자료: Huang, et al.(2025). p.6

또한 양자화된 가중치, bitmask와 같이 연산에 필요한 데이터와 LLM 모델의 추론 과정에서 필수적으로 고려해야 하는 KV cache의 거대한 양의 데이터를 대량으로 한 번에 전송하기 위해 HBM의 높은 메모리 대역폭 활용률을 유지한다. 이러한 높은 메모리 대역폭의 활용률은 제한된 메모리 대역폭으로 인해 노출된 지연 시간을 완화함으로써 전체 지연 시간 성능의 개선을 확인할 수 있다. 더 나아가, 다양한 LLM 모델의 효율적인 FPGA 상의 배포를 위하여 end-to-end 컴파일 환경을 제안하였다〔그림 4-61〕.

특히, 동적 컴파일을 통하여 입력 토큰 길이에 대처하면서 명령어 단위 파이프라인을 통해 CPU-FPGA 이기종 AI 가속 시스템 구축을 가능하게 했다. 이러한 기법들을 통하여, Xilinx VCU128 FPGA에서 구현된 가속기는 이전 연구 대비 10~24% 뛰어난 HBM 대역폭 활용률을 보여 LLM 모델 추론을 빠르고 효율적으로 가능하게 했다. 또한 NVIDIA A100 GPU 대비 6배 우수한 전력 효율성을 보였다.

〔그림 4－61〕 EdgeLLM의 컴파일, 배포 과정 개요도



자료: Huang, et al.(2025). p.10

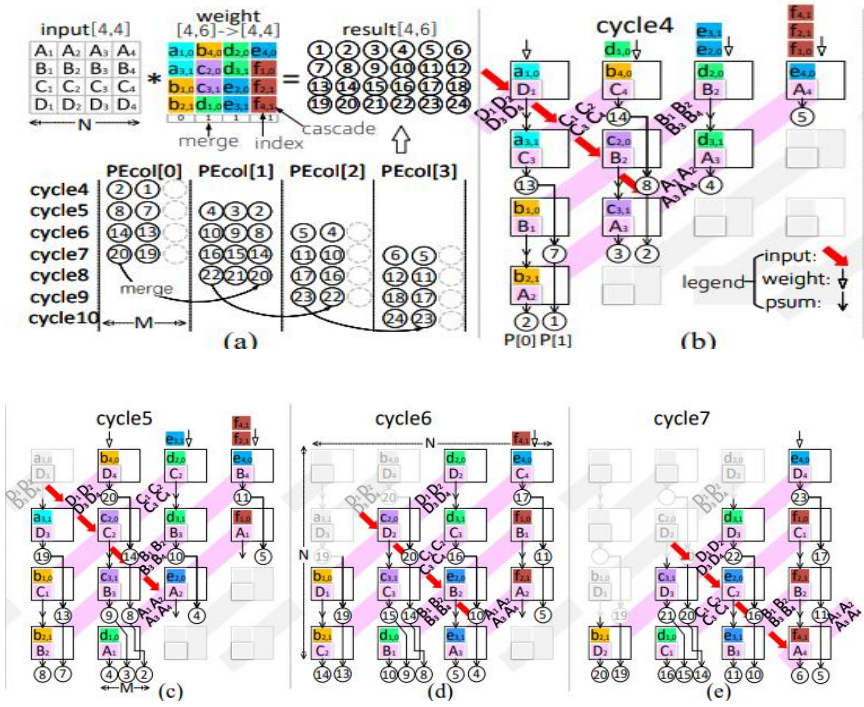
3) ChatOPU: An FPGA-based Overlay Processor for Large Language Models with Unstrutured Sparsity(Zhao, et al.(2024))(ICCAD'24)

해당 연구는 Decode-only LLM의 두 단계인 prefill과 decode가 갖는 서로 다른 특성을 균형적으로 고려하기 위하여, 비구조적 프루닝을 지원하는 멀티 코어 프로세서인 ChatOPU를 제안한다. 특히, Dense matrix-multiplication(DMM)과 sparse matrix-multiplication (SpMM)에서 모두 효율적인 데이터 재사용을 가능하도록 하는 대각 데이터 플로우 시스톨릭 어레이를 제안한다〔그림 4－62〕. DMM의 경우, 입력 데이터는 대각선 방향으로 전달하고 가중치 데이터는 수직 방향으로 전달한다. 이때, 입력 데이터 흐름의 수직으로 crossbar가 연결되어 있어, 입력 데이터 벡터가 crossbar에 도달하면 해당 crossbar에 연결된 각 연산기는 입력 데이터 벡터를 전달받고 가중치와 곱셈연산을 수행한다. 이후, 연산기는 계산

된 부분합을 수직 방향의 연산기로 전달하며 누적한다.

반면 SpMM의 경우, bitmask decoder와 non-zero loader를 통해 희소성(sparse) 높은 가중치의 bitmask를 시스톨릭 어레이로 전달함으로써 sparse 행렬 곱 연산의 전처리 과정을 제안하는 시스톨릭 어레이에 효율적으로 접목한다.

〔그림 4－62〕 ChatOPU의 대각 데이터 플로우 시스톨릭 어레이



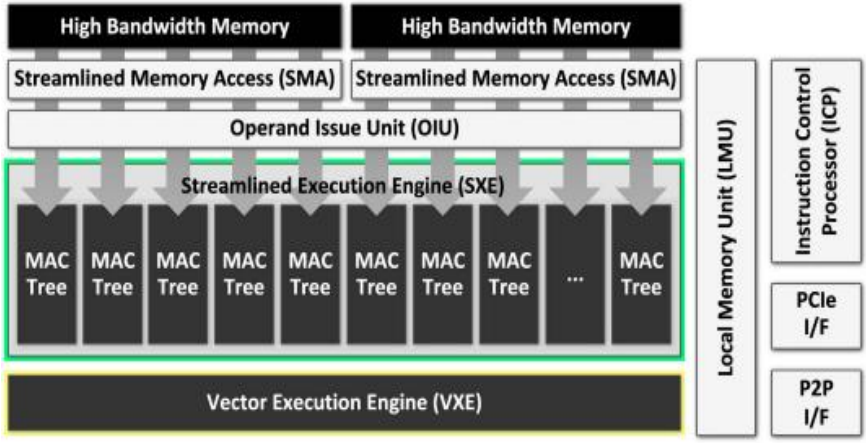
자료: Zhao, et al.(2024). p.5

Alveo U200 FPGA에 구현된 ChatOPU는 이전 프루닝 연구인 SparseGPT를 적용한 OPT-1.3B모델로 실험을 진행하였다. 그 결과, 제안하는 시스톨릭 어레이의 SpMM 처리량은 DMM 대비 최대 2배의 처리량 향상을 보인다. 더 나아가, 이전 연구 대비 33.3배 빠른 첫 토큰 생성 지연 시간을 보임과 동시에 2.7배의 처리량 향상을 달성하였다.

4) LPU: A Latency-Optimized and Highly Scalable Processor for Large Language Model Inference(Moon, et al.(2024)(IEEE Micro'24)

해당 연구는 HBM의 데이터 전송에 특화된 streamlined memory access(SMA)와 여러 고정 소수점 MAC, 그리고 wallace tree 기반의 adder-tree로 구성된 streamlined execution engine(SXE) 등의 연산 처리 요소들로 이루어진 latency processing unit(LPU)를 제안한다〔그림 4-63〕.

〔그림 4-63〕 LPU 시스템의 전체 블록 다이어그램 개요도



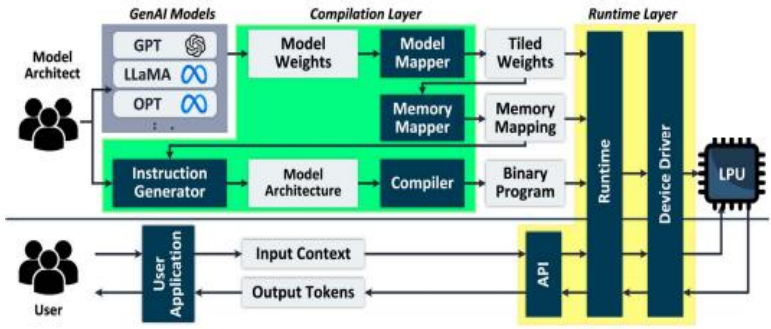
자료: Moon, et al.(2024). p.6

기존 LLM의 동작 과정에서 메모리의 제한된 메모리 대역폭으로 인한 낮은 하드웨어 활성 비율이 발생한다. 이러한 문제를 해결하기 위하여, LPU의 SXE 연산기와 HBM 메모리 대역폭 수치를 일치시켜 메모리 대역폭의 활용 비율을 높게 유지한다. 또한 여러 LPU를 데이터의 불일치 없이 효율적으로 동작시키기 위한 expandable synchronization link(ESL)를 제안하여, 다중 디바이스 시스템에서 발생하는 각 디바이스 내 데이터간의 synchronization 지연 시간을 일부 연산 시간과 동시에 실행하여 노출되는 지연 시간을 제거한다. 더 나아가, LPU를 위한 소프트웨어 컴파일 프레임워크인 HyperDex를 제안하고, 제안된 모든 기법들이 융합된 다중 LPU 가속 시스템인 Orion-cloud과 Orion-edge를 제안한

다〔그림 4－64〕.

삼성 4nm 공정으로 합성된 단일 LPU는 OPT-1.3B 모델에 대하여 1.25 ms/token의 토큰 당 지연 시간을 달성하며, 제안하는 LPU의 우수한 병렬 연산 능력 및 높은 메모리 대역폭 활용률을 입증한다. 또한 LPU의 확장성을 기반으로 여러 LPU가 탑재된 Orion-cloud는 NVIDIA H100 2장을 활용했을 때 대비 1.33배의 우수한 에너지 효율을 달성하였다.

〔그림 4－64〕 HyperDex의 컴파일 과정



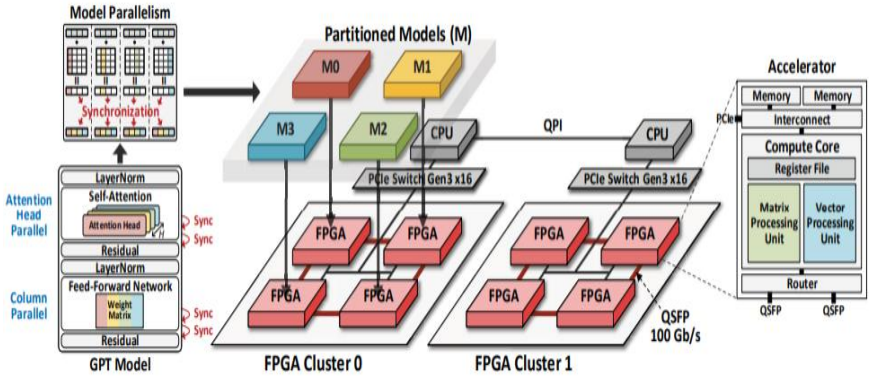
자료: Moon, et al.(2024). p.11

5) DFX: A Low-latency Multi-FPGA Appliance for Accelerating Transformer-based Text Generation(Hong, et al.(2022))(MICRO'22)

해당 연구는 GPU의 우수한 병렬 처리 능력은 큰 입력 토큰을 처리하기 위한 prefill 단계에 적합하지만, 단일 입력 토큰에 대해 반복적으로 이전 스텝의 결과 기반 연산을 수행하는 decode 단계에서는 병렬 처리 능력을 극대화하지 못한다는 점을 지적한다. 이를 해결하기 위해 dual-socket CPU와 Alveo U280 FPGA 4대로 구성된 DFX 서버 구조를 제안한다. 이는 모델 병렬 처리와 네트워크 최적화를 적용한 다중 FPGA 시스템을 통해 두 단계에서 모두 높은 처리량을 유지한다. 제한하는 DFX 시스템은 FPGA 간 최소한의 데이터 동기화만으로 최대 병렬 계산 성능을 달성한다. 구체적으로, 각 FPGA에 모델의 파라미터를 균등하게 분배하며, 이를 통해 다중 FPGA 기반 모델 병렬화 및 효율적인 연산 네트워크를 구성한다〔그림 4－65〕.

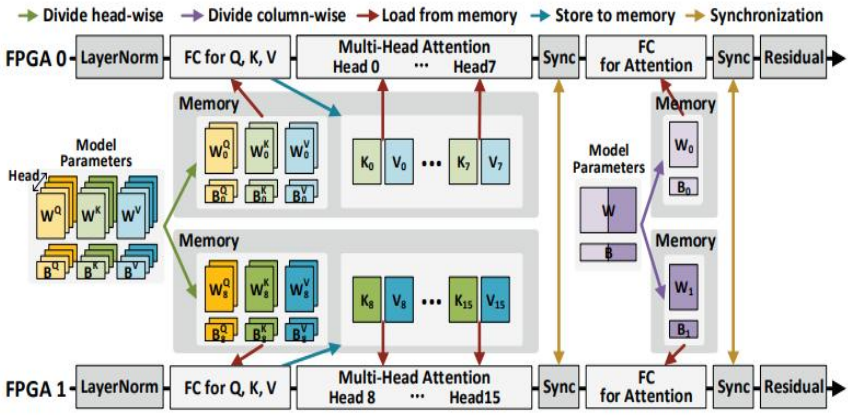
예를 들어, MHA의 가중치는 각 헤드별로 분할되고, FFN의 가중치는 각 FPGA가 할당된 파티션에서부터 개별 연산을 수행할 수 있도록 분할된다. 분할된 각 모델의 파라미터는 연산이 진행될 FPGA의 메모리에 나뉘어 저장되며, FPGA 내의 코어는 분할된 파라미터를 전달받아 연산을 수행한다. 연산 결과는 FPGA 사이의 링 네트워크를 통해 순차적으로 전달되며, MHA, FFN을 포함한 트랜스포머 구조의 연산에서 총 네 차례 데이터 동기화가 수행된다. 이러한 다중 FPGA 구조를 통해 연산 병목을 최소화하여 LLM 추론의 모든 단계에서 우수한 성능을 달성하였다〔그림 4-66〕. DFX의 다중 FPGA 가속 시스템은 GPT-2 1.5B 모델에 대하여 4개의 NVIDIA V100 GPU 대비 5.58배 빠른 추론 속도를 가지면서 3.78배 우수한 초당 토큰 개수를 보임과 동시에 3.99배 우수한 에너지 효율을 달성하였다.

〔그림 4-65〕 DFX의 다중 FPGA 시스템 개요도



자료: Hong, et al.(2022). p.5

〔그림 4－66〕 DFX의 다중 FPGA 가속 시스템의 추론 과정 개요도



자료: Hong, et al.(2022). p.5

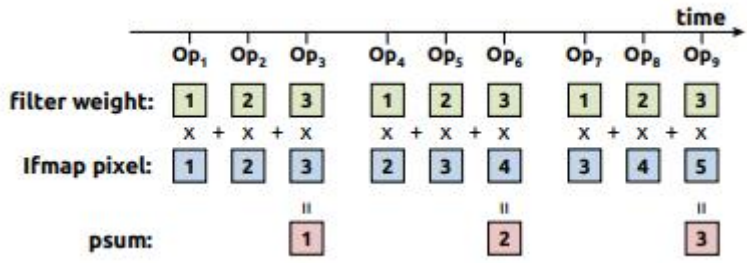
6) Eyeriss: A Spatial Architecture for Energy-Efficient Dataflow for Convolutional Neural Networks(Chen, et al.(2016))(SIGARCH'16)

해당 연구는 딥러닝의 대표적인 모델인 CNN에서 발생하는 과도한 연산량과 막대한 데이터 이동으로 인한 에너지 비효율 문제를 해결하기 위해, 데이터 이동을 최소화할 수 있는 row-stationary(RS) 데이터플로우 기반 spatial CNN 가속기 구조 Eyeriss를 제안한다. 기존의 weight-stationary(WS), output-stationary(OS), no local reuse(NLR) 데이터플로우가 각각 weight, input, partial sum 데이터 이동 및 재사용에만 초점을 맞추었기 때문에 하드웨어 자원이나 네트워크 구성에 따라 효율이 쉽게 저하되는 한계가 있다. 이에 비해 RS 데이터플로우는 filter weight, input feature map(ifmap), partial sum(psum) 등 CNN 연산에 필요한 모든 데이터의 이동을 통합적으로 고려하여 에너지 소모를 최소화하도록 설계한다.

제안된 구조는 복잡한 고차원 convolution 연산을 여러 개의 1D convolution primitive로 분해하여 연산을 병렬적으로 수행한다〔그림 4－67〕. 각 1D primitive는 filter weight의 한 row과 입력 데이터의 한 행을 입력으로 받아 한 줄의 부분합을 계산하고, 이후 여러 줄의 부분합이 합쳐져 최종 output feature map(ofmap)이 생성된다. 이러한 각 primitive는 하나의 processing element(PE)에 할당되어 연산이 수행되며, 이 과정에서 해당 PE 내부의 register file(RF)에 filter와 입력 데이터가 일정 시간 동안 고정되어 저장된다. 이를 통해

동일한 데이터를 반복적으로 외부 메모리에서 불러오지 않아도, PE 내부에서 여러 번 재사용할 수 있어 데이터 이동량이 크게 줄어든다. 이를 통해, PE 내부에서 수행되는 데이터의 local reuse를 향상하였다.

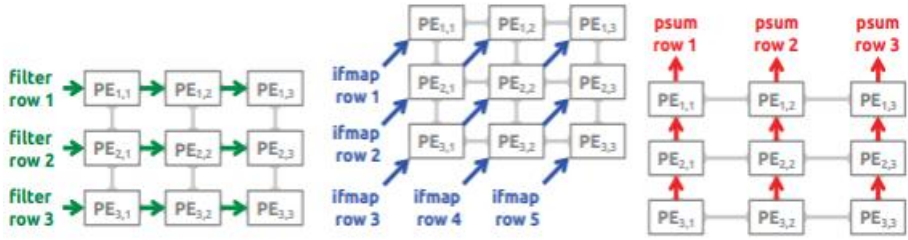
〔그림 4－67〕 Eyeriss의 1D convolution primitive 연산 과정 개요도



자료: Chen, et al.(2016). p.6

Eyeriss는 CNN의 연산 효율과 데이터 재사용을 동시에 극대화하기 위해 2단계 primitive mapping 기반의 RS 데이터플로우를 제안한다〔그림 4－68〕. 먼저 logical mapping 단계에서는 모든 1차원 연산 primitive를 동일 크기의 logical PE array에 배치하며, 각 2D convolution의 spatial locality을 고려해 인접 연산들을 logical PE set로 묶는다. 이후 physical mapping 단계에서는 실제 하드웨어 자원에 맞게 여러 logical PE set를 하나의 physical PE에 순차적으로 fold해 배치한다. 이 과정에서 set 내부의 convolution 재사용과 psum 누적을 유지하면서, set 간 filter, ifmap, psum의 공유 및 누적 기회를 극대화한다. 이를 통해 동일 위치의 연산들이 하나의 PE에서 interleaving 방식으로 수행되어, 연산 병렬성과 데이터 재사용이 동시에 향상된다.

〔그림 4－68〕 Eyeriss의 2D convolution logical PE set 데이터플로우



자료: Chen, et al.(2016). p.7

또한, RS 데이터플로우는 데이터 이동 에너지 최소화를 위해 계층적 저장 구조를 최적화하였다. 데이터는 DRAM → global buffer → PE → RF 순으로 점진적으로 이동하며, 각 단계에서 최대한 많은 데이터가 재사용되도록 설계되었다. 이와 같은 설계는 데이터가 상위 메모리에서 하위 메모리로 이동할수록 접근 비용이 낮아지고, 반대로 하위 메모리인 RF와 PE 내부는 접근 비용이 낮은 특징을 활용하여 데이터의 재활용을 위해 제안된다. 데이터 이동의 거리와 빈도를 줄이고, 저장 계층 간 접근 구조를 효율적으로 계층화함으로써 전체 에너지 소모를 최소화한 것이다.

Eyeriss는 CNN의 sparsity을 활용하여 연산 효율을 추가로 향상시켰다. 0이 아닌 값에 대해서만 연산을 수행하고, 데이터 압축 방식을 적용하여 실제 전송해야 하는 데이터량을 줄인다. 이러한 sparsity-aware processing는 RS 데이터플로우와 결합되어 연산기에서 불필요한 동작을 줄이는 동시에 데이터 이동 에너지를 더욱 절감한다. 더 나아가, 데이터의 종류에 따라 서로 다른 전용 network on-chip(NoC) 구조를 사용하여 데이터 전달 병목을 해소한다. 입력 ifmap과 filter는 global multicast NoC를 통해 여러 PE로 동시에 전송되고, 부분합 데이터는 Local PE-to-PE NoC를 통해 빠르게 누적된다. 이러한 설계 방식을 통해 Eyeriss는 CNN의 연산 처리 과정에서 발생하는 대부분의 데이터 이동을 내부에서 처리하고, 외부 메모리 접근을 최소화하였다.

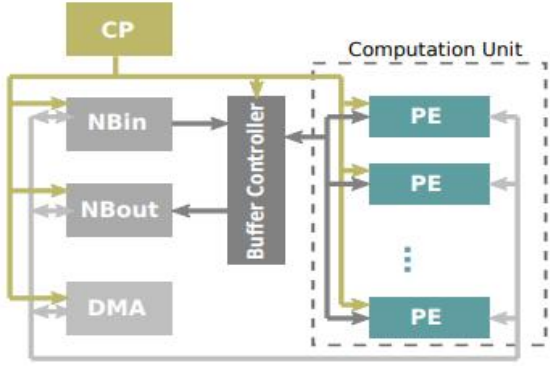
65nm 공정으로 제작된 Eyeriss chip은 AlexNet 모델을 활용한 실험에서 기존 데이터플로우 방식 대비 convolution layer에서 1.4~2.5배, fully-connected layer에서 1.3배 이상의 에너지 효율 개선을 달성하였다. 이를 통해, Eyeriss는 단순히 연산 속도를 높이는 것이 아

니라, CNN 가속에서 가장 큰 병목인 데이터 이동의 에너지 비용을 근본적으로 줄이기 위한 하드웨어 아키텍처 접근을 제시한 연구임을 입증한다.

7) Cambricon-X: An accelerator for sparse neural networks(Zhang, et al.(2016)) (MICRO'16)

해당 연구는 sparse neural network(SNN)의 비정형적 연결 구조로 인해 기존 가속기(Chen, et al.(2014), Chen, et al.(2014))가 가지는 불필요한 연산 수행으로 인한 연산 자원 낭비와 메모리 접근 비효율 문제를 해결하기 위해, sparsity와 irregularity를 효율적으로 활용하여 dense neural network 뿐만 아니라 SNN도 효율적으로 처리할 수 있는 맞춤형 가속기 Cambricon-X를 제안한다〔그림 4－69〕.

〔그림 4－69〕 Cambricon-X의 전체 구조 개요도

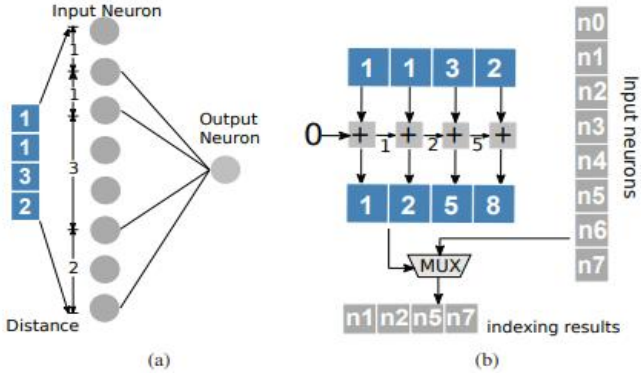


자료: Zhang, et al.(2016). p.3

제안된 가속기는 control processor(CP), buffer controller(BC), 두 개의 neuron buffer NBin과 NBout, direct memory access(DMA), 그리고 여러 processing elements(PE)로 구성된 computation unit(CU)으로 구성된다. 모든 PE는 Fat-tree 기반 network topology를 활용하여 연결되고, 각 PE는 synapse buffer(SB)와 neural network function units for PE(PEFU)로 이루어진다. 기존 연구들이 32-bit floating point(FP) 연산을 수행한 것과 달리, 해당 연구는 16-bit fixed-point 연산을 진행하는데, 이는 정확도는 FP32와 유사하게 유지되고 하드웨어 비용은 크게 절감할 수 있기 때문이다. PEFU는 주로 multiply-and-accumulate

(MAC) 연산을 수행하며, 여러 개의 multiplier와 adder-tree로 구성된다. 연산 속도를 향상하기 위해, PEFU는 곱셈과 덧셈 단계로 나누어 2단계 파이프라인으로 동작한다. BC는 각 PE로 필요한 뉴런을 전송하고, PE의 연산을 조율하며 비교적 연산량이 적은 연산을 수행한다. BC는 indexing module(IM)과 BC 전용 function unit(BCFU)로 구성된다. IM은 각 PE에 필요한 뉴런만 선택적으로 추출하여 전송함으로써 대역폭 요구량을 최소화한다. 효율적인 IM 구현을 위해 step indexing과 direct indexing 방법을 비교한다. RTL 구현 결과 step indexing이 direct indexing에 비해 10% 더 적은 면적과 40% 더 적은 전력을 소모하여 step indexing을 채택하였다〔그림 4－70〕. 분산된 synapse를 저장하는 SB는 메모리 접근 지연을 숨기면서, PEFU가 연산 중단없이 동작할 수 있는 크기인 2KB SRAM으로 구성된다. 뉴런별로 synapse 수가 불규칙하게 분포하고 있기 때문에, 각 PE의 SB는 비동기적으로 메모리에서 데이터를 가지고 와 연산 효율을 향상한다. 입력/출력 뉴런은 NBin/NBout buffer에 저장된다. On-chip buffer와 off-chip buffer는 DMA 기반 메모리 인터페이스를 통해 연결되고, 메모리 접근 병목을 완화하기 위해 synapse 데이터를 여러 chunk로 분할 전송한다. BC와 모든 PE는 다른 tree 구조 대비 데이터 이동 효율이 높고 병목이 적은 fat-tree topology를 통해 연결된다.

〔그림 4－70〕 Step indexing의 동작 과정 및 하드웨어 구현 개요도



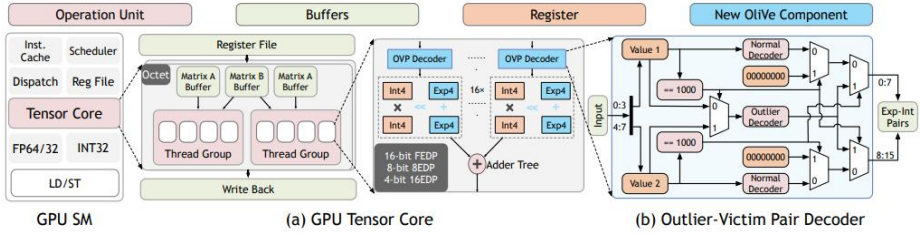
자료: Zhang, et al.(2016). p.5

LeNet-5, AlexNet, VGG-16과 같은 다양한 dense neural netrworks에서 실험한 결과 Cambricon-X는 state-of-the-art 가속기 DianNao(Chen, et al.(2014)) 대비 평균 7.23배의 속도 향상과 6.34배의 에너지 절감을 보였으며, GPU 대비 10.6배의 속도 향상, 29.43배의 에너지 효율 향상을 달성하였다. 또한, TSMC 65nm 공정으로 구현 시, 6.38mm^{2}의 면적, 954mW 전력 소모만을 보여 저전력, 고효율 가속기임을 입증하였다.

8) OliVe: Accelerating Large Language Models via Hardware-friendly Outlier-Victim Pair Quantization(Guo, et al.(2023))(ISCA'23)

기존의 outlier-aware 하드웨어 가속기인 OLAccel(Park, et al.(2018))과 GOBO(Zadeh, et al.(2020))는 행렬 내에서 outlier의 위치를 추적하기 위해 coordinate list 기반 sparsity 인코딩 방식을 사용한다. 그러나 이러한 방식은 unaligned memory access를 초래하여 복잡한 index 관리 로직과 높은 하드웨어 오버헤드를 유발하고, 결과적으로 전체 연산 효율이 저하되는 한계가 존재한다. 이러한 문제를 해결하기 위해 outlier를 전역적으로 분리하지 않고 local 수준에서 직접 처리할 수 있는 Outlier-Victim Pair Encoding(OVP) 기법과 이를 기반으로 하는 하드웨어 아키텍처 Outlier-Victim Pair Encoding(OliVe)를 제안한다〔그림 4-71〕.

〔그림 4-71〕 OliVe의 전체 구조 개요도



자료: Guo, et al.(2023). p.7

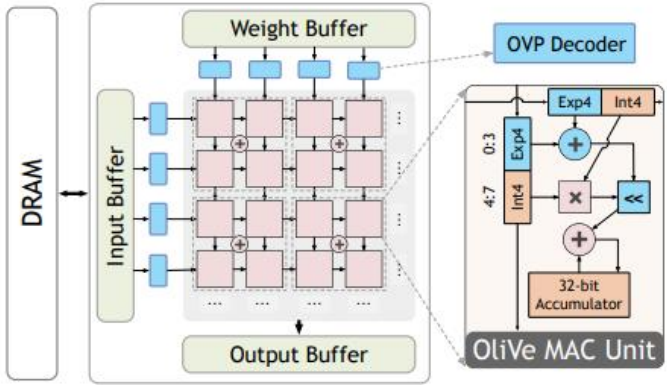
OliVe는 인접한 두 입력값을 하나의 OVP로 묶고, 이를 normal-normal, outlier-normal, outlier-outlier의 세 가지 경우로 분류한다. Normal-normal 쌍은 기존의 낮은 정밀도 양자화를 그대로 적용하고, outlier-outlier 쌍의 경우 더 큰 값을 유지한 채 작은 outlier를 제거

한다. 반면 outlier-normal 쌍의 경우 normal 값을 victim으로 설정하여 제거하고, 해당 자리를 fixed-point 기반의 고정밀 outlier 값으로 대체한다. 이러한 구조를 통해 별도의 sparsity indexing이나 전역 좌표 관리가 필요하지 않으며, 메모리 접근이 정렬된 형태로 단순화되어 하드웨어 복잡도를 크게 낮출 수 있게 된다.

또한 OliVe는 outlier가 가지는 값의 넓은 범위를 효율적으로 표현하기 위해 adaptive bias float(abfloat)이라는 새로운 데이터 타입을 제안한다. Abfloat은 exponent에 adaptive bias를 적용하여 normal 값과 outlier 값의 표현 영역이 겹치지 않도록 설계되었으며, 복잡한 floating-point 연산 대신 shift 기반의 float-to-fixed 변환을 수행한다. 이를 통해 별도의 고정밀 floating-point 연산기 없이도 단순한 shift 및 add 연산만으로 outlier 처리가 가능해지며, 하드웨어 구현의 효율성을 높인다.

OliVe는 기존 GPU Tensor Core 구조에 통합하기 위해 4-bit OVP decoder를 16 EDP(element-dot product)에, 8-bit decoder를 8 EDP에 추가하였다. OVP decoder는 1-byte 단위로 입력 데이터를 읽고 outlier identifier를 감지하여 victim 값을 0으로 처리한 뒤, outlier 값만 abfloat decoder로 전달한다. Abfloat decoder는 bias와 mantissa 정보를 이용해 exponent-integer pair 형태로 복원하며, 복호화된 값들은 모두(exponent, integer) pair

〔그림 4－72〕 OliVe의 MAC unit 전체 구조 개요도



자료: Guo, et al.(2023). p.8

로 변환되어 OliVe 전용 Multiply-Accumulate(MAC) 유닛에서 연산된다[그림 4-72]. OliVe MAC 유닛은 두 개의 exponent-integer pair의 곱셈을 단순한 shift 연산으로 대체함으로써 기존 INT MAC 구조에 소규모의 shifter와 adder만을 추가하여 구현할 수 있게 되었다.

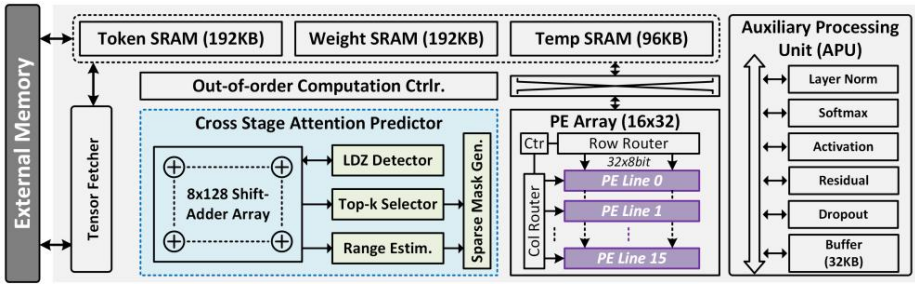
ASIC 수준에서 구현된 OliVe 아키텍처는 output-stationary systolic array 구조를 기반으로 하며, GPU와 달리 디코더를 전체 PE 배열에 분산하지 않고 threshold에만 배치함으로써 디코더 수를 약 90% 이상 절감한다. 각 PE는 4개의 4-bit 연산 unit으로 구성되고, 여기에 하나의 adder와 shifter를 추가하여 abfloat 및 int8 혼합 정밀도 연산을 지원한다. 이를 통해 4-bit 기반의 혼합 정밀도 연산이 가능하며, outlier와 normal 값이 함께 존재하는 실제 LLM 연산 환경에서도 높은 연산 효율을 유지할 수 있게 된다.

하드웨어 구현 결과, OliVe는 기존 outlier-aware 아키텍처 대비 뛰어난 성능 향상을 보인다. 4-bit 양자화 환경에서 OLAccel(Park, et al.(2018)) 대비 4.8배의 연산 속도 향상과 3.7배의 에너지 절감을 달성하였으며, 전체 면적 대비 4-bit 디코더와 8-bit 디코더의 점유율은 각각 2.2%와 1.5%로 매우 낮은 수준을 유지한다. 특히, sparsity 기반 접근법의 복잡한 제어 logic을 제거하고, 단순한 local 인코딩 방식과 하드웨어 친화적인 데이터 표현 방식을 결합함으로써, OliVe는 outlier-aware 양자화의 실질적인 ASIC 구현 가능성을 입증하였다.

9) FACT: FFN-Attention Co-optimized Transformer Architecture with Eager Correlation Prediction(Qin, et al.(2023))(ISCA'23)

해당 연구는 Transformer 기반 모델의 추론 과정에서 발생하는 과도한 연산량과 전력 소모 문제를 해결하기 위해, Eager Prediction(EP) 알고리즘과 이를 지원하는 ASIC 기반 가속기 FACT(Fully Accelerated Co-designed Transformer)를 제안한다[그림 4-73]. 기존 연구들이 주로 attention 모듈의 연산 최적화에 집중하는 반면 FACT는 Transformer의 세 주요 연산 모듈인 QKV 생성, attention, FFN을 모두 고려하여 효율적으로 가속화 하기 위해 알고리즘-하드웨어 co-design 방식을 도입한다.

〔그림 4－73〕 FACT의 전체 구조 개요도



자료: Qin, et al.(2023). p.7

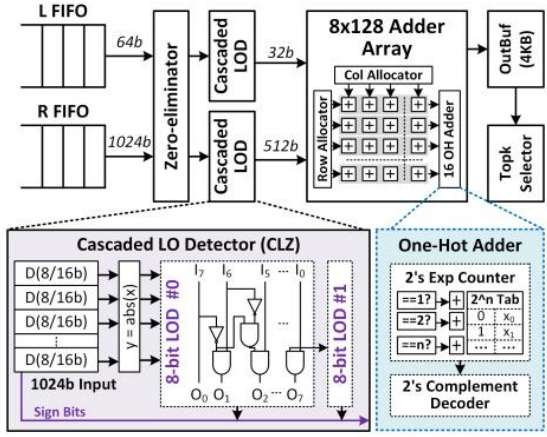
FACT는 Eager Prediction(EP) 알고리즘을 통해 입력 토큰과 weight 값을 기반으로, attention 행렬의 중요도를 근사적으로 예측한다. 이를 위해 EP는 Leading-One Detector(LOD)를 사용하여 각 값의 부호 및 크기를 나타내는 지표를 추출하고, 이를 shift와 add 연산으로 조합해 estimated attention matrix를 생성한다. 이러한 방식은 연산 비용이 큰 곱셈 없이도 데이터 간 상대적 크기를 추정할 수 있어, attention 계산에 필요한 연산량과 전력 소모를 크게 줄인다. 또한, 예측된 attention 행렬에서는 각 row에 대해 Top-k filtering을 수행하여 중요한 key 토큰만을 남기고, 중요도가 낮은 항목에 대한 연산은 생략한다. 이를 통해 EP는 attention 단계에서 불필요한 연산을 제거할 뿐만 아니라, KV sparsity와 Query sparsity를 동시에 활용할 수 있게 된다. KV sparsity는 Top-k에 선택되지 않은 column에 대응하는 key 및 value tensor를 제거하여 메모리 접근량을 줄이고, query sparsity는 softmax 확률 분포 특성을 이용해 특정 토큰이 지배적인 경우, 해당 row 전체의 attention 연산을 생략한다.

추가로 EP는 FFN 단계에서도 효율을 높이기 위해 token-wise mixed precision 기법을 적용한다. 예측된 attention에서 한 토큰이 Top-k에 포함된 빈도를 기반으로 중요도를 평가하고, 중요한 토큰에는 full precision을, 중요도가 낮은 토큰에는 상위 4-bit만을 사용하는 mixed precision 방식을 적용한다. 이를 통해 FACT는 transformer 전체의 연산량과 전력 소모를 효과적으로 줄이면서도 정확도를 유지할 수 있게 된다.

FACT는 EP 알고리즘을 하드웨어에서 효율적으로 수행하기 위해 전용 One-hot Adder Tree 기반 EP 유닛을 제안한다〔그림 4－74〕. EP 유닛은 LOD를 활용해 입력 데이터의 상

위 bit를 탐지하고, 이를 병렬 카운터 방식으로 누적하여 곱셈연산을 근사적으로 수행한다.

〔그림 4－74〕 Eager Prediction 유닛의 전체 구조 개요도



자료: Qin, et al.(2023). p.7

EP 수행과정에서 발생하는 latency를 제거하기 위해서는 out-of-order QKV generation scheduler 기법을 제안한다. Attention 행렬에서 하나의 row가 예측되면 해당 row에서 필요한 key와 value가 확정되기 때문에 바로 PE array에서 해당 key와 value에 대한 projection을 시작하고 결과를 temp buffer에 저장한다. 이후 새로운 attention 행이 예측될 때마다 새로운 KV만 추가로 연산을 진행해 attention 행렬의 예측이 끝날 때까지 기다리지 않고 병렬적으로 연산을 진행해 예측과 연산을 병렬적으로 수행한다.

FFN 연산의 경우, 중요도에 따라 8/4-bit의 mixed precision 연산을 진행하게 되고 모든 토큰 데이터는 INT8로 저장되게 된다. 중요도 낮은 토큰은 상위 4-bit만 사용되기 때문에, 데이터 접근 시 LSB까지 읽게 되어 메모리 대역폭에 낭비가 발생한다. 이를 해결하기 위해 diagonal storage pattern을 제안한다. 데이터를 MSB, LSB 4-bit로 분리하여 MSB를 bank 0의 address 0, LSB를 bank 1의 address 1로 저장하여, 8-bit를 읽을 때, 짝수 bank의 마지막 address bit만 반전시켜 MSB와 LSB를 한 번에 읽고 4-bit만 읽을 땐, 해당 address만 읽도록 한다. 이를 통해 FACT는 INT8/INT4와 같은 mixed-precision 연산 환경에서도 메모리 접근 낭비 없이 FFN 처리량을 극대화한다.

FACT는 제안된 EP 알고리즘의 효과를 검증하기 위해 22개의 다양한 벤치마크를 대상으로 성능 평가를 진행한다. 그 결과, FACT는 transformer의 세 가지 주요 모듈을 통합적으로 최적화하여 평균 3.59배 throughput 향상을 달성한다. FACT는 Nvidia V100 GPU 대비 EP 알고리즘의 하드웨어 최적화가 반영되지 않은 구현에서는 26.6배의 에너지 절감만을 보인다. 하드웨어 최적화를 반영한 FACT는 다양한 벤치마크 전반에서 평균 4,388 GOPS/W의 에너지 효율을 기록하였으며, 이는 Nvidia V100 GPU, ELSA(Ham, et al.(2021)), Sanger(Lu, et al.(2021)) 대비 94.98배, 3.91배, 22.86배 높은 성능을 달성했음을 보인다. 이를 통해, transformer의 QKV, Attention, FFN 모든 모듈을 가속화 한 고효율 ASIC 가속기의 우수성을 입증한다.

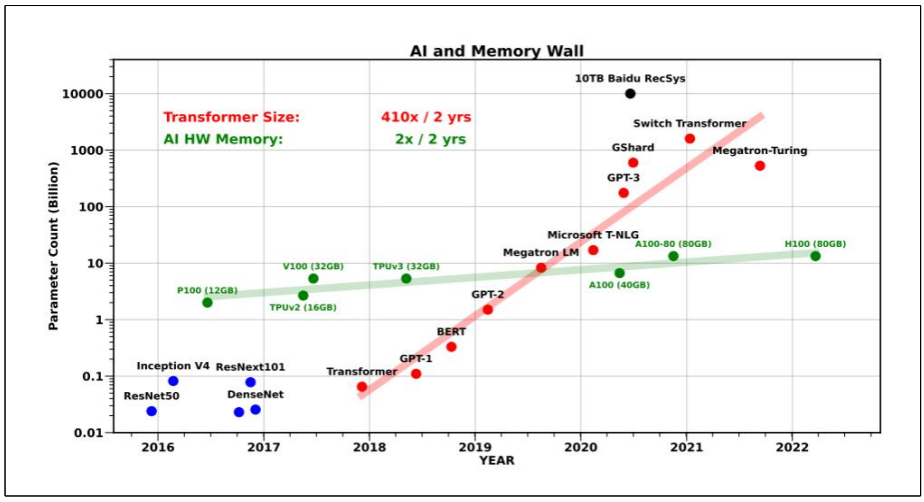
2. 메모리 아키텍처 최적화

가. PIM 아키텍처의 필요성

GPU와 CPU 기반의 AI 시스템은 대부분 계층형 캐시 구조와 외부 DRAM 메모리에 의존하고 있다. 그러나 대규모 AI 모델을 처리하는 과정에서 파라미터와 특징맵 데이터를 반복적으로 불러오고 저장해야 하며, 이 과정에서 빈번한 메모리 접근이 발생한다. 메모리 접근은 연산 자체보다 훨씬 높은 전력을 소모하며, 특히 외부 DRAM 접근은 전체 시스템 전력 소모의 상당 비중을 차지한다. 더욱이 메모리 대역폭의 한계로 인해 연산 유닛이 연산을 수행하는 시간보다, 메모리로부터 데이터가 전송되기를 기다리는 시간이 더 길다. 이는 하드웨어 자원의 비동작 상태를 늘리고 전체 에너지 효율을 크게 저하시킨다.

특히, LLM의 추론 과정에서는 기존의 CNN보다 수십~수백배 많은 파라미터를 요구하게 되어, 제한된 메모리 대역폭으로 인하여 모델의 추론 지연 시간 성능이 제한되는 문제가 발생한다. 즉, 이는 사용 가능한 연산기의 처리 능력은 우수하나, 연산기에 전달되는 데이터의 전송 속도가 상대적으로 매우 낮아 처리량이 제한된다는 것을 의미한다〔그림 4−75〕.

〔그림 4－75〕 연도별 AI 모델의 파라미터 수, 하드웨어 메모리



자료: Gholami, et al.(2024)

이러한 문제를 해결하기 위해서는 메모리 아키텍처 최적화가 필수적이다. 온칩 SRAM이나 대형 스크래치 패드 메모리를 적용하면 외부 DRAM 접근을 최소화할 수 있으며, 데이터 재사용 중심의 버퍼 설계와 데이터 압축·양자화 기법을 함께 활용하면 전력 효율을 극대화할 수 있다. 또한 계층 간 데이터 이동을 줄이는 구조를 도입하면 메모리 병목 현상을 완화하고 하드웨어 자원의 활용도를 높일 수 있다. 이와 같은 메모리 구조 개선은 동일 소비전력으로 더 많은 작업을 처리하게 하여, 성능과 에너지 효율을 동시 향상이 가능하다.

여기에 더해 메모리와 연산 유닛을 물리적으로 통합하는 Processing-In-Memory(PIM) 기술을 활용한 가속기는 저탄소 전환을 위한 또 다른 강력한 하드웨어 솔루션이 될 수 있다. PIM 기반 가속기는 오프칩 메모리(DRAM)과 가속기 간의 데이터 이동을 최소화하기 위해 메모리 내부 또는 근처에서 구현된 연산기를 통해 연산을 수행함으로써, 외부 버스 대역폭 요구를 줄이고 메모리 접근 전력을 획기적으로 절감할 수 있다.

특히, 대규모 행렬 연산이나 메모리 집약적 AI 모델에서 PIM 구조는 기존 GPU 대비 훨씬 높은 연산 효율과 에너지 절감을 동시에 제공한다. 또한 최근에는 넓은 대역폭을 기반으로 기존보다 빠른 데이터 전송을 가능하게 하는 HBM의 내부에 PIM 아키텍처를 결합함으로써 메모리 병목으로 인해 제한되는 성능의 저하를 완화한다.

구체적으로, PIM은 구현 위치에 따라 크게 두 가지 형태로 구분이 가능하다. 각각은 메모리 셀 내에서 연산을 수행하는 Processing-using-Memory(PUM) 구조와 메모리 주변 또는 로직 다이에서 연산을 수행하는 Processing-near-Memory(PNM) 구조라고 불린다.

PUM 구조는 메모리 셀 어레이 내부나 bitline 수준에서 직접 연산을 수행하는 방식으로, 주로 아날로그적 특성을 활용한다. Sense amplifier의 전하 공유나 bitline 전류 누산을 이용하여 AND, OR, COPY 등의 연산을 수행하고 이들을 결합한 형태의 곱셈과 덧셈연산을 구현한다. 동일한 메모리 기반 가속 연구 방식을 Compute-in-Memory(CIM)이라고 칭하기도 한다. 이 구조는 데이터가 저장된 위치에서 곧바로 연산이 이루어지므로 데이터 이동 에너지를 근본적으로 줄일 수 있으며, 대규모 병렬 연산이 가능하다.

PNM 구조는 메모리 셀 주변이나 별도의 로직 다이에 디지털 연산 로직을 배치해 데이터를 근처에서 읽어와 연산을 수행하는 방식이다. DRAM에서는 sense amplifier 근처에 간단한 ALU를 삽입하거나, HBM과 같이 logic die를 갖는 경우, 그곳에 연산기를 배치하여 연산을 수행하는 형태가 이에 해당한다. 이러한 구조는 디지털 정밀도와 제어 용이성을 확보하면서도 칩 외부 데이터 이동을 최소화할 수 있다는 장점이 있다.

한편, PIM은 사용되는 메모리 소자에 따라서도 다시 세분화 될 수 있다. SRAM, DRAM, RRAM, 그리고 MRAM과 같은 소자별 메모리 셀 구조와 공정 특성이 달라 연산 구현 방식과 효율성에 큰 영향을 미친다. SRAM은 로직 회로와 동일한 CMOS 공정에서 제작되며, 낮은 지연 시간과 높은 동작 주파수를 제공한다. 이러한 특성 덕분에 연산 유닛과의 통합이 용이하며 메모리 내부에서 직접 연산을 수행하는 CIM 구조로의 확장이 활발히 연구되고 있다. CIM 기반 SRAM은 bitline 전압이나 전류 누적을 이용해 연산을 병렬적으로 수행함으로써 매우 높은 연산 밀도와 에너지 효율을 제공한다. 다만, 공정 변동성과 아날로그 신호 정밀도 한계로 인해 보정 회로가 필요하며, 주로 경량화된 신경망 가속기나 엣지 AI 응용을 중심으로 연구가 진행되고 있다.

DRAM은 대용량 데이터 저장과 높은 접근 병렬성을 갖춘 구조로, PIM 구현을 위한 가장 현실적인 기반으로 평가된다. DRAM 셀 구조는 단일 트랜지스터-커패시터(1T1C)로 단순하지만 이를 기반으로 한 bank와 subarray는 대규모 병렬 접근을 지원하므로 연산 기능을 확장하기에 적합하다. 특히, 각 bank의 주변 회로에 연산 블록을 추가하면 데이터 이동 없이 동일 bank 내에서 연산을 수행할 수 있는 구조가 많이 연구되고 있으며, NPU와 함께

하나의 시스템을 이루어 AI 워크로드를 처리하는 연구도 진행되고 있다.

비휘발성 메모리(NVM) 기반 PIM은 차세대 저전력·고집적 연산 플랫폼으로 주목받고 있다. RRAM-PIM은 저항 상태 변화(High/Low)를 이용해 가중치 행렬을 아날로그 전도도로 표현하고, 입력 전압에 비례하는 전류 누적을 통해 GEMV를 지원한다. NVM 기반 구조는 낮은 대기 전력 특성 덕분에 엣지 AI 기기나 초저전력 응용에 적합하다.

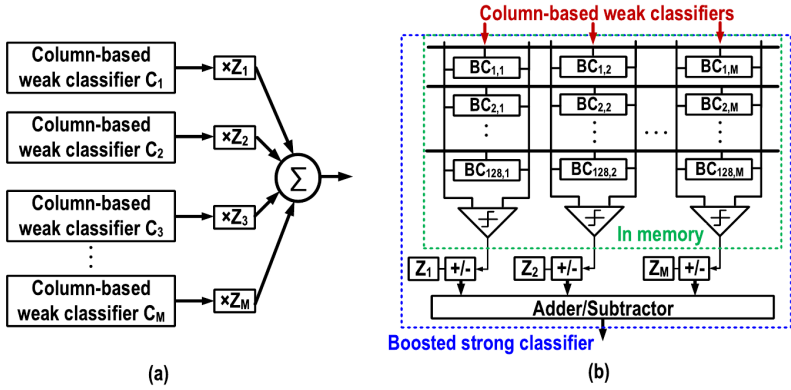
기존의 연산 중심 아키텍처에서는 데이터 이동이 전체 전력의 대부분을 차지했지만, PIM은 데이터를 이동시키지 않고 메모리 내부 혹은 근처에서 직접 연산함으로써 시스템 차원의 전력 소모를 획기적으로 줄일 수 있다. 특히 엣지 디바이스, 데이터센터, 자율주행, 금융, 과학 연산 등 대규모 데이터 처리가 필수적인 다양한 응용 분야에서 이러한 전력 절감 효과는 곧바로 탄소 배출 저감으로 이어진다. 즉, PIM은 단순히 속도 향상만을 목표로 하는 기술이 아니라, 지속 가능한 컴퓨팅의 핵심축으로 자리매김하고 있으며, 메모리 내부 연산을 통한 에너지 효율 개선은 산업 전반의 탄소중립 목표 달성에도 실질적인 기여할 수 있다.

나. PIM 연구 동향

1) In-memory computation of a machine-learning classifier in a standard 6T SRAM array(Zhang, N., et al.(2023))(JSSC'17)

해당 연구는 기존의 머신러닝 분류기가 연산보다 데이터 이동에 더 많은 에너지를 소모한다는 점에 주목하며, 표준 6T SRAM 셀을 그대로 활용하면서, 메모리 내부에서 입력 특징과 가중치의 MAC 연산을 아날로그 방식으로 수행하는 in-memory classifier 구조를 제안한다. 제안된 구조는 SRAM의 워드라인을 아날로그 전압으로 구동하여 저장된 bit 값에 따라 전류를 생성하고 이를 bitline을 통해 합산하는 방식으로 선형 분류를 수행한다. 이때 각 column은 하나의 약한 분류기로 동작하며, 여러 column의 출력을 결합해 최종 판정을 내린다. 또한, 실제 칩의 편차나 잡음에 의한 연산 오차 보정이 가능한 error-adaptive classifier boosting(EACB) 기법을 적용하여 학습 단계에서 회로 오차를 함께 고려하도록 설계하였다〔그림 4-76〕.

〔그림 4－76〕 EACB의 전체 구조도 (a) 학습 과정을 거쳐 얻어진 분류기 구조
(b) 메모리 내 분류기



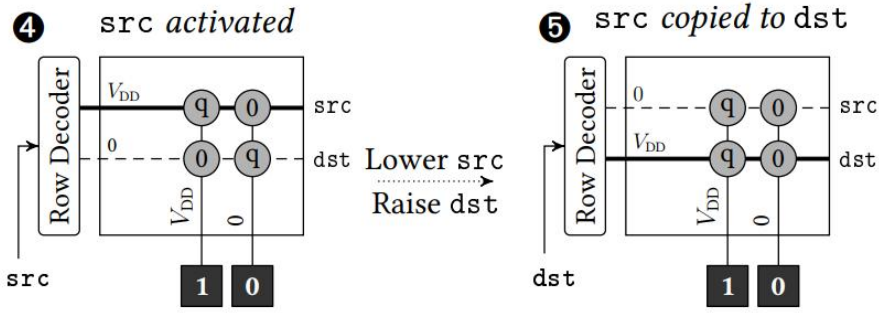
자료: Zhang, et al.(2023). p.3

130nm 공정으로 제작된 128×128 셀 어레이 칩을 이용해 선형 분류기 기반의 MNIST 추론 실험을 수행한 결과, 130nm CMOS 기반의 10-bit 디지털 MAC 시스템과 유사한 분류 정확도를 유지하면서 에너지 소모를 113배 이상 줄였다.

2) RowClone: Fast and Energy-Efficient In-DRAM Bulk Data Copy and Initialization (Seshadri, Vivek, et al.(2013))(MICRO'13)

이 연구는 메모리 내의 데이터 COPY를 수행하고자 할 때, 주소 할당 등을 위해 CPU로 데이터가 전달되는 과정에서 발생하는 지연과 에너지 소모 문제를 해결하고자 한다. RowClone은 2가지 모드를 제공하며 fast parallel mode로는 같은 subarray 안에서 한 row의 데이터를 바로 다른 row으로 복사한다. 이때는 단 두 번의 활성화 명령만으로 복사가 이루어지며, 데이터를 CPU로 보내지 않는다. 또 다른 모드인 pipelined serial mode(PSM)는 서로 다른 bank 간의 복사에 사용되며, DRAM 내부의 공유 버스를 통해 읽기와 쓰기를 겹쳐 실행한다〔그림 4－77〕.

〔그림 4－77〕 RowClone의 Pipelined Serial Mode.



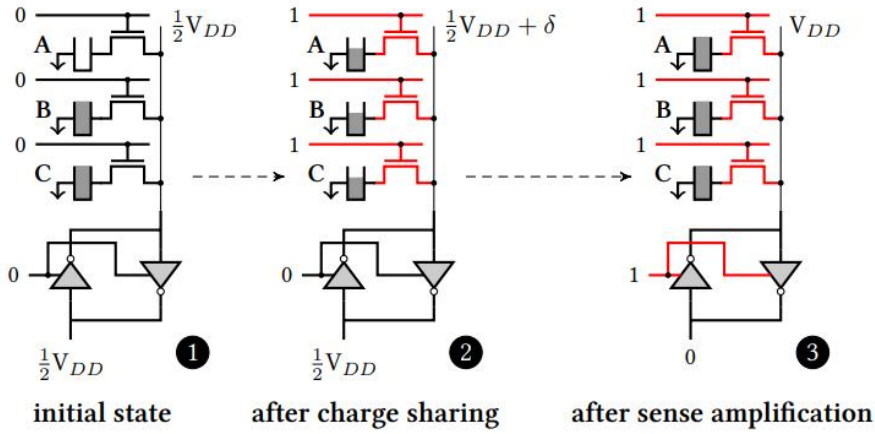
자료: Seshadri, Vivek, et al.(2013). p.3

실험 결과, RowClone은 4KB 데이터를 복사할 때 기존 방식보다 최대 11배 빠르며, 에너지는 70배 이상 절감하고, 데이터 초기화 작업에서는 6배의 속도 향상, 에너지는 40배 이상 줄였다. 더불어 하드웨어 변경이 거의 필요하지 않아 칩 면적이 0.01% 정도만 증가한다.

3) Ambit: In-Memory Accelerator for Bulk Bitwise Operations Using Commodity DRAM Technology(Seshadri, Vivek, et al(2017))(MICRO'17)

RowClone(Seshadri, Vivek, et al.(2017))을 확장한 기존 DRAM이 한 번에 한 row만 활성화하는 것과 달리, Ambit은 triple-row activation(TRA)을 통해 세 row을 동시에 열어 bitline에서의 전하 공유를 아날로그 연산에 활용한다. 구체적으로, 세 가지 bit 중 둘 이상이 '1'이면 전압이 상승해 sense amplifier가 '1'로 판정하고, 둘 이상이 '0'이면 방전되어 '0'으로 감지된다. 이를 이용하여 Ambit은 한 row을 미리 0 또는 1로 초기화함으로써, C=0일 때는 나머지 두 행의 AND, C=1일 때는 OR 연산을 구현한다〔그림 4－78〕.

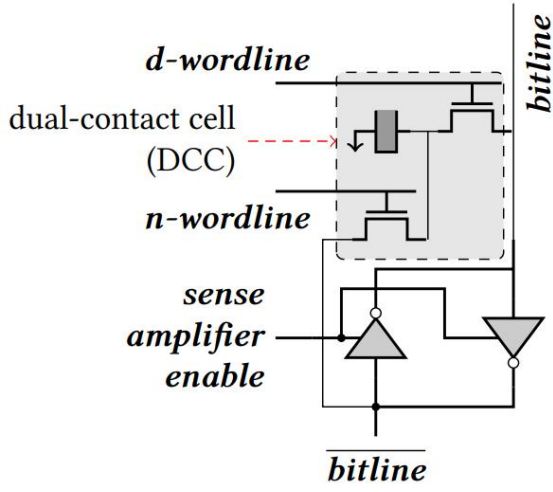
〔그림 4－78〕 Ambit의 Triple-row Activation



자료: Seshadri, Vivek, et al.(2017) p.4

반면 NOT 연산은 sense amplifier 내부의 두 개 인버터를 활용하여 양쪽 신호를 모두 연결할 수 있는 dual-contact cell(DCC)을 추가해 반전된 신호를 저장한다〔그림 4－79〕.

〔그림 4－79〕 Sense amplifier에 연결된 Dual Contact Cell



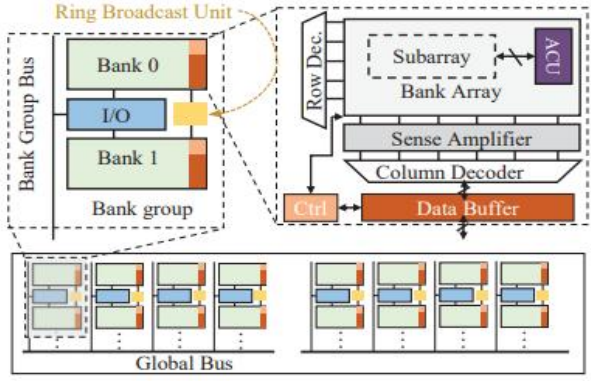
자료: Seshadri, Vivek, et al.(2017). p.6

32MB 규모의 벡터에 대해 NOT/AND/OR을 반복 실행하는 마이크로벤치마크에서, Ambit(8-bank)은 Intel Skylake CPU 대비 44.9배, GTX 745 대비 32.0배, HMC 2.0 대비 2.4배 높은 처리량을 보였고 3D-stacked DRAM에 통합한 Ambit-3D(256-bank)는 HMC 2.0 대비 9.7배까지 향상되고, 에너지는 25.1~59.5배 절감되는 모습을 보였다.

4) TransPIM: A Memory-based Acceleration via Software-Hardware Co-Design for Transformer(Zhou, et al.(2022))(HPCA'22)

해당 연구는 LLM의 큰 메모리 사용량과 낮은 데이터 재사용률에서 발생한 과도한 메모리 병목 현상으로 인해 연산기가 낮은 활용률로 동작하는 문제점을 해결하고자 토큰 기반 데이터 공유(token-based data sharding)과 in-bank auxiliary computing unit를 활용한 HBM-PIM 가속기를 제안하였다〔그림 4－80〕.

〔그림 4－80〕 TransPIM의 전체 개요도



자료: Zhou, Minxuan, et al.(2022). p.6

Token-based data sharding은 입력 토큰을 기반으로 메모리 내의 존재하는 bank를 할당해 병렬적으로 연산을 수행하고 연산이 끝난 뒤 인접한 bank에 재사용 가능한 데이터를 이동시키는 ring broadcast를 적용함으로써 데이터 재사용 비율을 크게 높였다. 그리고 bank 내부에 재구성 가능한 데이터 버퍼, broadcast unit 구현을 통해 인터커넥트 버스와 비동기적으로 데이터 이동을 가능하게 함으로써 데이터 로드 과정에서 발생하는 오버헤드

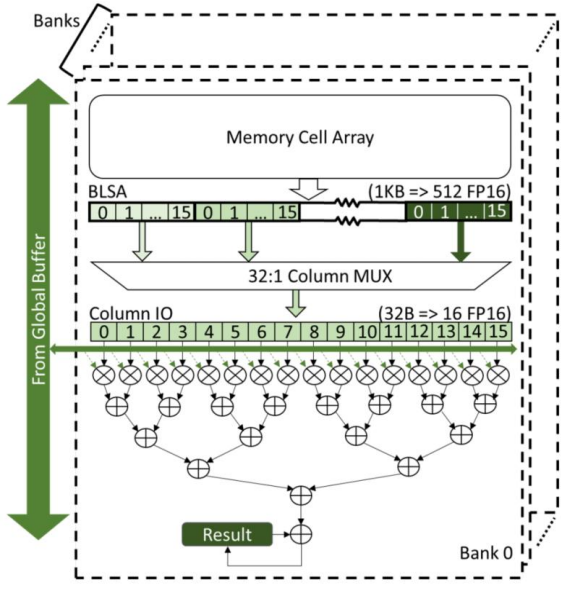
를 줄였다. In-bank auxiliary computing unit은 메모리 셀 행을 column-wise bit-serial 방식으로 읽어 adder tree를 거침으로써 뺄셈 연산을 처리하고 테일러 시리즈 확장 근사를 통해 메모리 내부에서 근사화된 softmax 함수를 처리할 수 있도록 구현하였다.

TransPIM는 Roberta모델에 대해 NVIDIA RTX 2080Ti GPU, Google TPU v1 대비 22.1~114.9배 / 8.7~57.4배 빠른 지연 시간을 가져 PIM 구조의 우수한 병렬처리 능력을 입증하였으며, 138.1~666.6배 / 39.5~376.7배 뛰어난 에너지 효율 수치를 보였고 Pegasus 모델에 대한 성능 실험을 통해 이전 연구 대비 2.0~3.3배 높은 최대 처리량 수치를 달성하였다.

5) Newton: A DRAM-maker's accelerator-in-memory(AiM) architecture for machine learning(He, Mingxuan, et al.(2020))(MICRO'20)

해당 연구는 딥러닝 추론에서 메모리와 호스트 사이의 협소한 채널 대역폭으로 인해 연산 자원이 비효율적으로 활용되는 문제를 해결하기 위해, DRAM 내부에 최소한의 연산기

〔그림 4－81〕 Newton의 bank별 logic 및 datapath 개요도

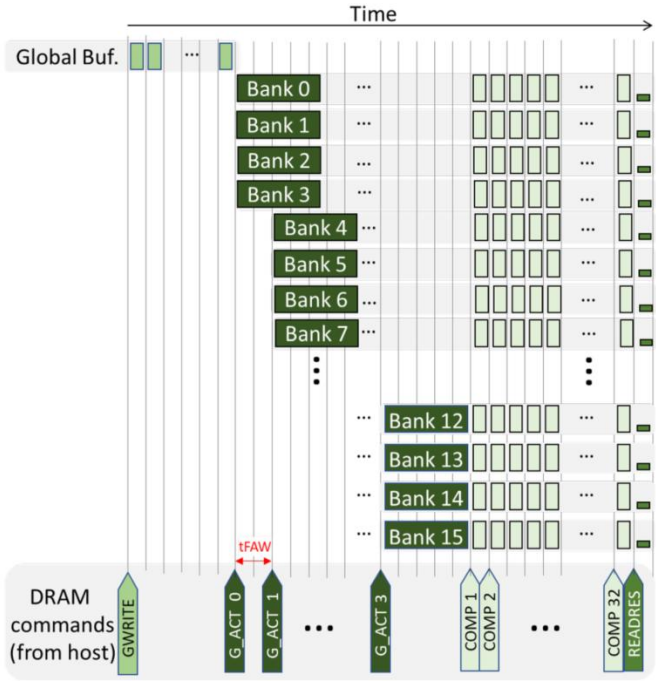


자료: He, Mingxuan, et al.(2020). p.4

와 버퍼만을 배치한 디지털 PIM 아키텍처를 제안한다. Newton의 각 bank에는 DRAM column I/O 대역폭(256-bit)과 rate-match 되도록 16개의 BF16 MAC 유닛과 파이프라인 된 adder tree가 포함되며, 최대 대역폭을 활용하여 연산을 수행한다〔그림 4－81〕.

Newton은 DRAM 내부 연산을 제어하는 과정에서 명령어 대역폭 병목을 줄이기 위해 여러 bank를 묶어 하나의 명령으로 동시에 동작시킨다. 예를 들어, G_ACT 명령은 네 개의 bank를 한 번에 활성화하고, 이후 COMP 명령은 동일한 입력 sub-chunk를 전체 bank에서 동시에 연산하도록 지시한다. 이러한 방식은 명령어 대역폭을 최대 16배 절감하며, 동기적 병렬 연산이 가능하게 한다〔그림 4－82〕.

〔그림 4－82〕 Newton 실행 시 timing diagram



자료: He, Mingxuan, et al.(2020). p.7

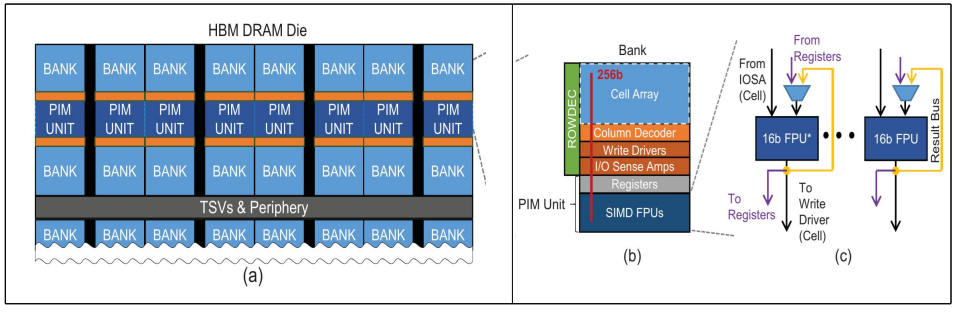
실험은 BERT, GNMT, DLRM 등 메모리 바운드 신경망의 fully-connected layer를 대상으로 수행되었으며, Newton은 HBM2E 기반 환경에서 Titan-V GPU 대비 평균 54배, 이상적

non-PIM 시스템 대비 10배 빠른 성능을 보였다. 또한, 일반 DRAM보다 약 2.8배 높은 전력을 소모하지만, 데이터 이동 감소와 병렬 연산 효율 향상으로 인해 전체 에너지 효율은 약 3.5배 개선되었다.

6) Hardware architecture and software stack for PIM based on commercial DRAM technology: Industrial product(Lee, Sukhan, et al.(2021))(ISCA'21)

해당 연구는 추천 모델 등에서 발생하는 높은 메모리 대역폭 요구와 낮은 데이터 재사용성으로 인해 시스템 전체 에너지의 대부분이 데이터 이동에 소모되는 문제를 해결한다. HBM-PIM은 각 DRAM bank의 I/O 경계에 PIM 연산 유닛을 배치하여 DRAM 내부의 병렬 대역폭을 직접 활용할 수 있어, 기존의 메모리-호스트 간 데이터 이동 병목을 근본적으로 완화한다. 또한, 모든 연산 제어는 기존 DRAM 명령(READ/WRITE)으로 수행되어 JEDEC 표준과 완벽히 호환되므로, 별도의 하드웨어 변경 없이 기존 시스템에 바로 적용할 수 있는 실용성이 큰 장점이다〔그림 4-83〕.

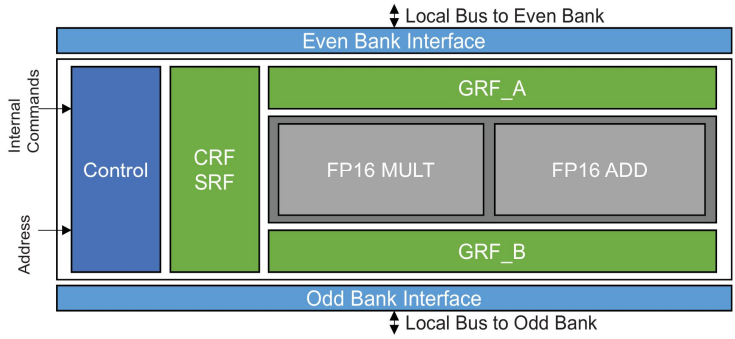
〔그림 4-83〕 HBM-PIM의 전체 구조도 및 datapath (a) HBM DRAM die, (b) PIM 유닛이 결합된 bank, (c) PIM datapath



자료: Sukhan, et al.(2021). p.3

〔그림 4-84〕의 PIM 실행 유닛은 두 개의 bank(Even/Odd)가 하나의 로직을 공유하도록 구성되어, 면적과 전력 소모를 최소화하면서도 bank 간 병렬성을 유지한다. 내부의 SIMD 연산기는 DRAM의 column access 주기와 동기화되어 동작하므로, 추가적인 파이프라인 지연 없이 메모리 접근과 연산을 병행할 수 있다.

〔그림 4－84〕 PIM 실행 유닛의 마이크로아키텍처



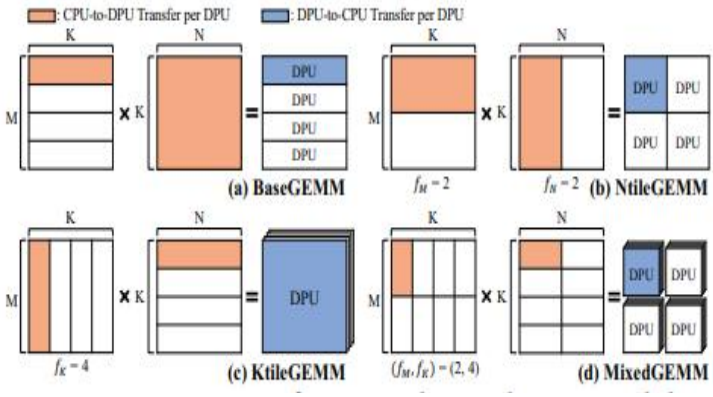
자료: Sukhan, et al.(2021). p.3

실험 결과, GEMV 연산에서 최대 11.2배, DeepSpeech2에서 3.5배, GNMT에서 1.5배, CNN의 fully-connected layer에서 1.4배의 성능 향상을 확인했다. 전력 소비는 기존 HBM 대비 약 5.4% 증가하지만 성능 개선 대비 미비한 수준으로 전체 에너지 효율성이 증가한다.

7) Accelerating LLMs using an Efficient GEMM Library and Target-Aware Optimizations on Real-World PIM Devices(Kim, et al.(2025))(CGO'25)

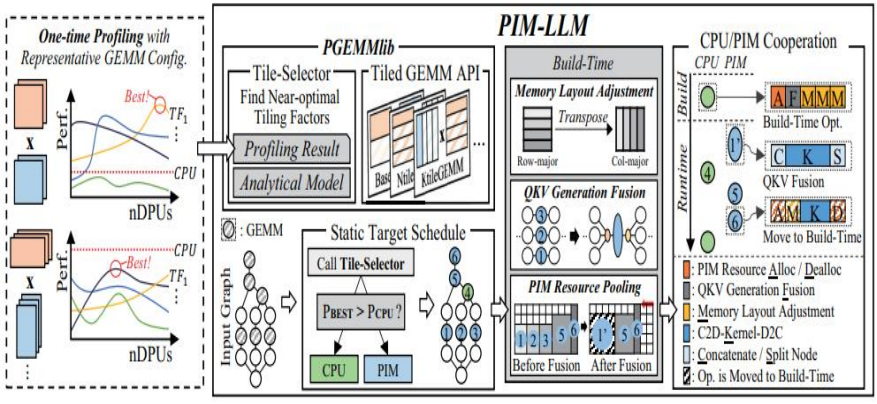
해당 연구는 PIM에 최적화된 타일링 기법을 제공하는 PGEMMlib 라이브러리와 LLM 실행의 최적화 기법을 포함하는 end-to-end 프레임워크인 PIM-LLM을 제안하였다. 타일링은 연산에서의 데이터 재사용률을 극대화하기 위해 행렬 곱 연산을 분해하여 가장 높은 데이터 재사용 비율을 가지도록 설계하여 가속기의 데이터 재사용 효율을 극대화할 수 있는 기법이다. 제안하는 PIM 가속 솔루션에서 제공하는 PGEMMlib은 라이브러리의 tile selector를 통해 주어진 DRAM 연산 처리 유닛과 GEMM의 행/열 크기를 기반으로 예측한 커널 실행 타임과 메모리 대역폭에 의해 결정되는 데이터 전송 시간을 고려해 가장 낮은 지연 시간을 달성할 수 있는 타일링 인자를 결정한다〔그림 4－85〕. 또한 LLM 추론 실행 과정에서 최적화 기법에는 QKV generation fusion, memory layout adjustment를 활용한다.

〔그림 4－85〕 타일링 방법에 따른 행렬 곱 분할 방법 예시



자료: Kim, et al.(2025). p.4

〔그림 4－86〕 PIM-LLM의 전체 동작 개요도



자료: Kim, et al.(2025). p.7

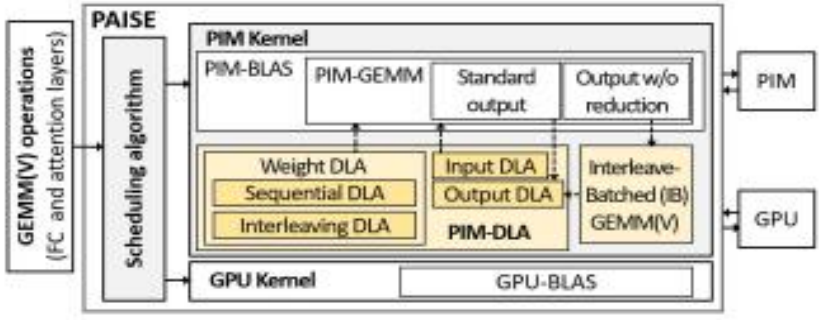
구체적으로, QKV generation fusion은 컴파일 과정 중 QKV generation의 각 연산이 독립적이라는 점을 이용해 통합된 GEMM 연산으로 병합함으로써 커널 수를 줄여 연산 시간을 감소시켰다. Memory layout adjustment는 빌드 과정 중 PIM에 적합한 memory layout으로 바꾸고자 파라미터를 행 단위 우선순위 기준 혹은 열 단위 우선순위 기준으로 변환해 저장함으로써 변환 시간과 전송 효율을 높였다〔그림 4－86〕.

PIM-LLM은 LLaMA-7B 모델에서 이전 연구 및 GPU 가속 연구 대비 각각 45.75배, 14.65배 우수한 처리량 향상을 보였다. 이러한 결과는 PIM 기반 시스템의 GPU 대비 우수한 처리량 및 에너지 효율성으로부터 저탄소/고효율의 시스템을 구현할 수 있는 솔루션임을 보인다.

8) PAISE: PIM-Accelerated Inference Scheduing Engine for Transformer-based LLM (Lee, et al.(2024))(HPCA'25)

해당 연구는 제한된 메모리 대역폭으로 인해 발생하는 메모리 병목 현상이 LLM의 decode 단계에서 크게 발생함을 확인하여, 이를 해결하고자 PIM 기반 프레임워크인 PAISE를 제안하였다. PAISE는 interleave batched GEMM 연산에 특화된 PIM 커널과 모델 및 PIM 하드웨어를 고려한 스케줄링 알고리즘으로 구성되어 있다〔그림 4-87〕.

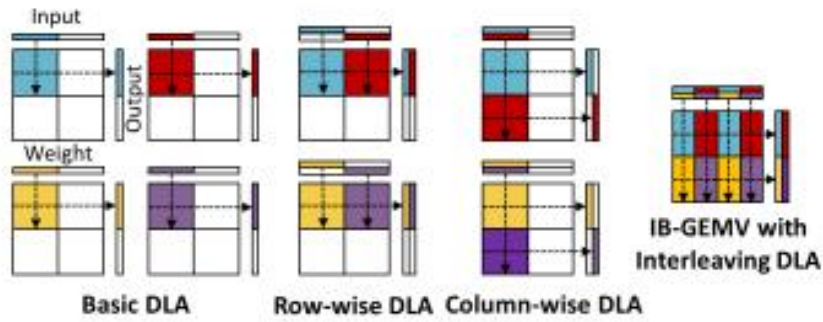
〔그림 4-87〕 PAISE 시스템의 개요도



자료: Lee, et al.(2024). p.4

구체적으로, PIM kernel은 data layout adjustment(DLA) 모듈을 통해 행/열 별 가중치 로드와 배치 크기를 고려한 열 단위 interleave 가중치 로드 기법을 지원함으로써 interleave-batched GEMM 연산을 효율적으로 가속할 수 있음을 보였다〔그림 4-88〕.

〔그림 4－88〕 DLA에 따른 GEMM 연산 흐름 개요도



자료: Lee, et al.(2024). p.5

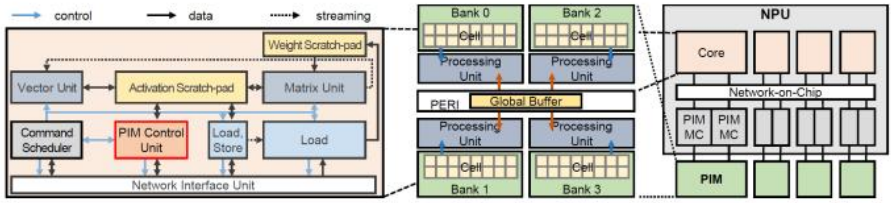
또한 스케줄링 알고리즘은 연산의 크기, 재사용 비율, DLA의 정렬 과정에서 발생하는 오버헤드, 메모리 효율 등을 수치화해 LLM 내 주요 메모리 병목 연산을 판단하고, PIM과 GPU 중 offload 대상 하드웨어를 결정한다. 즉, GPU과 PIM의 장점을 동시에 극대화할 수 있도록 LLM의 주요 병목을 분석한 후, 이에 따라 효율을 극대화할 수 있는 플랫폼에 해당 연산을 할당함으로써 효율적인 LLM의 추론 가속을 가능하게 한다.

PAISE를 AMD MI100 GPU와 HBM-PIM로 구성된 PIM-GPU 이기종 시스템을 이용해 대표적인 LLM 모델인 GPT-2와 LLaMA2-7B 모델에 대해 지연 시간을 측정하였다. 그 결과, MHA 연산을 GPU만 활용하여 연산했을 때 대비 PIM에 offload한 경우 최대 48.3%만큼의 추론 속도 향상을 보여 효율적으로 각 전체 시스템을 활용했음을 입증했다.

9) IANUS: Integrated Accelerator based on NPU-PIM Unified Memory System(Seo, et al.(2024))(ASPLOS'24)

해당 연구는 기존 NPU는 GEMM 연산에는 효율적이지만, GEMV 연산에서는 낮은 연산 집약도로 인해 자원 활용도가 떨어지고 반대로 PIM은 대역폭 집약적인 GEMV에 효과적이지만 연산 집약적인 대규모 GEMM을 처리하기에는 연산 자원이 부족하다는 한계를 극복하기 위해 NPU와 PIM을 결합한 이기종 가속 시스템 IANUS를 제안한다〔그림 4－89〕.

〔그림 4－89〕 NPU-PIM 아키텍처인 IANUS의 전체 아키텍처 개요도

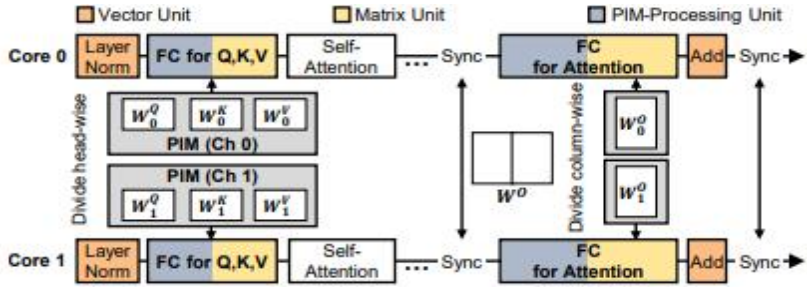


자료: Seo, et al.(2024). p.5

구체적으로, 가중치를 DRAM bank와 row 단위로 타일링하고 분해된 타일들을 각 PIM에 할당하여 병렬 접근이 가능하도록 구현하여 효율적인 GEMV를 가능하게 하였다. 또한 NPU, PIM 각각 효율적인 워크로드 분배를 위하여, 연산과 메모리 접근 과정에서 LLM의 병목 현상을 분석하고, 이를 기반으로 워크로드 스케줄링을 지원하는 PIM access scheduling을 제안한다〔그림 4－90〕.

이러한 통합 메모리/가속기 구조는 PIM 메모리를 NPU의 메인 메모리로 활용할 수 있도록 설계하여, 메모리 효율성을 높였다. PIM과 NPU의 연산 유닛을 효율적으로 사용하기 위해 PIM 메모리에 head 단위로 분할하여 연산에 필요한 파라미터를 할당하였으며, MHA 연산의 각 self-attention 연산을 파이프라이닝 하기 위해 병렬적으로 처리 가능한 연산들을 matrix unit과 PIM에 각각 할당하였다. 또한 정규화, 전치 연산과 같이 데이터 프리패치가 필요한 연산은 NPU의 벡터 연산 유닛에서 처리한다.

〔그림 4－90〕 IANUS의 NPU-PIM 워크로드 스케줄링 예시



자료: Seo, et al.(2024). p.7

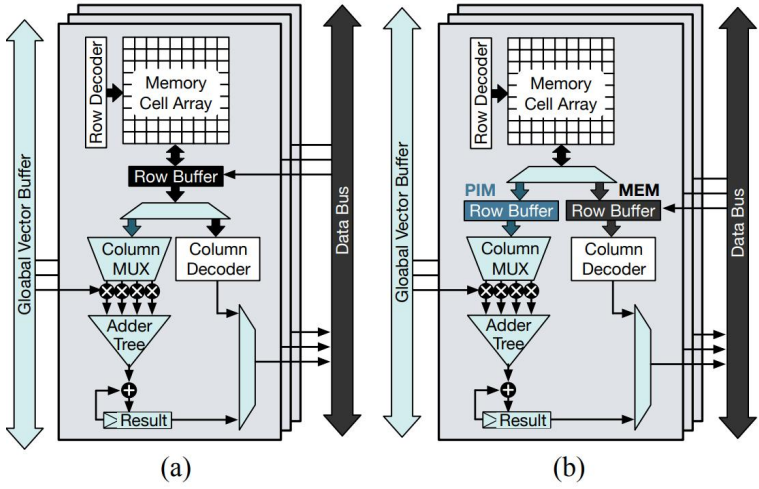
이때, 별도의 맵핑 알고리즘을 적용해 각 플랫폼별 처리 지연 시간을 기준으로 행렬 곱 연산 유닛과 PIM에 적절한 연산을 할당한다.

IANUS는 GPT2-M/L/XL/2.5B 모델에 대해 NVIDIA A100 GPU 대비 평균 4.3~11.3배의 적은 지연 시간으로 모델의 추론을 수행함을 확인하였으며, 특히 GPT2-XL 모델에 대해 Multi-FPGA system인 DFX(Hong, et al.(2022)) 대비 평균 3.2배 우수한 지연 시간 개선을 달성하였다.

10) NeuPIMs: NPU-PIM Heterogeneous Acceleration for Batched LLM Inferencing(Seo, et al.(2024))(ASPLOS2'24)

해당 연구는 IANUS(Seo, et al.(2024))에서와 마찬가지로, NPU와 PIM의 각각의 장점을 활용하기 위해 NeuPIMs 시스템을 제안한다. NeuPIMs의 각 메모리 bank에는 두 개의 row buffer가 존재하며, 하나는 일반적인 메모리 접근, 다른 하나는 GEMV 연산을 위해 사용된다. 이 dual row buffer 구조를 통해 NPU의 메모리 접근과 PIM 연산을 동시에 수행할 수 있어, 기존 PIM의 blocked mode(호스트와 PIM이 동시에 동작할 수 없음) 문제를 해결한다〔그림 4-91〕.

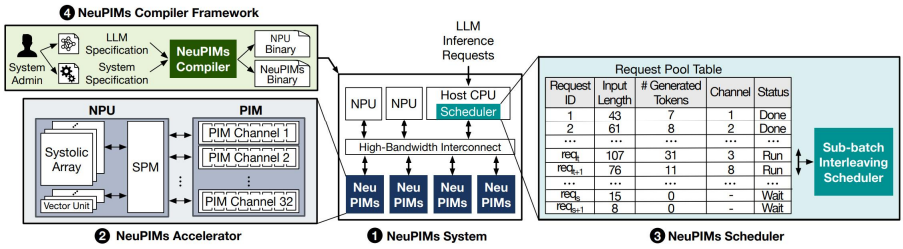
〔그림 4-91〕 NeuPIMs의 bank별 logic 개요도



자료: Heo, et al.(2024). p.7

NeuPIMS 시스템은 다수의 NeuPIMs 디바이스와 독립적인 NPU를 고대역폭 인터커넥트(PCIe, CXL 등)과 연결하며 sub-batch interleaving 스케줄링을 통해 서로 다른 하위 배치를 NPU와 PIM에서 교차 실행시킨다〔그림 4－92〕.

〔그림 4－92〕 NeuPIMs 시스템의 개략도



자료: Heo, et al.(2024). p.6

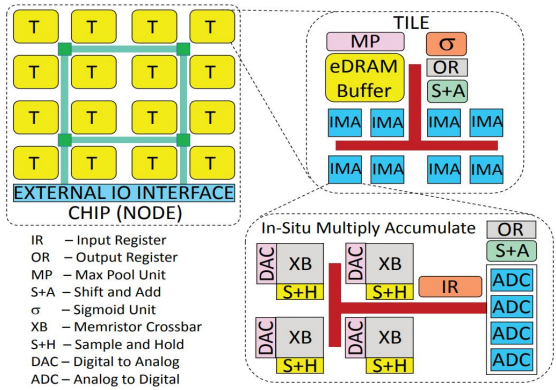
NeuPIMs는 다양한 크기의 GPT-3 모델(7B / 13B / 30B / 175B)을 대상으로 한 실험에서 GPU-only, NPU-only, 그리고 단순 NPU와 PIM을 통합한 시스템 대비 각각 3.0배, 2.4배, 1.6배의 처리량 향상을 보였다. 자원 활용률 또한 NPU는 12.3%에서 64.9%로, PIM은 17.0%에서 26.4%로 개선되었으며 동일한 메모리 용량에서 약 2.4배 높은 성능과 25% 낮은 에너지 소비를 기록했다.

11) ISAAC: A Convolutional Neural Network Accelerator with In-Situ Analog Arithmetic in Crossbars(Shafiee, Ali, et al.(2016))(CAnews'16)

ISAAC은 가중치를 저장하는 memristor crossbar를 그대로 아날로그 dot-product engine으로 활용해, MAC 연산을 메모리 내부에서 수행하는 in-situ 연산 기반 CNN 가속기다. 각 레이어에 전용 crossbar를 배치하고, 레이어 사이에는 소형 eDRAM 버퍼만 두는 레이어-파이프라인을 구성하여 대규모 데이터 이동을 줄이고 처리 효율을 높였다. ISAAC의 각 타일은 다수의 In-situ Multiply-Accumulate(IMA) 유닛으로 구성되어 있으며, 각 IMA는 crossbar(XB), 샘플·홀드(S+H), 아날로그-디지털 변환기(ADC), 그리고 shift-and-add(S+A) 블록으로 이루어진다. 이때 crossbar의 row에는 입력 전압을 인가하고, 각 교차점 cell의 도전도로 가중치를 표현하여 곱셈이 아날로그로 수행되며, columns에 합산된 전류는 S+H

와 ADC를 거쳐 디지털 누산기로 전달된다. 이러한 아날로그 도메인에서의 직접 곱셈 및 합산(in-situ MAC) 구조를 통해 별도의 연산 유닛으로 데이터를 이동시킬 필요가 없어, 데이터 이동 에너지를 크게 절감하고 연산 밀도를 극대화한다〔그림 4-93〕.

〔그림 4-93〕 ISAAC의 아키텍처 개략도



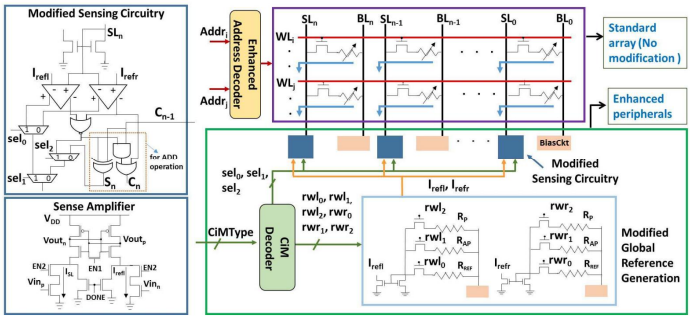
자료: Shafiee, Ali, et al.(2016). p.6

ISAAC은 단일 칩 기준 면적 효율은 약 1.73 TOPS/mm², 전력 효율은 400 GOPS/W, 저장 효율은 34 TOPS/TB를 보이며 Titan X GPU 대비 24.3배 높은 처리량과 84배 높은 에너지 효율을 보였다.

12) Computing in Memory With Spin-Transfer Torque Magnetic RAM(Shubham, et al. (2018))(VLSI'16)

이 연구는 STT-MRAM의 비휘발성 특성과 높은 내구성을 기반으로, 기존의 1T-1R 셀을 그대로 사용하면서 감지 회로와 참조 전류만을 수정하여, 한 번의 메모리 접근으로 AND, OR, XOR, ADD 연산을 처리할 수 있도록 한다. 두 개의 주소 디코더를 OR 결선하여 두 워드라인을 동시에 활성화하고, 감지 입력에 연산별 참조 전류를 공급해 한 번의 column access로 bitwise 연산/올림 신호를 구한다〔그림 4-94〕.

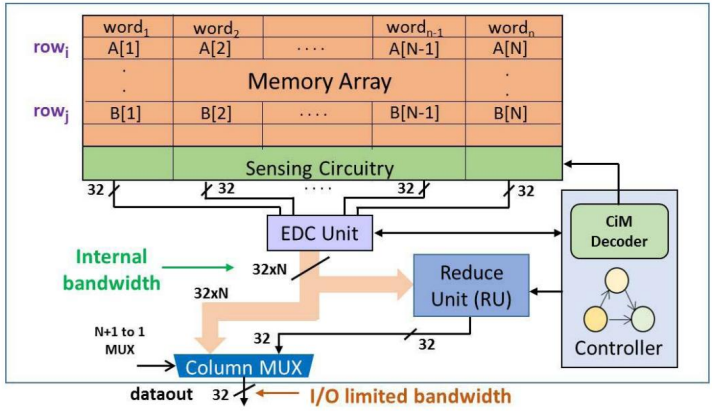
〔그림 4－94〕 STT-CiM의 셀 어레이 구조



자료: Shubham, et al.(2018). p.5

배열 내부는 32×N data width로 동작하지만, 외부 data width는 32-bit로 제한된다. 이를 보완하기 위해 column MUX 앞에 EDC Unit과 Reduce Unit(RU)를 배치, N개 부분 결과를 sum / norm / compare 등으로 축약해 한 번의 배열 접근으로 최종 연산결과를 생성한다〔그림 4－95〕.

〔그림 4－95〕 STT-CiM의 벡터 연산 개요



자료: Shubham, et al.(2018). p.7

실험 결과, STT-CiM은 동일한 메모리를 연산 기능 없이 Nios Ⅱ CPU를 사용하는 기존 시스템과 비교했을 때, 동일한 연산을 수행할 때 메모리 접근 에너지는 평균 3.8배, 최대

12.4배 감소, 전체 실행 시간은 평균 3.9배, 최대 10.4배 단축되었다.

3. 하드웨어-소프트웨어 공동설계

가. 하드웨어-소프트웨어 공동 설계 필요성

기존의 AI 가속기 설계 방식은 하드웨어와 소프트웨어를 독립적으로 개발하는 경우가 많았다. 그러나, 이러한 독립적 설계는 실제 응용 환경에서 설계 목표 대비 성능과 전력 효율이 저하되는 문제가 빈번히 발생한다. 즉, GPU와 같은 범용 하드웨어의 배포 과정을 고려하지 않은 소프트웨어 알고리즘을 하드웨어 상에서 구동하는 과정에서는 모델 구조, 연산 정밀도, 메모리 접근 패턴 등에서 비효율적인 연산을 요구하며, 이로 인한 하드웨어 자원과 모델 특성 간의 불균형이 발생한다.

하드웨어-소프트웨어 공동 설계는 이러한 문제를 해결하기 위한 핵심 전략이다. AI 모델의 구조와 하드웨어 아키텍처를 상호 보완/최적화하면, 성능과 전력 효율을 동시에 높일 수 있다. 예를 들어, 모델 정밀도를 4-bit 수준의 low-bit 정수형으로 축소하고, 이에 최적화되어 동작하는 전용 연산 유닛을 설계하면 연산량, 연산기의 면적, 데이터 이동량 등의 측면에서 이점을 가질 수 있다.

또한 LLM 모델의 파라미터, 특징맵의 대부분은 많은 비율의 0을 포함하므로, 가중치, 특징맵 분포에서 0의 비율을 의미하는 모델의 희소성을 분석하고, 효율적인 압축 규격을 정한 뒤 이를 하드웨어 상에서 디코딩, 연산 가속을 지원하는 하드웨어-소프트웨어 공동 설계 또한 적은 전력 소모량, 면적, 우수한 처리량을 가질 수 있다. 이러한 최적화는 데이터센터, 고성능 컴퓨팅 등 다양한 환경에서 저탄소·고효율 디지털 대전환을 가속화할 수 있는 솔루션이다.

이와 더불어, Transformer 및 대규모 딥러닝 모델의 추론 효율을 결정짓는 핵심 요소 중 하나는 비선형 함수의 계산 방식이다. Softmax, GELU, LayerNorm과 같은 비선형 함수들은 연산량 자체는 전체 모델에서 차지하는 비중이 크지 않지만, 곱셈, 나눗셈, 지수, 제곱근 연산과 같은 floating-point 기반의 고비용 연산을 포함하고 있어, 실제 하드웨어 구현 시 연산 병목뿐 아니라 과도한 하드웨어 자원 사용과 전력 소모를 초래한다. 특히 이러한 함수들은 대부분 piecewise linear(PWL) 기반 LUT 근사 또는 다항식 근사 방식으로 대체

되지만, 입력 범위의 변화나 양자화 환경에서의 정밀도 손실로 인해 성능 저하가 발생하기 쉽다. 결국, 비선형 함수 근사화 기법은 고비용 연산을 정수 기반의 경량 연산으로 대체함으로써, FPGA 및 ASIC과 같은 가속기에서 연산기 밀도 향상, 배선 및 메모리 접근량 감소, 그리고 전력 소모 절감 효과를 동시에 달성할 수 있다. 이를 통해 한정된 하드웨어 자원 내에서 더 높은 병렬 처리 성능과 에너지 효율을 확보할 수 있으며, 결과적으로 모델의 온디바이스·엣지 추론 환경에 최적화된 하드웨어 구현을 가능하게 한다.

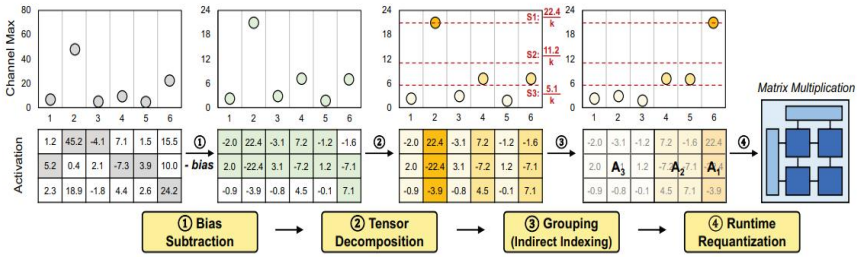
나. low-bit 양자화-호환 연산기 공동 설계 연구 동향

1) Tender: Accelerating large language models via tensor deocomposition and runtime requantization(Lee, et al.(2024))(ISCA'24)

Tender는 LLM 양자화 과정에서 low-bit에서 동작하며 열 단위 세분화 양자화의 우수한 성능을 하드웨어 상에서 효율적으로 구현하기 위한 decomposed quantization 기법을 제안하며, 이를 효율적으로 배포하기 위하여 MAC, rescale operation을 모두 지원하는 연산기 구조를 제안한다.

구체적으로, Tender의 decomposed 양자화 연산 흐름은 bias subtraction, tensor decomposition, grouping, runtime requantization으로 구성된다. 이 과정을 통하여 특징맵을 2의 배수 기반 스케일링 팩터와 함께 분류한 후, 이상치 그룹의 연산 수행을 위해 비이상치의 연산에서부터 점진적으로 누적 값을 시프트 연산자로 매 단계 스케일링한다〔그림 4－96〕. Tender는 이러한 하드웨어-소프트웨어 공동 설계를 통하여, OPT, LLaMA-2 모델에서 이전 연구 대비 평균 2.63배 우수한 처리량 성능을 보였다. 또한 이전 공동 설계 연구인 Olive(Guo, et al.(2023)) 대비 에너지 효율성 지표에서 1.24배 뛰어남을 보이며 제안하는 하드웨어-소프트웨어 공동 설계의 우수성을 입증하였다.

〔그림 4－96〕 Tender의 decomposed 양자화 연산 흐름 개요도



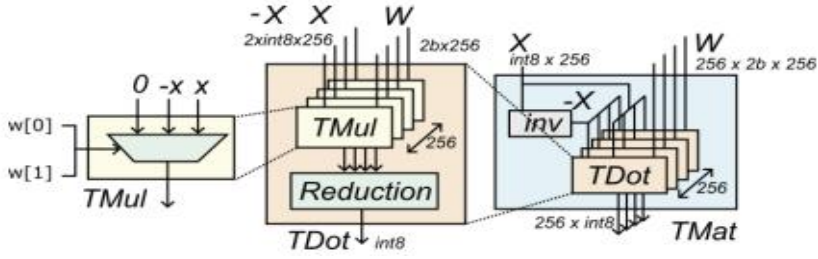
자료: Lee, et al.(2024), p.5

2) TerEffic: Highly Efficient Ternary LLM Inference on FPGA(Yin, et al.(2025)) (Arxiv'25)

해당 연구는 -1, 0, 1 으로만 데이터를 표현, 저장하는 형식인 ternary 데이터 형식을 완전하게 활용할 수 있는 메모리 아키텍처와 ternary matrix multiplication unit을 제안한다.

구체적으로, ternary matrix multiplication unit은 내부적으로 256개의 Ternary Dot(TDot) 유닛이 있고, 각 TDot 유닛은 256개의 Ternary Multiplication(TMul) 유닛으로 내적 연산을 수행한다. 이때, 각각의 TMul 유닛 내부에서 가중치에 의한 변환 연산 시, 효율성이 떨어지기 때문에 TMat 유닛에서 기존 입력값과 미리 연산을 해 놓은 반대의 값을 모두 TDot에 입력으로 넣고, 이를 바탕으로 TMul에서 가중치의 bit를 mux 신호로 출력을 결정한다〔그림 4－97〕.

〔그림 4－97〕 Ternary Matrix Multiplication(TMat) Core 개요도



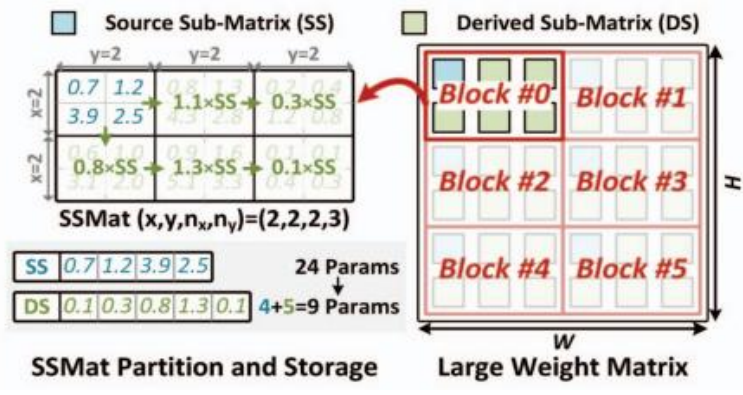
자료: Yin, et al.(2025), p.4

TerEffic은 Ternary 기반 경량화와 하드웨어 최적화를 통해 LUT 사용량을 13.2% 줄였다. Alveo U280 FPGA에 구현된 가속기는 2.7B Ternary 모델에서 NVIDIA A100보다 3배 빠른 초당 727 토큰 생성 속도와 8배 높은 16 token/sec/W의 전력 효율을 달성함으로써, 하드웨어-소프트웨어 공동 설계의 뛰어난 효과를 입증했다.

3) MECLA: Memory-compute-efficient llm accelerator with scaling sub-matrix partition (Qin, et al.(2024))(ISCA'24)

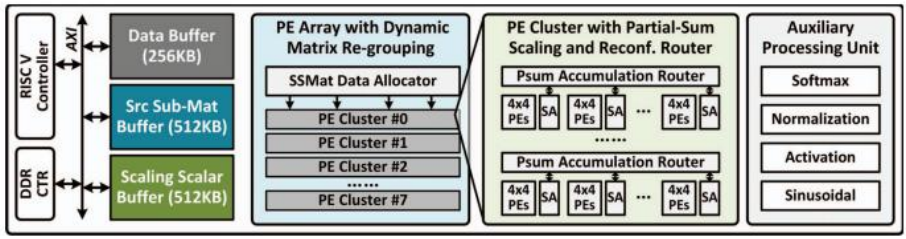
해당 연구는 LLM의 과도한 가중치 메모리에서 발생하는 오버헤드를 줄이기 위하여, 가중치 행렬의 일부와 적은 양의 스케일링 파라미터만으로 전체 가중치 행렬을 표현하는 Scaling Sub-Matrix Partition(SSMP)를 제안한다. 구체적으로, SSMP는 가중치 행렬을 Source Sub-matrix(SS), Derived Sub-matrix(DS)로 분할하고, SS를 스케일링하여 여러 DS를 생성해 낸다〔그림 4－98〕. 즉, 가중치 행렬의 일부만 오프칩에서 로드하고, SS를 기반으로 여러 DS를 온 칩에서 생성함으로써 오프 칩에서 로드할 파라미터의 수를 대폭 감소시킨다.

〔그림 4－98〕 MECLA 연구에서 제안하는 SSMP 기법 개요도



자료: Qin, et al.(2024), p.5

〔그림 4－99〕 MECLA processor의 전체 구조 개요도



자료: Qin, et al.(2024), p.6

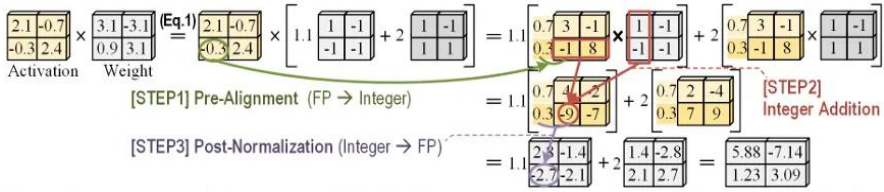
SSMP 기법으로 인해 생기는 오차를 최소화하기 위해, grid search와 successive halving 기법을 활용해 task에 따라 SS, DS의 크기를 다르게 설정하고 스케일링 파라미터는 파인튜닝을 통해 최적화한다. 또한 MECLA는 4x4 연산 구조와 스케일링 누적기가 포함된 프로세싱 유닛을 활용하여 SSMP 기법을 효율적으로 구현할 수 있는 하드웨어 구조를 제안하였다〔그림 4－99〕. 입출력 차원에 따라 inner/outer 연산 방식을 선택함으로써 데이터 재사용을 극대화하며, 이를 지원하기 위해 on-the-fly matrix regrouping과 re-association 기법을 함께 제안한다. SSMP의 연산량 감소 및 메모리 접근 감소 효과를 확인하기 위해 LLaMA-2 등의 모델에서 CoLA, wikitext2, QNLI 등 다양한 task에 실험을 진행한 결과, 이전 경량화 연구 대비 평균적으로 2.9배 우수한 연산량 감소량과 69.2%의 메모리 접근 횟수 감소율을 달성하였다. 또한 SSMP가 적용된 MECLA 가속기는 7,088 GOPS/W의 전력 효율 수치를 달성하여 Spatten(Wang, et al.(2021)) 대비 각각 18.6배 우수한 에너지 효율 수치를 보여 하드웨어-소프트웨어 공동 설계에서 메모리 최적화 알고리즘의 성능 개선 효과를 확인한다.

4) Winning Both the Accuracy of Floating Point Activation and the Simplicity of Integer Arithmetic(Kim, et al.(2023))(ICLR'23)

해당 연구는 FP activation과 양자화된 weight를 대상으로, FP MatMul 연산을 INT 연산으로 재구성하여 하드웨어 상에서 효율적인 DNN 연산을 수행할 수 있는 FP-INT 연산기 구조를 제안한다. 구체적으로, FP MAC 연산을 INT 연산으로 치환하기 위해 weight matrix를 binary-coded matrix와 scaling factor로 분해하고, 각 bit plane 별로 activation과 weight 간의 곱셈을 weight 부호에 따라 덧셈 또는 뺄셈 연산으로 대체한다. 이후 각 bit plane의

결과를 병합할 때는 소수의 scaling factor 곱셈만 수행하면 되므로, 전체 MatMul을 기존 FP MatMul 대비 극히 적은 수의 FP 곱셈만으로 처리할 수 있게 된다. 특히 제안된 방식은 연산 중 pre-alignment 단계를 도입하여, 각 dot product 단위로 입력 activation의 exponent를 공유하는 shared exponent를 사용함으로써 significand를 가장 큰 exponent를 기준으로 align 한다〔그림 4－100〕. 이를 통해 기존 FP 연산마다 수행해야 했던 alignment 과정을 MatMul 단위로 한 번만 수행하게 되어, 연산 오버헤드를 크게 줄였다. 또한 기존 연구인 MSFP(Rouhani et al., 2020)가 shared exponent를 저장하고 이를 불러와 연산하는 구조인 것과 달리 본 연구에서는 연산 중에 exponent를 공유하여 기존 FP 표현력의 손실을 최소화하였다. 이렇게 정렬된 activation은 INT 덧셈을 통해 accumulate 되며, 이후 normalization 과정을 거쳐 다시 FP 값으로 복원된다. 다만 pre-alignment 시 significand bit 수가 과도하게 커지면 INT 연산의 장점이 감소할 수 있으므로, 제안된 구조에서는 significand를 기존 significand bit 수 + bitwidth(2-bit)으로 truncation 하여 연산 오차와 하드웨어 자원 효율 간의 균형을 유지한다. 이를 통해 기존 FP 연산과 유사한 수준의 오차를 유지하면서도, 덧셈 시 필요한 에너지를 크게 절감한다. 제안된 iFPU는 binary 양자화, shared exponent, truncation 기반 정밀도 조절을 통합하여, 양자화된 DNN 모델을 효율적으로 처리할 수 있는 하드웨어를 구현하였다. BERT-base, VGG-9, ResNet-18, ResNet-50, RegNet-3.2GF, MnasNet-2.0, 그리고 OPT-1.3B 등 7가지 DNN 모델을 대상으로 실험한 결과, FP32 및 BF16 activation을 사용할 경우 기존 FPU와 유사한 수준의 DNN 정확도를 유지하였다. 또한 28nm CMOS 기술로 합성한 결과, FP Add와 FP Mul로 구성된 기존 FP-MAC 및 FP-ADD 대비, BF16 activation 기준 최대 9.9배의 throughput-per-area 향상과 최대 11.9배의 energy efficiency 향상을 달성하여, 제안된 iFPU의 우수성을 입증하였다.

〔그림 4－100〕 iFPU 전체 연산 흐름 개요도

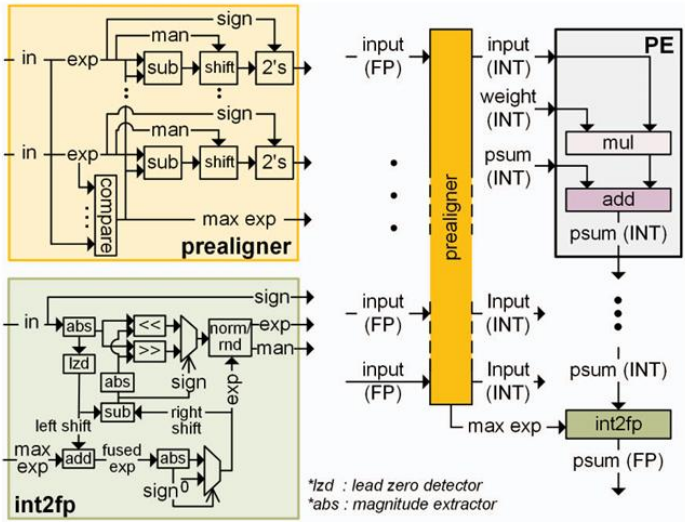


자료: Kim, et al.(2023). p.4

5) FIGNA: Integer Unit-based Accelerator Design for FP-INT GEMM Preserving Numerical Accuracy(Jang, et al.(2024))(HPCA'24)

해당 연구는 LLM의 많은 연산량으로 인한 메모리 및 연산 부담을 줄이기 위해 주로 사용되는 weight-only 양자화를 적용함으로 인해 발생하는 FP-INT 연산을 INT 기반 연산으로 재구성하여 연산을 수행하는 하드웨어 구조를 제안한다〔그림 4-101〕. Shared exponent 개념을 활용하여 INT 연산을 수행하면서도 기존 FP 연산 수준의 수치 정확성을 유지할 수 있는 FP-INT GEMM 가속기를 제안하였다. 이 방법은 기존 Block Floating Point(BFP) 방식의 한계를 극복하기 위해, 계산 과정 중에 지수를 공유하는 on-the-fly exponent sharing 기법을 도입하여 별도의 재학습이 필요 없게 한다. 구체적으로, FIGNA는 pre-alignment 단계를 통해 FP activation의 mantissa를 동일한 exponent scaling으로 정렬하여, FP-INT MAC 연산을 INT-INT 기반의 MAC 연산으로 수행할 수 있게 된다. Weight 행렬과의 모든 내적 연산이 완료되면 INT 결과를 normalization을 거쳐 다시 FP 형식으로 복원한다〔그림 4-102〕. 이를 통해, 기존 FP 연산마다 수행되던 align 및 normalization 과정을 MatMul 단위로 한 번만 수행하도록 하여, 연산 오버헤드를 감소시킨다. 다만, pre-aligned mantissa

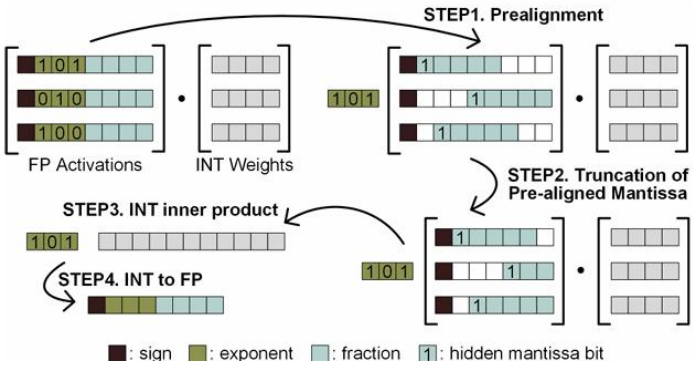
〔그림 4-101〕 FIGNA 전체 구조 개요도



자료: Jang, et al.(2024). p.4

의 bitwith가 과도하게 커지는 경우 곱셈기 크기와 에너지가 INT 기반 연산 수행을 통해 제거한 align 및 normalization 오버헤드 절감 효과보다 더 큰 문제를 발생시킬 수 있기 때문에 FIGNA는 pre-aligned mantissa의 bitwith를 줄이면서도 기존 FP-INT GEMM 연산의 정확성을 유지할 수 있는 mantissa truncation을 적용하여 정밀도와 효율 간의 균형을 조정한다.

〔그림 4－102〕 FIGNA 전체 연산 흐름 개요도



자료: Jang, et al.(2024). p.4

또한, INT weight 내의 0 값이 연산 오류를 유발하는 문제를 해결하기 위해, 기존의 signed INT 형식과 호환되면서도 0 값을 제거하는 0-less signed INT 형식을 새롭게 제안한다. 이를 통해 곱셈 결과의 scale 변화를 예측할 수 없는 문제를 해결하고, INT 기반 FP-INT MAC 연산에서도 수치적 안정성을 확보한다. 추가로, chunk-based mantissa allocation 방식을 도입하여 사용되지 않는 bit를 제거하고, 유효한 mantissa chunk만을 동적으로 할당함으로써 곱셈기 크기를 최소화한다. 이를 통해, BF16 activation 기준으로 FIGNA PE는 기존 FP-FP PE 대비 area 및 power 효율을 각각 2.27배, 2.04배 향상시켰고 FP32 activation의 경우 최대 89%의 area 및 power 절감을 달성하였다. 제안된 FIGNA 아키텍처는 Google TPU의 2D systolic array 구조를 기반으로 설계된 weight-stationary dataflow 방식을 채택하여, 데이터 재사용률을 극대화하고 연산 효율을 향상한다. 또한 unified buffer, weight buffer, accumulation buffer로 구성된 계층적 버퍼 구조를 통해 데

이터 이동 오버헤드를 최소화한다. 결과적으로 FIGNA는 pre-alignment, mantissa truncation, 0-less INT 형식, chunk-based 할당 기법을 통합하여 FP-INT GEMM 연산에서 FP 수준의 수치 정확성을 유지하면서도 하드웨어 자원 소모를 획기적으로 절감할 수 있음을 입증하였다. Activation FP32, FP16, BF16과 weight INT4, INT8을 사용하여 OPT, BERT, BLOOM family 모델에서 실험을 진행한 결과, 기존 naive한 FP-FP engine 및 FP-INT engine과 비교하였을 때 최대 14.3배 더 높은 area efficiency를 보임을 확인하였다. Energy efficiency의 경우, FIGNA가 모든 경우에서 가장 높은 성능을 달성하였다.

6) FIGLUT: An Energy-Efficient Accelerator Design for FP-INT GEMM Using Look-Up Tables(Park, et al.(2025))(HPCA'25)

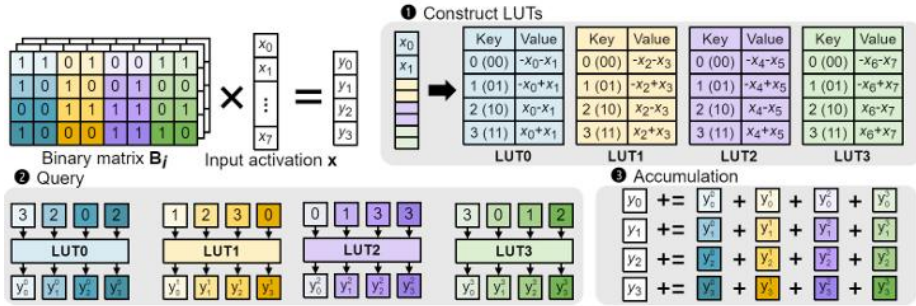
해당 연구는 FP-INT GEMM 연산에서 발생하는 연산량과 전력 소모를 줄이기 위해, 연산을 Look-Up Table(LUT)로 대체하는 LUT 기반 FP-INT GEMM 가속기 구조 FIGLUT를 제안한다. FIGLUT은 기존 LUT 기반 연산에서 병렬 접근 시 발생하는 bank conflict 문제를 제거하기 위한 전용 LUT 구조를 설계하고, fan-out에 따른 전력 증가를 최소화하는 최적의 PE 구조를 도입하여 높은 에너지 효율을 달성한다. 구체적으로, FIGLUT은 입력 activation과 binary weight 행렬을 기반으로 가능한 모든 partial sum을 사전에 계산하여 LUT에 저장하고, 연산 시 LUT에 저장된 값을 조회하여 결과를 얻는 방식으로 연산을 수행한다〔그림 4－103〕. LUT의 key는 μ개의 binary weight로 구성되며, μ 값이 커질수록 한 번의 조회로 대체되는 연산 수가 증가하지만, LUT 생성 시 필요한 메모리 크기가 커지기 때문에 최적의 μ를 설정하는 것이 중요하게 된다.

하드웨어 아키텍처 측면에서, FIGLUT은 Google TPU와 유사한 2D systolic array 구조를 기반으로 하며, weight-stationary dataflow를 적용하여 weight를 고정하고 activation 값만 순차적으로 전달한다. 각 PE는 LUT와 Read-and-Accumulate(RAC) unit 들로 구성되며, RAC unit은 기존의 MAC unit을 대체하여 LUT에서 key를 기반으로 값을 읽고 누적하는 역할을 한다. 특히, FIGLUT은 하나의 LUT을 여러 RAC이 공유할 수 있도록 설계되어, μ 값과 fan-out(k) 간의 power-performance trade-off를 분석하여 최적의 구조인 k=32, μ=4를 사용한다. 기존의 register file 기반 LUT(RFLUT)는 port 수의 제약으로 병렬 접근 시 높은 전력 소모와 bank conflict가 발생한다. 이를 해결하기 위해 FIGLUT는 flip-flop 기반

LUT(FFLUT) 구조를 제안한다. FFLUT은 multiplexer를 통해 여러 RAC이 동시에 접근할 수 있도록 구성되어 bank conflict 없이 병렬 조회를 지원하며, FP 덧셈 대비 낮은 전력으로 LUT 데이터를 읽을 수 있다. 또한 LUT 크기와 전력을 절반으로 줄이기 위해 half-LUT(hFFLUT) 기법을 도입하였다. hFFLUT은 LUT 내 부호만 반대인 값들이 존재하는 특성을 이용하여 값의 절반만 저장하고, key의 최상위 bit를 이용해 부호 반전을 수행함으로써 모든 값을 생성한다. 이를 통해 LUT 전체 전력 소비를 약 50% 감소시켰으며, 이를 위해 필요한 추가적인 회로의 오버헤드는 전체 전력 대비 미미한 수준으로 분석된다.

결과적으로, FIGLUT은 FFLUT 및 hFFLUT 기반의 conflict-free LUT 구조, 다중 RAC 공유 최적화를 통하여, FP-INT GEMM 대비 높은 에너지 효율과 낮은 연산 오버헤드를 달성한다. FIGLUT은 FP 기반 가속기 대비 전력 소모를 절반 이하인 약 52%로 줄이면서도 처리량을 유지하였으며, weight를 3-bit로 양자화하는 경우 기존 SoTA 가속기 대비 59% 높은 TOPS/W와 20% 낮은 perplexity를 달성한다.

〔그림 4－103〕 LUT 기반의 FIGLUT 연산 개요도



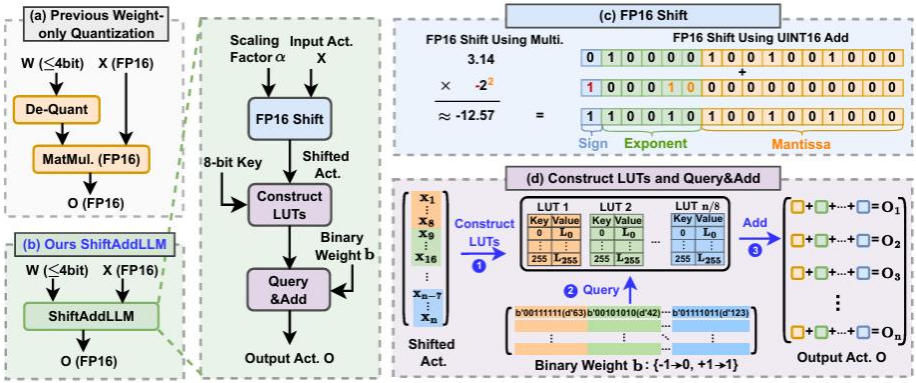
자료: Park, et al.(2025). p.4

7) ShiftAddLLM: Accelerating Pretrained LLMs via Post-Training Multiplication-Less Reparameterization(You, et al.(2024))(Neurips'24)

해당 연구는 사전 학습된 LLM을 GPU 상에서 실행 가능한 multiplication-free 구조로 re-parametrization 하여, 기존의 고비용 FP MAC 연산을 shift 및 add 연산으로 완전히 대체할 수 있는 ShiftAddLLM 기법을 제안한다. ShiftAddLLM은 binary-coding quantization

(BCQ)를 기반으로, weight 행렬을 여러 개의 binary 행렬과 대응하는 scaling factor로 분해하고 각 scaling factor를 2의 거듭제곱 형태로 양자화하여 모든 곱셈을 하드웨어 상에서 곱셈에 비해 간단한 shift 연산으로 대체한다. Activation은 FP16으로 유지되며, scaling factor에 따른 shift 연산과 binary weight의 부호를 활용한 덧셈과 뺄셈만 수행된다.

〔그림 4－104〕 ShiftAddLLM 전체 연산 흐름 개요도



자료: You, et al.(2024). p.4

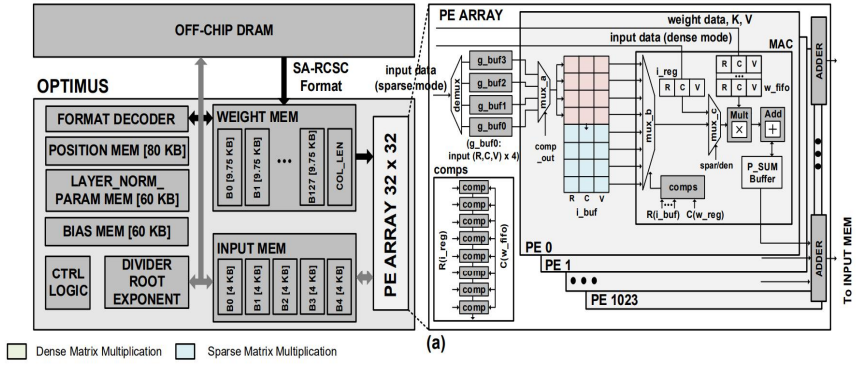
또한, 기존 weight-only 양자화 기법들이 weight objective 또는 activation objective 중 하나만을 최적화 대상으로 삼는 한계를 극복하기 위해 ShiftAddLLM은 multi-objective optimization을 도입하여 weight와 activation 양쪽에서 발생하는 re-parametrization 오류를 동시에 최소화한다. 이를 통해 정확도를 유지하면서도 연산량과 에너지 소모를 줄일 수 있다. 추가로 column-wise 및 block-wise scaling factor를 제안하여, 정확도와 지연 시간 사이의 균형에 따라 적절한 scaling factor를 사용한다.

하드웨어 효율 향상을 위해, ShiftAddLLM은 LUT기반의 연산 구조를 도입한다. LUT의 key는 binary weight 조합으로, value는 미리 scaling factor 값을 이용해 계산된 shift 된 activation의 partial sum으로 구성된다. 이를 통해 weight-activation 곱셈을 LUT query와 덧셈으로 대체할 수 있으며, GPU 상에서도 dequantization 과정 없이 연산을 수행할 수 있게 된다〔그림 4－104〕. 또한, 각 layer의 re-parametrization sensitivity를 분석하여, 손실에 민감한 layer에는 상위 bit를, 덜 민감한 layer에는 하위 bit를 자동으로 할당하는 mixed-bit

allocation 기법을 제안한다. 해당 기법을 통해 2, 3, 4비트 중 sensitivity에 따라 layer별로 비트를 자동으로 할당하고, 설정한 평균 비트 threshold를 넘어가지 않도록 설정해준다.

ShiftAddLLM을 OPT, Llama-1/2/3, Gemma, Mistral, BLOOM 5개의 LLM 모델에서 실험한 결과 기존 양자화 모델 대비 우수한 성능을 보이는 것을 확인하였다. 특히, 2/3-bit 설정에서 기존 최적 양자화 기법 대비 각각 평균 5.6 및 22.7의 perplexity 감소를 달성하였다. 또한, 기존 LLM 대비 메모리 사용량과 에너지 소비를 80% 이상 절감하면서도 지연 시간은 동일하거나 더 낮은 수준을 유지하여, 하드웨어 친화적인 양자화 기법을 GPU 상에서 실행해 효율적인 성능을 달성함을 입증하였다.

〔그림 4－105〕 고성능 Transformer 추론 엔진인 OPTIMUS의 전체 아키텍처



(a)

State	State0	State1	State2	State3			State4			State5		
Data Fetch	Bias, Position Layer norm, x	W^Q, W^K, W^V	W^O	W^{F1}			W^{F2}			W^Q, W^K, W^V (for next layer)		
Data Size		$3 \cdot d^2_{model}$	d^2_{model}	$2048 \cdot d_{model}$			$d_{model} \cdot 2048$			$3 \cdot d^2_{model}$		
Computation		x + Position = X	$[W^Q, W^K, W^V] \cdot X = Q, K, V$	Softmax $(K^T \cdot Q/\sqrt{d_k}) = S$	$V \cdot S = Z^{h-7}$	$W^O \cdot Z^{h-7} = Z$	X + Z = A0	Layer Norm (A0) = L0	ReLU $(W^{F1} \cdot L0 + B^{F1}) = F0$	$W^{F2} \cdot F0 + B^{F2} = F1$	A0 + F1 = A1	Layer Norm (A1) = L1
Number of Operations		$t \cdot d_{model}$	$3 \cdot d^2_{model} \cdot t$	$d_{model} \cdot t^2$	$d_{model} \cdot t^2$	$d^2_{model} \cdot t$	$t \cdot d_{model}$	$d_{model} \cdot t$	$2048 \cdot d_{model} \cdot t$	$2048 \cdot d_{model} \cdot t$	$t \cdot d_{model}$	$d_{model} \cdot t$

(b)

자료: Park, et al.(2020). p.7

다. 희소성 정보 추출 - 압축 형식 기반 연산기 공동 설계 연구 동향

1) OPTIMUS: Optimized Matrix Multiplication Structure for Transformer Neural Network Accelerator(Park, et al.(2020))(MLSys'20)

본 연구는 Transformer 추론 시 발생하는 메모리 병목과 연산 자원 비효율 문제를 해결하고자 한다. 특히 대규모 가중치 이동, 희소 행렬 연산의 부하 불균형, 디코더의 순차적

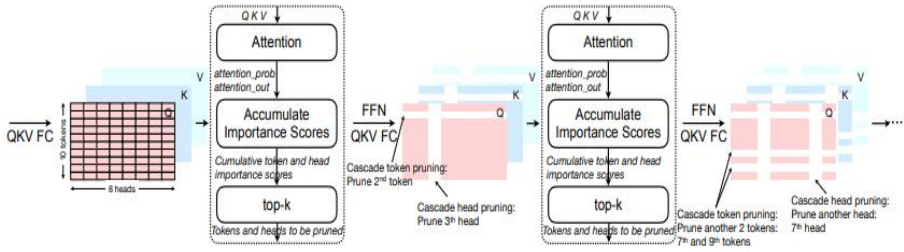
의존성으로 인한 병렬성 저하가 주요 한계로 지적되며, 이는 기존 GPU 기반 시스템에서 실시간 디코딩을 어렵게 만든다. 이를 해결하기 위해 제안된 OPTIMUS는 〔그림 4-105〕에 제시된 것처럼 밀집(초록색)과 희소(파란색) 행렬 곱셈 경로를 분리한 구조를 기반으로, 연산과 데이터 로드를 정밀하게 제어하는 파이프라인을 구성한다. 제어 흐름에 따라 더블 버퍼링을 활용하여 연산과 데이터 전송을 병렬 처리함으로써 DRAM 접근을 최소화하고, 단일 배치 추론에서도 모든 중간 데이터를 로컬 SRAM에 상주시킨다. 핵심 구성 요소로는 (1) 디코딩 단계에서 이전 타임 스텝의 K, V값을 재활용하여 불필요한 연산을 제거하는 redundant computation skipping, (2) 여러 PE간 입력 불균형을 완화하기 위해 데이터 접근 패턴을 재배치하는 Set-Associative RCSC(SA-RCSC) 희소 포맷, (3) 연산과 메모리 접근을 완전히 동기화해 병목을 줄이는 더블 버퍼링 기반 파이프라인 제어 로직이 있다. 또한 OPTIMUS는 1,024개의 MAC 유닛을 포함한 대규모 PE 어레이와 독립적인 입력/가중치 메모리 뱅크를 통해 높은 병렬성과 데이터 재사용률을 달성하며, 연산 State 1~5가 순차적으로 진행되면서 밀집·희소 연산이 구분되어 처리된다. 이러한 구조 덕분에 OPTIMUS는 디코더 단계의 연산 병목을 최소화하면서도, 다단계 연산을 타일 단위로 처리할 수 있다. WMT15 EN-DE 벤치마크를 사용한 실험 결과, OPTIMUS는 단일 배치 추론에서 Titan X GPU 대비 약 24배의 지연 시간 단축을 달성하였으며, 에너지 효율 또한 1,400배 이상 향상되었다.

2) Spatten: Efficient sparse attention architecture with cascade token and head pruning (Wang, et al.(2021))(HPCA'21)

본 연구는 대규모 LLM에서 발생하는 과도한 연산량과 메모리 접근 병목 문제를 해결하기 위한 알고리즘-하드웨어 공동 설계 기반의 가속기 구조를 제안한다. 먼저, cascade token pruning은 입력 문장에서 중요도가 낮은 토큰을 layer 단위로 점진적으로 제거하는 방식으로, attention 확률 기반의 누적 점수를 통해 토큰의 중요도를 평가한다. 중요도가 낮은 토큰은 이후 모든 layer에서 연산 대상에서 제외되어 불필요한 계산과 데이터 이동을 줄인다. 다음으로, cascade head pruning은 attention head별 출력 크기를 누적 비교하여 중요도가 낮은 head를 선택적으로 제거하는 기법이다. 이를 통해 feature 차원을 축소하고, 전체 네트워크 수준에서 연산 효율을 높여 attention뿐 아니라 후단의 FFN 단계까지 최적화한다. 마지막으로, progressive quantization은 softmax 확률 분포 형태에 따라 입력

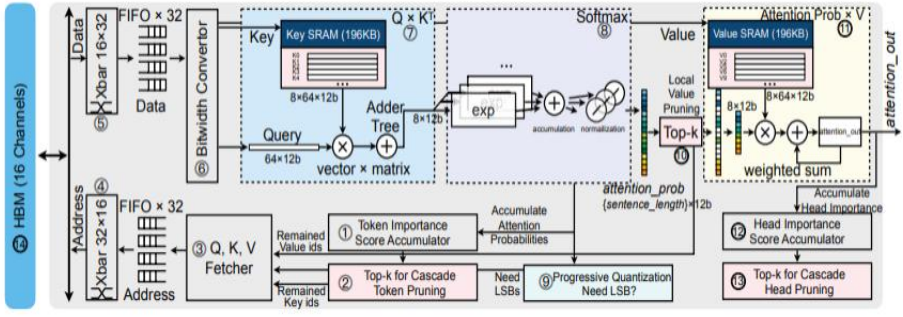
bitwidth를 동적으로 조정하는 방법이다. 분포가 특정 토큰에 집중되면 상위 bit만 사용하고, 평탄할 경우에만 하위 bit를 추가로 불러와 재계산을 수행함으로써 메모리 접근량을 줄이면서도 연산 정확도를 유지한다. 이 세 가지 기법은 하드웨어적으로 통합되어 동작하며, 실시간으로 토큰과 head 중요도를 선별하기 위한 O(n) 복잡도의 top-k engine이 함께 설계되었다. 또한 HBM 기반 병렬 접근 구조와 on-chip SRAM, crossbar 인터커넥트를 활용하여 데이터 이동 효율을 극대화하였다〔그림 4－106, 107〕. 이러한 과정은 decode 단계에서 토큰과 헤드 단위의 프루닝을 추론 단계에서 수행하여 동적으로 생략할 토큰/헤드를 제거함으로써 전반적인 가속기의 성능을 향상시킨다.

〔그림 4－106〕 Spatten의 제안하는 cascade token pruning, head pruning 기법



자료: Wang, et al.(2021). p.4

〔그림 4－107〕 Spatten 가속기의 전체 개요도



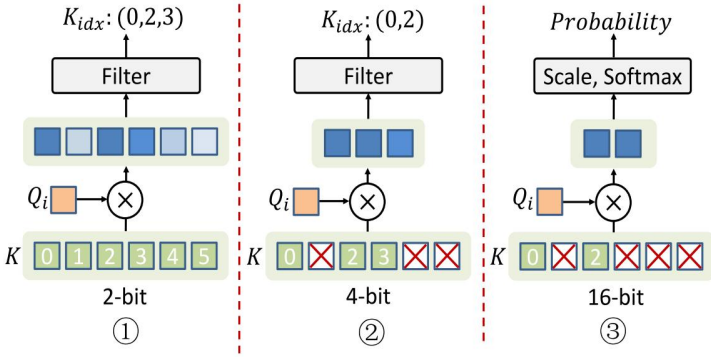
자료: Wang, et al.(2021), p.6

제안하는 기법을 통하여 Spatten은 정확도를 유지하면서 연산량과 외부 메모리 접근 횟수를 평균적으로 각각 2.1배, 10.0배 제거한다. 또한 TSMC 40nm 공정에서 합성된 Spatten 가속기는 NVIDIA TITAN Xp 대비 처리량과 에너지 효율 성능에서 각각 162배, 1,193배 우수한 성능을 보인다.

3) Energon: Towards Efficient Acceleration of Transformers Using Dynamic Sparse Attention(Zhou, et al.(2022))(TCAD'22)

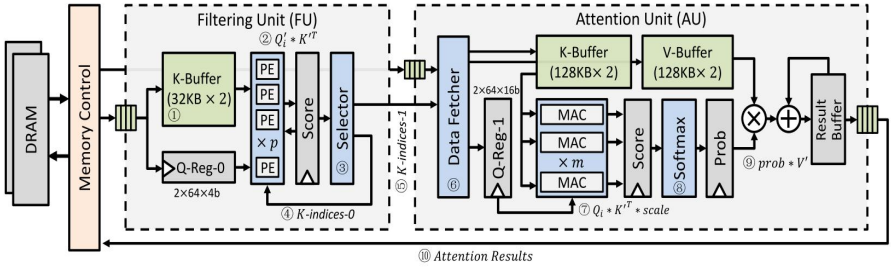
본 연구는 Transformer의 attention 연산이 가지는 복잡한 데이터 이동과 연산 복잡도로 인해 실시간 추론이 어려운 문제를 해결하기 위해, 동적 희소성 기반 알고리즘-하드웨어 통합 가속기를 제안한다. 기존 연구인 Spatten은 attention 희소성을 이용했지만, coarse-grained pruning 또는 모든 QKV 데이터를 DRAM에서 불러오는 구조로 인해 오버헤드가 컸다. Energon은 〔그림 4－108〕에 제시된 것처럼 attention에서 실제로 높은 중요도를 갖는 query-key 쌍만을 실시간으로 식별해 고정밀 연산에 사용하고, 나머지는 저정밀 필터링 단계에서 제거한다. 구체적으로 라운드 1에서 2-bit 텐서로 1차 필터링을 수행하고, 라운드 2에서 4-bit 텐서로 통과 후보를 추가로 거른 뒤, 라운드 3에서 남은 key만을 사용해 고정밀 sparse attention을 수행한다. Energon의 Filtering Unit(FU)은 MP-MRF를 수행해 중요 key 인덱스를 찾아내며, Attention Unit(AU)은 선택된 key와 value만을 이용해 고정밀 sparse attention을 계산한다. 각 연산은 ping-pong buffer 구조로 파이프라이닝 되어 메모리 접근을 숨기며, result-reusable multi-precision PE를 통해 두 필터링 단계를 동일한 하드웨어 자원 위에서 효율적으로 처리한다〔그림 4－109〕. 실험 결과, SQuAD, Wikitext-2와 ImageNet, CIFAR-100 모두에서 최대 11.5×의 pruning ratio를 달성하면서 정확도 손실은 0.5% 이하였다. 특히 NVIDIA V100 GPU 대비 8.7× 처리량, 103× 에너지 절감, Spatten 대비 1.7× 처리량 향상, 1.6× 에너지 효율 개선을 보였다.

〔그림 4－108〕 Energon의 MP-MRF 전략의 예시



자료: Zhou, et al.(2022), p.4

〔그림 4－109〕 Energon 가속기 전체 개략도



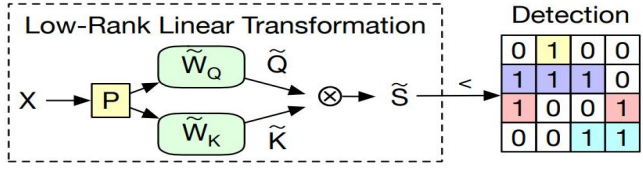
자료: Zhou, et al.(2022), p.5

4) DOTA: Detect and Omit Weak Attentions for Scalable Transformer Acceleration(Qu, et al.(2022))(ASPLOS'22)

본 연구는 긴 시퀀스를 처리할 때 self-attention의 계산 복잡도와 메모리 소모가 지수적으로 증가하여 막대한 연산 비용이 발생한다는 점을 문제로 삼는다. DOTA는 attention 그래프의 모든 연결이 동일하게 중요한 것은 아니라는 통찰에 기반하여, 런타임 동안 약한 attention 연결을 정확하게 탐지하고 생략할 수 있도록 경량 Detector와 Transformer 본체를 공동 학습 구조로 설계한다. 구체적으로, 〔그림 4－110〕에서처럼 Detector는 입력 query/ key 행렬을 저차원 공간으로 투영한 뒤, 연산량이 훨씬 적은 상태에서 약한 attention

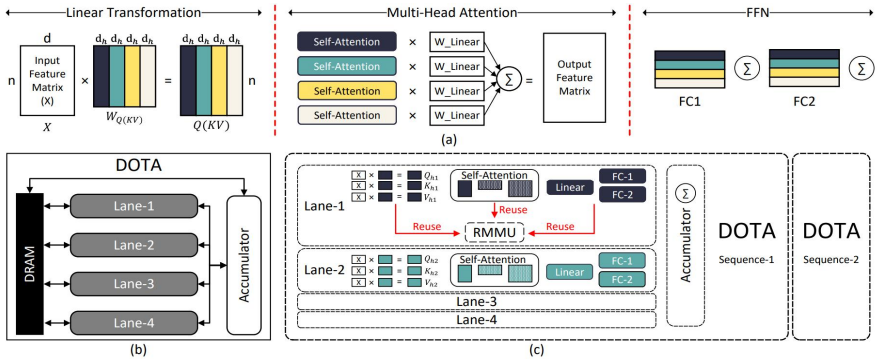
연결을 탐지하고, 이 결과로 생성된 binary mask를 기반으로 연산을 수행한다. Detector는 Transformer와 공동 학습되어 감지 정확도와 모델 품질을 동시에 보장하며, 감지 연산은 INT2~INT4 수준의 저정밀도 근사 곱셈으로 수행되어 오버헤드를 최소화한다. DOTA의 추론 과정은 세 개의 연속된 stage로 분해되며, 각 stage의 GEMM 연산은 여러 chunk로 나뉘어 서로 다른 compute lane에 매핑된다. 각 lane은 DRAM으로부터 입력 activation을 스트리밍하고, 중간 결과는 accumulator에 집계된다. 핵심 연산 경로는 〔그림 4－111〕에 나타난 것처럼 Reconfigurable Matrix Multiplication Unit(RMMU)이 담당하며, multi-precision 연산을 지원하여 고정밀 계산을 효율적으로 수행한다. 또한 token-parallel dataflow와 locality- aware scheduler를 적용하여 여러 query의 key/value 재사용을 극대화하고, 각 DOTA accelerator가 입력 시퀀스 하나를 전담하여 lane 단위로 chunk를 병렬 처리함으로써 메모리 접근 횟수를 크게 줄이면서도 긴 시퀀스를 높은 처리량으로 소화한다. 실험 결과, QA, LRA-Image, Text, Retrieval, GPT-2 LM 등 다양한 벤치마크에서 전체 attention score의 3-10%만 유지해도 dense 모델과 동등하거나 오히려 더 높은 정확도를 달성하였다. 또한 성능 측면에서는 GPU 대비 최대 152.6배의 attention 연산 가속과 5,185배에 달하는 에너지 효율 개선을 보였다.

〔그림 4－110〕 DOTA detector의 bitmask 생성과정



자료: Qu, et al.(2022), p.4

〔그림 4－111〕 DOTA의 가속기 시스템 개략도

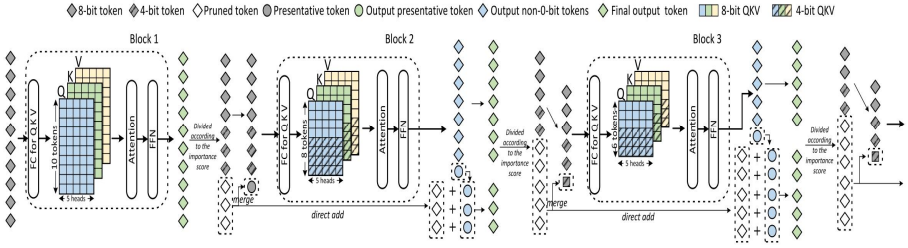


자료: Qu, et al.(2022), p.5

5) DTATrans: Leveraging Dynamic Token-Based Quantization with Accuracy Compensation Mechanism for Efficient Transformer Architecture(Yang, et al.(2022)) (TCAD'22)

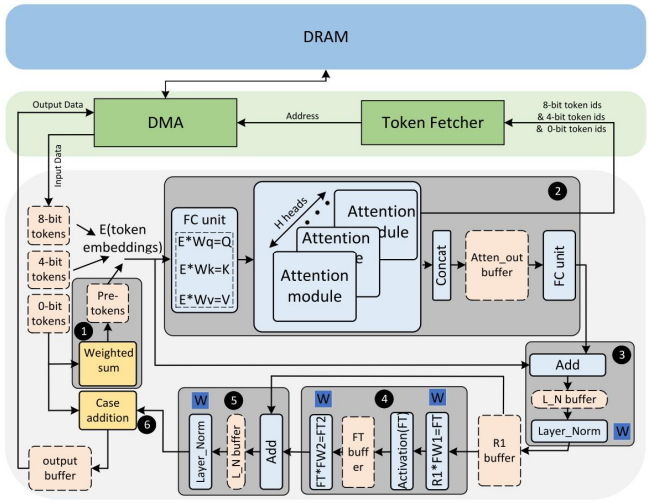
DTATrans는 Transformer 모델의 과도한 연산량과 메모리 소모 문제를 해결하기 위해 제안된 알고리즘-하드웨어 공동 설계 기반 가속기 구조이다. 본 연구는 토큰별 중요도 차이에 주목하여, 모든 입력 토큰을 동일한 bitwidth로 양자화하던 기존 방식의 한계를 개선한다. DTATrans는 토큰의 중요도에 따라 비트 정밀도를 동적으로 조정하는 동적 토큰 기반 양자화 기법과 정확도 손실을 최소화하는 경량 보상 메커니즘을 함께 적용한다. 〔그림 4－112〕에 제시된 압축 절차는 세 단계로 이루어진다. 우선 attention score를 기반으로 각

〔그림 4－112〕 토큰 중요도 기반 양자화와 대표 토큰 병합 과정



자료: Yang, et al.(2023), p.5

〔그림 4－113〕 DTAtrans의 가속기 전체 개략도

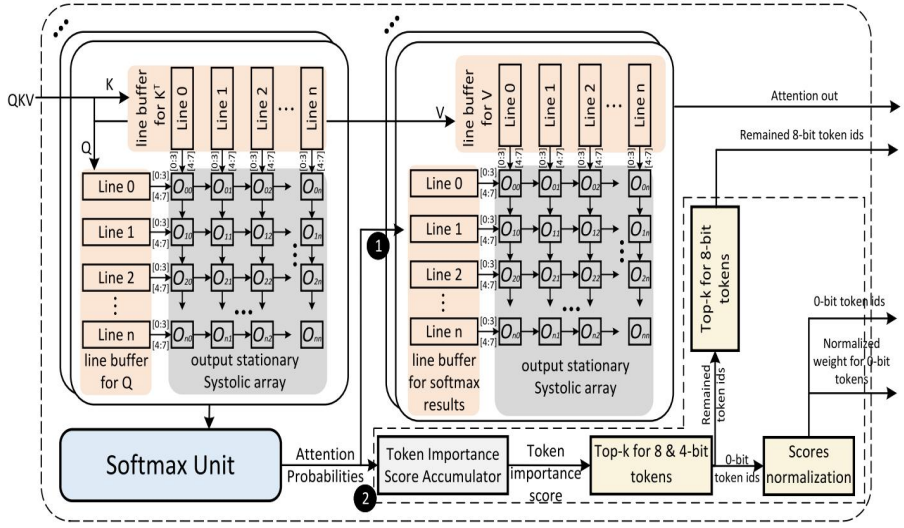


자료: Yang, et al.(2022), p.6

토큰의 중요도를 계산하고, 이를 바탕으로 토큰을 8-bit(중요), 4-bit(보통), 0-bit(비중요)로 구분한다. 이후 0-bit로 양자화된 토큰은 단순히 제거하지 않고, 중요도 가중치를 반영한 대표 토큰으로 병합해 연산에 포함한다. 생성된 대표 토큰은 이후 Transformer 블록 내 연산에 참여해 정보 손실을 최소화한다. 〔그림 4－113〕은 전체 하드웨어 구조를 나타낸다. 입력 토큰은 중요도에 따라 재배열되어 multi-head attention 모듈로 전달되며, 연산 후에는 feed-forward network와 두 단계의 residual 블록을 거쳐 출력된다. 전단의 병합 부분에서는 0-bit 토큰을 대표 토큰으로 결합하고, 후단에서는 다시 분배하여 원래의 토큰 집합으로 복원한다. 이러한 병합·복원 구조는 정확도를 유지하면서도 연산량을 대폭 줄이는 핵심 역할을 한다. 중심 구성 요소인 〔그림 4－114〕의 attention 모듈은 4-bit와 8-bit 연산을 혼합 처리할 수 있는 가변 속도 행렬 연산 배열(VSSA, Variable-Speed Systolic Array)으로 설계되었다. 각 연산 유닛은 bit-witdh에 따라 연산 속도를 조정하며, 데이터 흐름의 동기화 문제로 인한 병목을 방지하기 위해 토큰 재배치 및 정밀도 기반 클러스터링 기법이 함께 적용된다. 이 방식은 파이프라인 정체 비율을 약 39%에서 15% 이하로 줄여 하드웨어 활용도를 크게 개선했다. 실험 결과에 따르면, DTATrans는 평균적으로 모델 연산량

을 약 1.6배 줄이고, Spatten 대비 4.2배 높은 에너지 효율을 달성하였다. BERT-Base와 GPT-2 모델에서 평균 40% 이상의 0-bit 토큰 비율을 확보하면서도 정확도 손실은 1% 미만으로 유지되었다. 또한 28nm CMOS 공정 기반 칩 구현에서 1.49mm² 의 면적과 162mW 수준의 전력으로 동작하며, 1,622 GOPS/W의 에너지 효율을 기록하였다.

〔그림 4－114〕 DTATrans의 attention module의 개요도

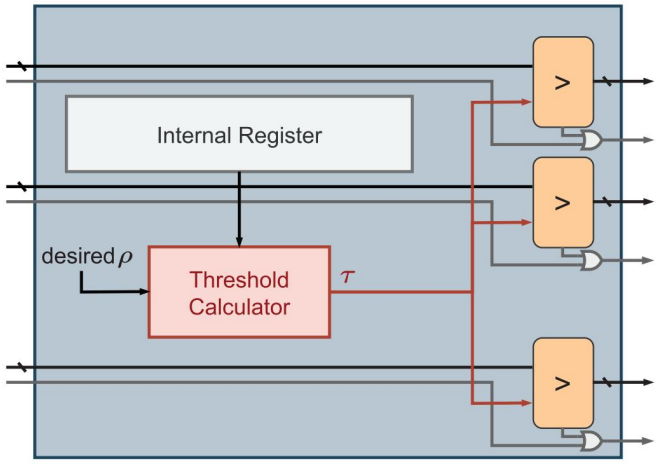


자료: Yang, et al.(2022), p.7

6) AccelTran: A Sparsity-Aware Accelerator for Dynamic Inference with Transformers (Tuli, et al.(2023))(TCAD'23)

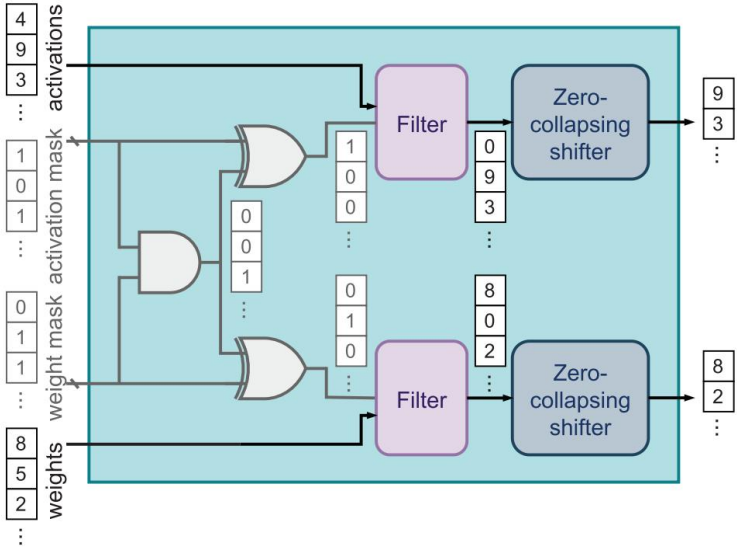
기존 Spatten, Energon, OPTIMUS 등의 연구는 attention 행렬을 pruning하는 방식으로 sparsity를 활용하지만, 복잡한 top-k 연산으로 인해 오히려 연산 오버헤드가 커지는 문제가 있었다. 이를 해결하기 위해 AccelTran은 runtime 중 활성화(activation) 값을 기반으로 비효율적 연산을 실시간 제거하여 전체 Transformer 추론 효율을 극대화하는 DynaTran 구조를 제안한다. 〔그림 4－115, 116〕에서는 threshold 기반 sparsity 검출과 bitmask 생성 과정을 나타낸다. 내부의 Threshold Calculator는 목표 sparsity에 맞게 활성화 및 가중치 행렬의 임계값(τ)을 결정하며, 각 요소를 비교하여 bitmask를 생성한다. 이후 bitmask를 기반

〔그림 4－115〕 AccelTran의 threshold 기반 희소화 모듈



자료: Tuli, et al.(2023), p.5

〔그림 4－116〕 AccelTran의 bitmask 기반 연산 filtering

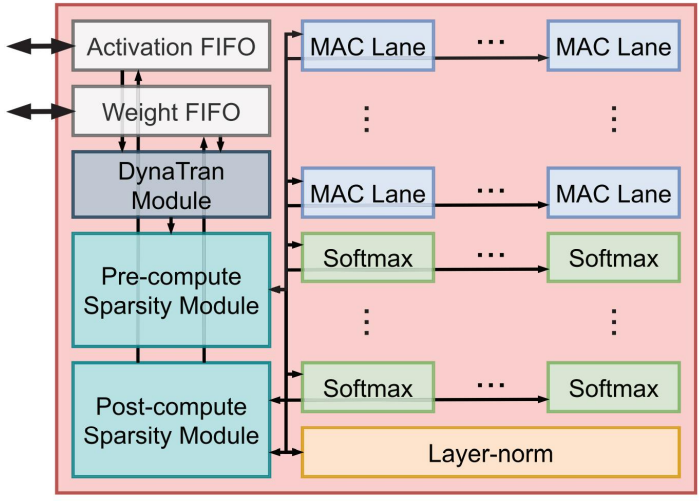


자료: Tuli, et al.(2023), p.6

으로 0이 아닌 데이터만 filtering하여 연산을 수행함으로써, 메모리 접근과 MAC 연산을 동시에 최소화한다. 〔그림 4－117〕은 AccelTran의 전체 시스템 아키텍처를 나타낸다. 내부 DynaTran Module은 activation과 weight FIFO에서 데이터를 받아 runtime pruning을 수행하고, pre-/post-compute sparsity module을 통해 zero element를 제거 및 복원한다. 이후 각 MAC Lane, Softmax, Layer-norm 모듈이 병렬로 동작하여 Transformer의 attention-FFN 연산을 순차적으로 수행한다. 특히 monolithic 3D RRAM을 이용한 고대역폭 메모리 통합으로 데이터 전송 병목을 줄였으며, staggered scheduling을 통해 MAC과 softmax가 병렬로 실행된다. 이로써 BERT-Base 기준 1.2×의 sparsity 향상과 96%의 정확도 유지를 달성하였다.

AccelTran-Edge는 저전력 LP-DDR3 메모리와 64개의 PE를 탑재한 경량형 아키텍처로, Raspberry Pi 대비 약 330,000배 높은 throughput과 93,000배 낮은 에너지 소모를 보였다. AccelTran-Server는 monolithic 3D RRAM과 512개의 PE, 32개의 Softmax 엔진을 갖춘 고성능 버전으로, NVIDIA A100 GPU 대비 5.7× 높은 처리량과 3.7× 낮은 에너지 소모를 보였다.

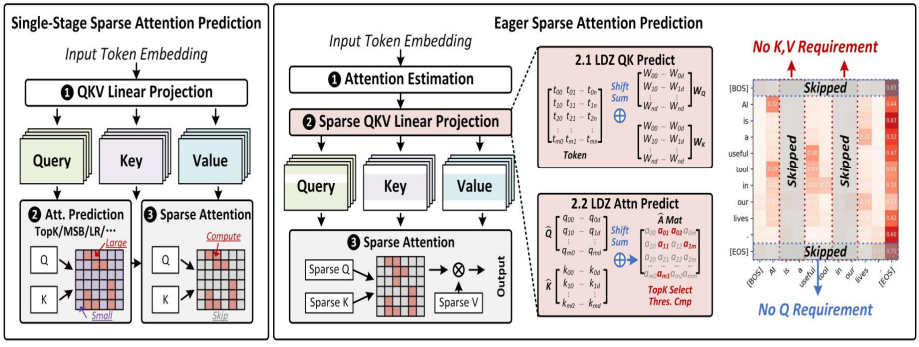
〔그림 4－117〕 AccelTran의 가속기 시스템 개략도



자료: Tuli, et al.(2023), p.5

7) FACT: FFN-Attention Co-optimized Transformer Architecture with Eager Correlation Prediction(Qin, et al.(2023))(ISCA'23)

〔그림 4－118〕 EEager prediction 수행 절차와 single-stage prediction 방식 비교

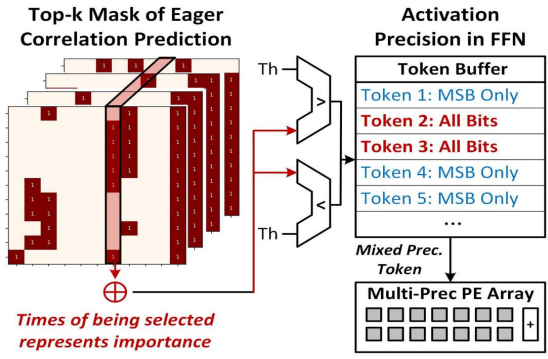


자료: Qin, et al.(2023), p.4

기존의 Transformer 가속기들은 attention 연산의 sparsity만을 활용하여 효율을 개선했지만, 실제로는 QKV 생성과 FFN이 전체 전력의 70% 이상을 차지한다는 점을 간과하였다. FACT는 이러한 병목을 해결하기 위해 attention 발생 이전 단계에서 correlation을 예측하는 Eager Prediction(EP) 알고리즘과 이를 지원하는 전용 하드웨어를 제안한다. 〔그림 4－118〕은 기존 단일 단계 예측과 FACT의 cross-stage eager prediction 구조를 비교한다. 기존 방식은 Q, K 생성 이후 attention을 예측하지만, FACT는 입력 토큰과 projection weight만으로 attention 행렬의 중요도를 사전에 예측하여 불필요한 QKV 연산을 건너뛴다. EP는 log-domain에서의 leading-one detection(LOD) 기반 add-only 연산을 사용해 곱셈기를 완전히 제거하므로, 매우 낮은 전력으로 상관도를 추정할 수 있다. 이렇게 생성된 attention mask는 〔그림 4－119〕에서처럼 Q, K, V sparsity와 token-wise mixed-precision FFN으로 확장된다. 중요도가 낮은 토큰은 FFN 연산을 INT4 정밀도로 수행하고, 중요한 토큰만을 INT8로 유지하여 전력 소모를 줄인다. 또한 FACT는 예측 결과를 기반으로 out-of-order QKV scheduling을 적용하여 EP 과정 동안에도 PE array가 idle 상태에 머무르지 않도록 했다. 마지막으로 diagonal SRAM storage pattern을 도입해 MSB와 LSB를 분리·대각선 형태

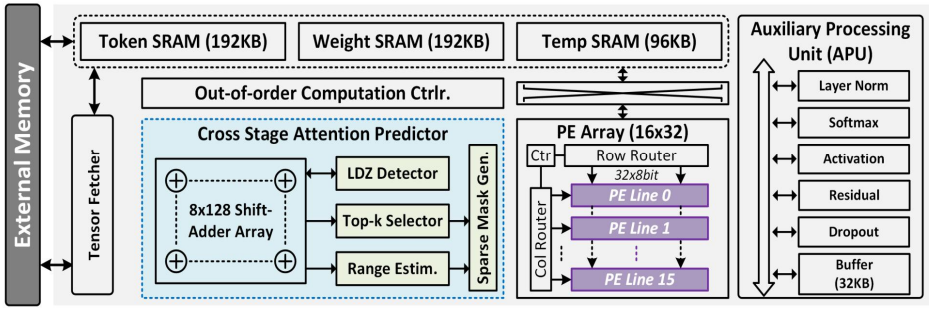
로 배치함으로써, INT4/8 혼합 정밀도 접근 시 메모리 대역폭 낭비를 완전히 제거하였다. 하드웨어 측면에서 〔그림 4－120〕에 제시된 FACT 가속기는 EP 유닛, out-of-order scheduler, PE array, APU(softmax·LayerNorm 모듈) 등으로 구성된다. 실험 결과, FACT는 BERT 및 ViT 계열 22개 벤치마크에서 평균 3.6배의 처리량 향상, 최대 100배의 에너지 효율 개선을 달성했다. 특히 attention 연산에서 90% 이상의 연산량을 제거하였고, FFN 단계에서는 token별 중요도에 따른 정밀도 축소를 통해 5.65×의 에너지 절감을 이끌었다. Nvidia V100 GPU 대비 94.98× 높은 에너지 효율을 달성하였으며, 기존 ELSA 대비 3.9×, Sanger 대비 22.8× 우수한 성능을 달성했다.

〔그림 4－119〕 FFN에서의 token별 혼합 정밀도 사용 방법



자료: Qin, et al.(2023), p.6

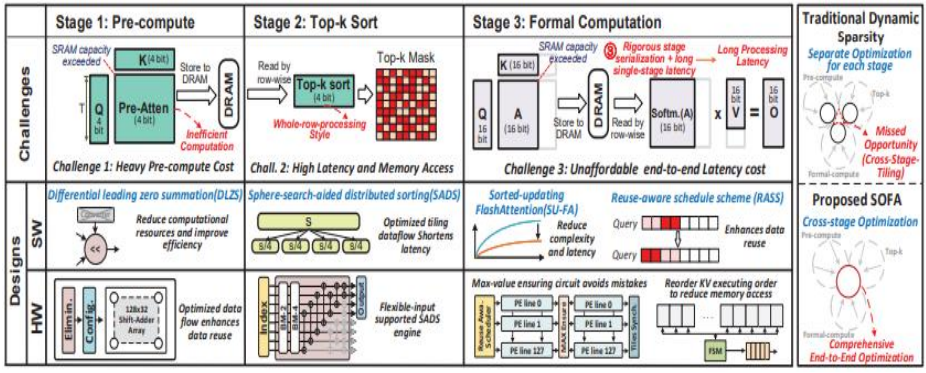
〔그림 4－120〕 FACT의 가속기 시스템 개략도



자료: Qin, et al.(2023), p.7

8) SOFA: A compute-memory optimized sparsity accelerator via cross-stage coordinated tiling(Wang, et al.(2024))(MICRO'24)

〔그림 4－121〕 동적 희소성을 위한 SOFA의 하드웨어-소프트웨어 공동 설계 개요도



자료: Wang, et al.(2024), p.2

해당 연구는 기존의 동적 희소성 특화 가속기 연구들은 LLM의 prefill, decode 단계의 연산을 분할하여 처리하지 못한다는 점을 근거로, large-scale token processing parallelism (LTPP) 상황에서 불가피한 지연 시간이 발생하는 문제를 지적한다. 이를 해결하기 위하여, Sparse Optimization Framework Architecture(SOFA)를 제안하여, LTPP 시나리오에서 메모리 접근 횟수를 줄임과 동시에 모델 전체 추론의 지연 시간을 줄인다. SOFA는 attention 희소성 예측 과정에서 발생하는 오버헤드를 줄이기 위하여 log-domain multiplication-free 전략인 differential leading zero summation을 제안한다. 또한 동적 희소성의 처리에서 발생하는 정렬 복잡도를 줄이기 위하여 attention 행렬의 행별로 하위 영역을 나누어 정렬하고, 정렬된 하위 영역을 기반으로 KV의 중요도를 판별하는 sphere-search aided distributed sorting(SADS) 기법과, 이를 위한 처리 모듈을 제안한다〔그림 4－121〕. SOFA는 제안하는 기법들을 tile 단위로 동작함과 동시에 cross-stage 연산이 가능하게 하므로, LTPP 상황에서 기존 동적 희소성 특화 가속기보다 유연하고 효율적으로 연산을 처리한다. SOFA는 토큰 희소성의 연산량 감소 효과를 확인하기 위해 정확도 저하율에 따른 연산 감소율을 측정하였다. LLaMA 모델에서 SQuAD, RTE, wikitext2, QNLI 등 다양한 task에 실험

을 진행한 결과, 0%, 1%, 2%의 정확도 저하율에 대해 평균적으로 56.8%, 62.6%, 67.4%의 QKV, Attention 연산량 감소시켰다. 또한 SOFA는 LLaMA-7B에 대해 Spatten(Wang, et al.(2021))보다 8.5배 우수한 지연 시간 감소율을 달성하고 7,183 GOPS/W 에너지 효율과 4,292 GOPS/mm2 의 면적 대비 처리량 효율 성능을 달성한다.

〈표 4－6〉 LLM 가속 솔루션 연구별 성능 비교

연구	기술 보유국 / 보유 기관	처리량	전력 효율
EdgeLLM (Huang, et al. 2025)	중국 / Shenzhen Univ.	85.8 Token/s	1.51 Token/s/W
LPU (Moon, et al. 2024)	한국 /HyperAccel	769 Token/s	2707.7 Token/s/W
DFX (Hong, et al. 2022)	한국/ KAIST	93.1 Token/s	1.61 Token/s/W
TransPIM (Zhou, et al. 2022)	미국 / Univ. of California	734 GOPS	17.48 GOPS/W
PAISE (Lee, et al. 2025)	한국 / Samsung SDS	103.1 Token/ms	103.1 Token/s/W
IANUS (Seo, et al. 2024)	한국 / Seoul National Univ.	263.1 Token/s	6.69 Token/s/W
Tender (Lee, et al. 2024)	한국 / Seoul National Univ.	53.3 Token/s	33.3 Token/s/W
TerEffic (Yin, et al. 2025)	싱가폴 / National Univ. of Singapore	727 Token/s	15.74 Token/s/W
MECLA (Qin, et al. 2024)	중국 / Tsinghua Univ.	14.0 TOPS	7.08 TOPS/W
OPTIMUS (Park, et al. 2020)	한국 /Hanyang Univ.	500.05 GOPS	110 sentence/J
Spatten (Wang, et al. 2021)	미국 /Cambridge Univ.	360 GOPS	382 GOPS/W
Energon (Zhou, et al. 2022)	중국 /Pekign Univ.	612 GOPS	225 GOPS/W
DTATrans (Yang, et al. 2022)	중국 /Shanghi Univ	1304.2 GOPS	1622.5 GOPS/J
DOTA (Qu, et al. 2022)	미국 /UC Santa Barbara Univ.	12 TOPS	-

연구	기술 보유국 / 보유 기관	처리량	전력 효율
AccelTran (Tuli, et al. 2023)	미국 / Princeton Univ.	15.05(Edge) TOPS	2.219 TOPS/W
FACT (Qin, et al. 2023)	중국 /Tsinghua Univ.	928 GOPS	4.388 TOPS/W
SOFA (Wang, et al. 2024)	중국 / Tsinghua Univ.	24.4 TOPS	7.18 TOPS/W

라. 하드웨어 친화적 근사화 알고리즘 및 연산기 연구 동향

1) PLAC: Piecewise Linear Approximation Computation for All Nonlinear Unary Functions (Dong, et al.(2020))(VLSI'20)

해당 연구는 기존의 error-flattened 기반 PWL 근사화 방법들이 특정 함수 특성에 종속되고, 하드웨어 구현 시 maximum absolute error(MAE)를 제어할 수 없어 설계 반복과 자원 낭비가 발생한다는 점을 지적한다. 이를 해결하기 위해 piecewise linear approximation computation(PLAC) 프레임워크를 제안한다. 구체적으로 PLAC 프레임워크는 소프트웨어 단계에서 정의된 MAE 기준을 만족하면서 최소 segment 수를 탐색하는 고성능 segmenter와 실제 회로 동작을 완전히 모사하여 MAE 조건 하에서 필요한 coefficient 및 output bit-width를 자동 결정하는 quantizer로 구성된다. Segmenter은 piecewise linear approximation 적용 시 인접 segment간의 경계점을 중복 없이 갱신하고 bisection-seeking method를 도입해 각 구간의 입력 폭을 최대화함으로써, 로그 함수에 대해 이론적 한계에 수렴하는 segmentation fidelity를 구현한다. Quantizer은 quantization factor(QF) 를 조정하여 소프트웨어 MAE를 하드웨어 허용 오차 범위에 맞게 스케일링하며, 최소 fractional bit width를 산출해 불필요한 리소스 낭비를 제거한다. PLAC 프레임워크의 하드웨어 아키텍처는 기존 연구에서 주로 활용한 multi-stage indexing logic을 단일-stage multiplexer 기반의 unified indexing 구조로 재설계한다. 또한, parallel comparator에서 생성된 sign bit vector를 직접 addressing key로 활용함으로써 인덱싱 경로의 복잡도를 최소화하여 area와 latency를 동시에 줄인다. 구현 결과, TSMC 40/65/90nm 공정 기반 ASIC 합성 시 PLAC 프레임워크는 log 함수에서 기존 PWL 방식 대비 면적 2.8%, 전력 3.8% 감소와 함께 동일한 지연 시간 내에서 더 낮은 오차를 달성하였으며, tanh, sigmoid, softsign 함수에서도 각각

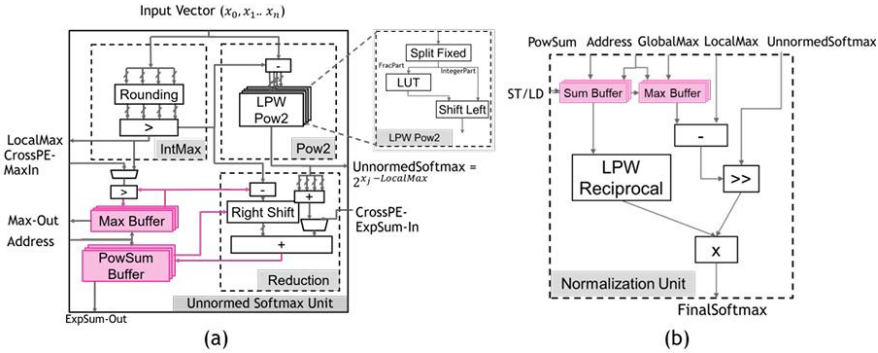
6.3%, 16.5%, 17.3%의 면적 절감과 4~12% 수준의 전력 및 지연 개선을 보였다. 결과적으로 PLAC은 다양한 비선형 함수에 적용 가능한 범용 근사화 프레임워크로써 기존 PWL 대비 적은 세그먼트 수로 동일한 정확도를 유지하면서도 회로 수준의 MAE 제어와 설계 반복 최소화를 동시에 달성한다.

2) Softermax: Hardware/Software Co-Design of an Efficient Softmax for Transformers (Stevens, et al.(2021))(DAC'21)

해당 연구는 Transformer 기반 네트워크에서 Softmax 연산이 전체 실행 시간의 큰 비중을 차지하는 문제점을 해결하기 위하여, 하드웨어 친화적 Softmax 근사화 프레임워크인 Softermax를 제안한다. Softermax는 base replacement, low-precision Softmax computation, online normalization의 세 가지 기법을 통해 Softmax의 연산 복잡도와 하드웨어 오버헤드를 동시에 줄인다. Softermax는 먼저 exponentiation을 power-of-two 형태로 단순화함으로써 LUT 크기와 연산 오버헤드를 줄인다. 또한 exponential, accumulation, division 연산을 모두 low-precision fixed-point 연산으로 대체하고 softermax-aware fine-tuning을 적용하여 정확도 손실을 최소화한다. 마지막으로 numerically-stable Softmax에서 필요한 max 탐색 과정을 제거하기 위해 online normalization 방식을 도입하고, base-2 특성을 이용해 renormalization 연산을 단순한 shift로 처리해 하드웨어 효율을 극대화한다.

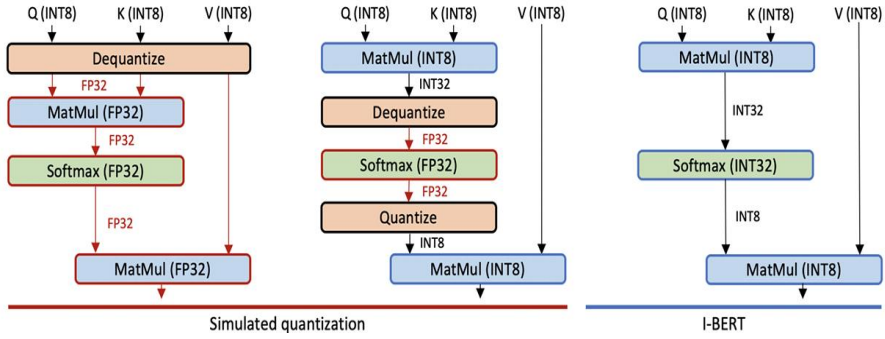
하드웨어 구조는 unnormed Softmax unit과 normalization unit으로 구성된다 〔그림 4-122〕. 전자는 local maximum 계산, 2의 거듭제곱 연산, 누적 합산을 수행하고, 후자는 denominator renormalization와 division을 담당한다. Softermax는 이러한 구조를 MAGNet 등 기존 딥러닝 가속기의 post-processing stage에 쉽게 통합할 수 있어, tensor core 기반의 TPU 아키텍처와의 호환성이 높다. BERT-Base/-Large 실험 결과, Softermax는 8-bit quantized baseline 대비 평균 정확도 손실이 0.5% 미만이었으며, 일부 task에서는 오히려 정확도가 향상되었다. TSMC 7nm 공정 기반 합성 결과, 기존 Softmax 대비 logic area 10% 감소 및 에너지 효율이 2.35배 향상되는 효과를 보였다. 결과적으로 Softermax는 Transformer 가속기에서 Softmax 연산 병목을 근본적으로 완화하는 하드웨어 친화적 설계로 기존 full-precision Softmax 대비 동등 이상의 정확도를 유지하면서 latency, area, power efficiency 측면에서 대폭 개선된 효율을 달성한다.

〔그림 4－122〕 Softermax 연산기 개요도



자료: Stevens, et al.(2021), p.4

〔그림 4－123〕 기존 양자화 기법과 I-BERT의 정수 도메인 연산 플로우 비교



자료: Kim, et al.(2021), p.3

3) I-BERT: Integer-only BERT Quantization(Kim, et al.(2021))(ICML'21)

해당 연구는 기존 Transformer 양자화 방법들이 floating-point 연산을 포함한 형태로 수행되어, 실제 정수 연산 기반 하드웨어 배포 시 연산 효율을 충분히 확보하지 못한다는 문제점을 지적한다. 본 연구에서는 이를 해결하기 위해 BERT 모델의 Softmax, LayerNorm, GELU 등 비선형 함수를 정수 도메인으로 근사화하여 효율적인 연산을 수행할 수 있는 I-BERT 프레임워크를 제안한다〔그림 4－123〕. I-BERT는 floating-point 연산을 대체하기 위해, 각 비선형 연산에 대해 integer polynomial 및 piecewise linear approximation 기반의

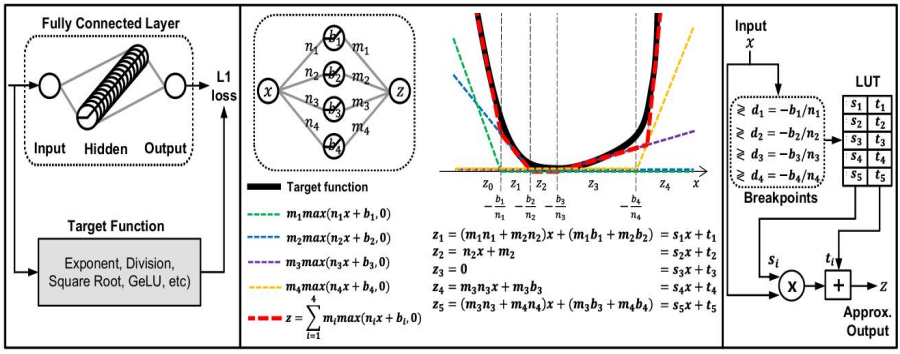
근사 함수를 설계한다. 구체적으로, Softmax는 입력값의 차이를 이용한 scaling 및 clipping을 통해 overflow를 방지하여 정수 도메인에서 정밀도를 유지한다. GELU 근사화의 경우, fixed-point polynomial 근사화 방식을 적용해 기존 floating-point 대비 우수한 연산 효율성을 달성한다. 또한 LayerNorm 함수에서는 reciprocal square root 함수를 integer arithmetic으로 근사하고, quantization range를 layer별로 정규화하여 overflow를 방지한다.

모델 학습 과정에서는 기존의 quantization-aware training 방법을 변형하여 integer arithmetic consistency를 보장하는 gradient calibration 단계를 추가한다. 이를 통해 학습 중에도 정수 연산으로 근사화된 함수들이 floating-point 기준 학습과 일관된 방향으로 수렴하도록 보정한다. GLUE 및 SQuAD benchmark를 기반으로 평가한 결과, I-BERT는 기존 8-bit quantization 모델 대비 평균 정확도 손실이 1% 미만이며, 실제 CPU 및 FPGA 상에서 4.3×의 throughput 향상과 2.3×의 latency 감소를 달성하였다. 또한 모든 연산이 integer domain에서 수행되므로, 추가적인 floating-point unit 없이도 on-device inference가 가능하며, 모델 사이즈와 전력 소모를 대폭 절감할 수 있음을 보였다. 결과적으로 I-BERT는 Transformer 구조를 완전한 정수 연산 체계로 재구성한 최초의 프레임워크로, 하드웨어 친화적 정밀도 제어를 통해 정확도 손실 없이 고효율 inference를 가능하게 하며, integer-based Transformer 가속기 연구들의 기반이 된 핵심적인 접근법으로 평가된다.

4) NN-LUT: Neural approximation of non-linear operations for efficient transformer inference(Yu, et al.(2022))(DAC'22)

해당 연구는 기존 근사화 기법들이 연산별 맞춤형 설계에 의존함으로써 하드웨어 설계 복잡도가 증가하고, 확장성 및 일반화 성능이 제한되는 문제를 해결하기 위해 범용적인 비선형 함수 근사화 프레임워크인 neural network-generated look-up table(NN-LUT) 를 제안한다. NN-LUT은 비선형 연산을 근사하기 위해 소형 신경망 네트워크를 학습한 뒤, 이를 동일한 근사 정밀도를 가지는 LUT 형태로 변환함으로써, 다양한 비선형 함수를 단일 하드웨어 구조에서 효율적으로 처리할 수 있다〔그림 4-24〕. 기존 LUT 기반 근사법이 사전에 정의된 분할 구간을 고정적으로 사용해 approximation fidelity가 제한되지만 NN-LUT은 학습된 신경만 네트워크의 breakpoints를 LUT의 구간 경계로 변환하여, 입력 범위 내에서 동적으로 최적화된 세분화를 수행한다. 또한 NN-LUT은 approximation 정확도를 높이기 위해 세 가지 training 방법론을 제안한다. 첫째로, 다양한 비선형 함수의 특성에 따른

〔그림 4－124〕 NN-LUT 프레임워크 개요

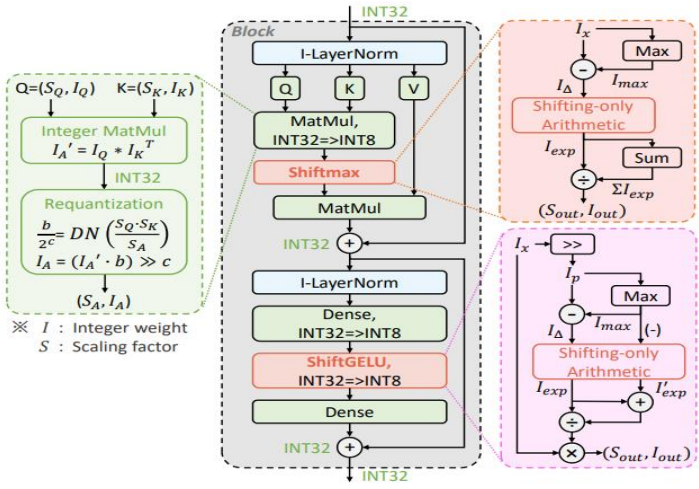


자료: Yu, et al.(2022), p.3

입력 범위와 초기화 방식을 차별화함으로써 신경망 네트워크 학습의 안정적인 수렴을 보장한다. 둘째로 input scaling method를 도입하여 LayerNorm의 reciprocal square root 연산처럼 넓은 동적 범위를 갖는 함수에 대해 안정적인 근사를 가능하게 한다. 셋째, 별도의 데이터셋이나 추가 학습 없이도 수행 가능한 dataset-free lightweight calibration 기법을 통해, fine-tuning된 Transformer의 각 레이어에서 발생하는 동적 범위 변화를 신속히 보정할 수 있도록 하였다. RoBERTa 및 MobileBERT 모델을 대상으로 한 실험에서, NN-LUT은 기존 Linear-LUT 대비 동일한 입력 수(16 entries) 조건에서 FP32 수준의 정확도를 유지하며, GLUE와 SQuAD task에서의 정확도 손실은 거의 없었다. 특히 LayerNorm 근사화에서 발생하는 오차는 calibration 단계를 통해 완전히 상쇄되었고, INT32 정밀도 환경에서도 I-BERT 대비 동등 혹은 더 높은 정확도를 달성하였다. 추가로, 하드웨어 합성 결과, NN-LUT은 기존 I-BERT 기반 연산 유닛 대비 면적 2.6배 감소, 전력 36배 감소, 지연 3.9배 단축을 달성하였다. 또한 Transformer 전용 NPU에 통합 시, 비선형 연산에 소요되는 사이클 비중을 최대 43% 감소시켜, 전체 시스템 실행 시간에서 최대 26%의 속도 향상을 보였다. 결과적으로 NN-LUT은 비선형 연산을 범용적이고 경량화된 LUT 구조로 통합한 첫 번째 neural-approximation 기반 프레임워크로, 기존 연산별 전용 근사화 방식을 대체할 수 있는 통합적 비선형 연산 가속 솔루션으로 평가된다.

5) I-vit: Integer-only quantization for efficient vision transformer inference(Li, et al. (2023))(ICCV'23)

〔그림 4－125〕 I-ViT 프레임워크 개요



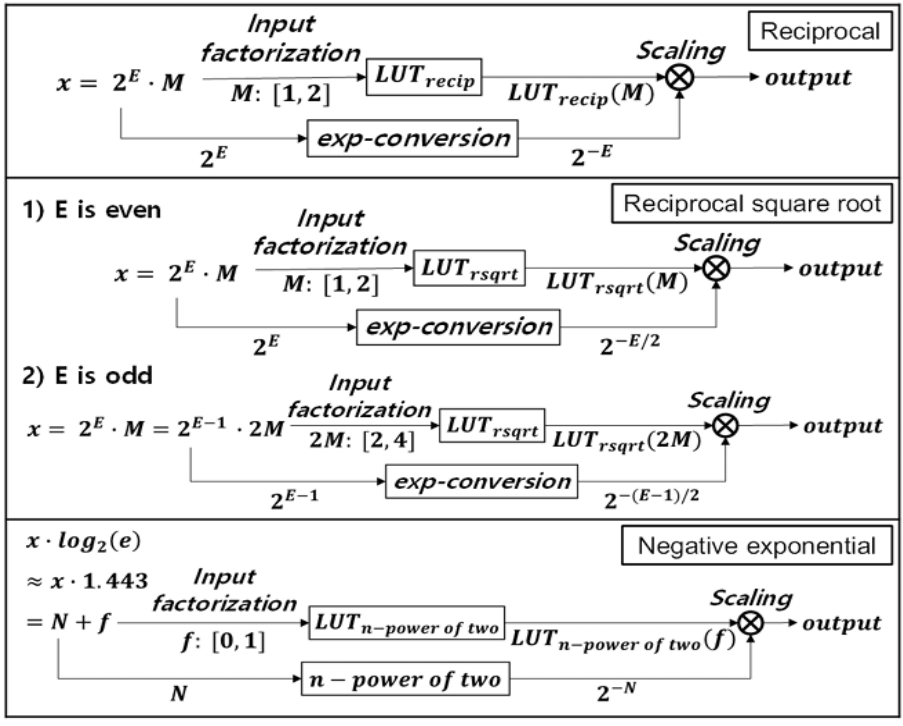
자료: Li, et al.(2023), p.3

해당 연구는 ViT의 높은 연산량과 메모리 사용량으로 인해 엣지 디바이스에서 효율적인 추론이 어려운 문제를 해결하기 위해, I-BERT의 정수 연산 기반 접근 방식을 ViT 구조에 확장한 integer-only quantization framework인 I-ViT를 제안한다〔그림 4－125〕. 기존의 dyadic arithmetic 기반 integer pipeline은 선형 연산에 한정되어 비선형 함수에는 적용되지 못했으며, 따라서 ViT의 완전한 정수화가 불가능했다. 이러한 한계를 극복하기 위해 I-ViT은 비선형 연산을 모두 정수-기반 근사함수로 대체하였다. 구체적으로, Softmax는 지수 계산을 하드웨어 친화적 시프트 연산으로 근사함으로써 연산 효율성을 대폭 향상하였다. 다음으로, 시프트 연산 기반의 지수 근사화를 통해 sigmoid 함수를 정수화하고, 이를 이용해 GELU 함수를 하드웨어 친화적으로 구현함으로써, 기존 I-BERT의 다항식 근사 대비 정확도와 효율을 모두 개선하였다. 마지막으로, LayerNorm의 normalization 항에 포함된 reciprocal square root 연산을 iterative integer approximation 방식으로 구현하여, floating-point 연산 없이도 완전한 정수 도메인에서 LayerNorm을 수행할 수 있도록 하였

다. 실험적으로 I-ViT은 ViT, DeiT, Swin Transformer 등 다양한 ViT 계열 모델에서 INT8 정수 기반 추론 시 FP32 precision 모델의 정확도를 유지하면서, 최대 4.1배 추론 속도 향상을 달성하였다. 또한 TVM 기반으로 RTX 2080Ti GPU의 Turing Tensor Core에서 효율적인 integer-only inference가 가능함을 보였다.

6) Range-invariant approximation of non-linear operations for efficient BERT fine-tuning(Kim, et al.(2023))(DAC'23)

〔그림 4－126〕 Range-invariant LUT Approximation 기반의 비선형 함수 근사화 구조



자료: Kim, et al.(2023), p.3

해당 연구는 기존 LUT 기반 비선형 함수 근사화 방법들이 입력 데이터의 동적 범위 변화에 따라 근사 정밀도가 급격히 저하되어, BERT 파인튜닝 과정에서 레이어 및 태스크별로 LUT를 별도로 최적화해야 하는 비효율이 존재한다는 점을 지적한다. 이를 해결하기 위

해 입력 범위의 변화에 영향을 받지 않는 범용 근사화 프레임워크인 range-invariant look-up table(RI-LUT)을 제안한다 〔그림 4-126〕. RI-LUT은 입력값을 power-of-two scaling과 range-invariant resolution으로 분해하여, 입력 분포의 변화와 무관하게 단일 LUT로 비선형 연산을 근사할 수 있도록 설계되었다. 이를 통해 reciprocal, reciprocal square root, negative exponential 연산을 모두 단일 LUT 기반으로 구현할 수 있으며, LUT 크기를 기존 256-entry에서 1-entry 수준으로 축소하더라도 안정적인 근사 정밀도를 유지한다. 특히 reciprocal square root 연산에서는 floating-point 데이터의 지수와 가수를 분리하여 짝수와 홀수 지수 구간에 맞게 정규화함으로써, 다양한 입력 분포에서도 approximation error를 최소화한다. 또한 sigmoid 및 GELU와 같이 복합 비선형 연산이 중첩된 구조를 단일 근사 단계로 단순화할 수 있는 Simple GELU 방식을 제안하였다.

이 방식은 기존 두 단계의 LUT 근사화 접근을 하나의 RI-LUT 접근으로 병합하여 연산 복잡도를 줄이고 하드웨어 효율을 향상시킨다. 학습 과정에서는 LUT 근사화로 인한 손실 지형 왜곡을 완화하기 위해 상위 일부 Transformer 레이어를 초기화하는 weight re-initialization 기법을 함께 적용하여, 데이터가 제한된 파인튜닝 환경에서도 안정적인 수렴을 유도한다. GLUE benchmark의 다양한 다운스트림 태스크에서 RI-LUT은 기존 linear LUT 방식 대비 평균 정확도 손실 없이 BERT-base 및 BERT-large 모델의 파인튜닝을 모두 지원하였으며, 단일-entry LUT 구성에서도 baseline 수준의 성능을 유지하였다. 하드웨어 합성 결과, RI-LUT은 기존 256-entry LUT 대비 최대 52%의 area 절감과 4% 이내의 추가 logic overhead 만으로 2-stage 근사화 연산을 1-stage로 단순화하였다. 결과적으로 RI-LUT은 입력 데이터의 범위 변화에 영향을 받지 않는 range-invariant 근사화 기법으로, Transformer 기반 모델의 training, fine-tuning의 모든 과정에서 비선형 함수를 단일 LUT 구조로 통합 구현할 수 있는 범용적이고 하드웨어 효율적인 근사화 프레임워크로 평가된다.

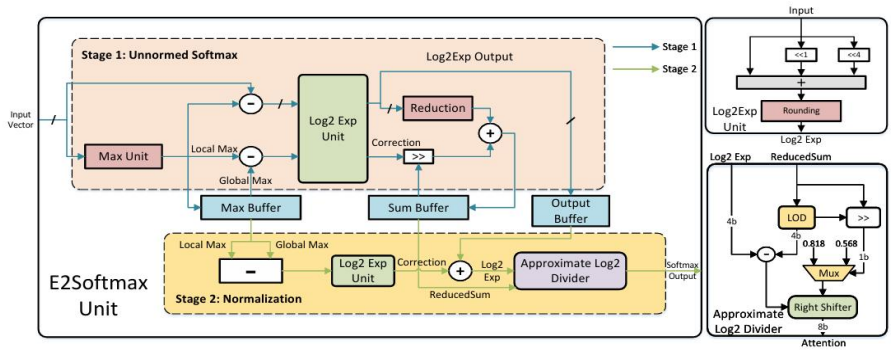
7) SOLE: Hardware-Software Co-design of Softmax and LayerNorm for Efficient Transformer Inference(Wang, et al.(2023))(ICCAD'23)

해당 연구는 Transformer의 Softmax와 LayerNorm 연산이 전체 추론 지연 시간과 에너지 소비의 주요 병목으로 작용함에도 불구하고, 기존 근사화 및 가속화 연구들이 연산 복잡도만을 줄이는 데 집중하여 메모리 접근 오버헤드와 정밀도 손실 문제를 해결하지 못한

다는 점을 지적한다. 이를 해결하기 위해, 하드웨어-소프트웨어 협업 기반으로 Softmax와 LayerNorm 연산을 동시에 최적화하는 통합 프레임워크인 Softmax and LayerNorm Co-design Framework(SOLE)를 제안한다. SOLE은 Softmax 근사 연산을 위한 E2Softmax와 LayerNorm 근사 연산을 위한 AILayerNorm 두 가지 핵심 모듈로 구성된다. 먼저, E2 Softmax는 지수 함수 계산을 위한 곱셈, 나눗셈 연산을 제거하고, log quantization을 적용하여 exponent 연산을 시프트 기반 정수 연산으로 근사화한다. 또한 정규화 항 계산 시 log-domain 기반의 approximate division 방식을 도입하여 LUT나 재학습 없이도 Softmax를 효율적으로 근사화한다〔그림 4－127〕. 다음으로, AILayerNorm 모듈은 입력 통계 계산(평균, 분산) 과정에서 dynamic compression과 power-of-two factorization을 적용하여, 기존 floating-point 연산을 저정밀 정수 연산으로 대체하였다.

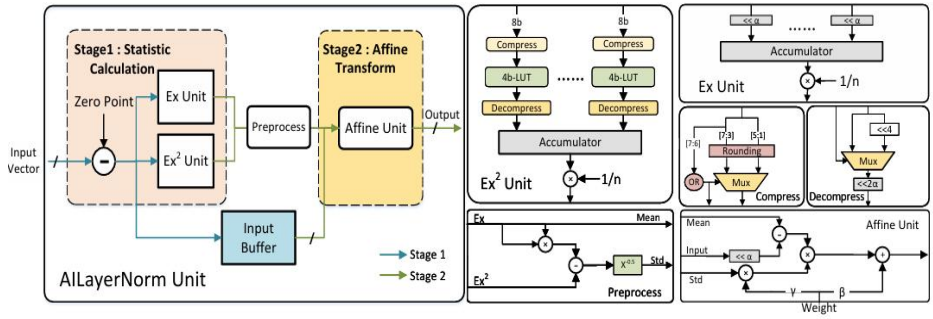
이를 통해 분산의 제곱근 계산을 LUT와 시프트 연산만으로 수행할 수 있으며, normalization 단계까지 완전한 integer-only 형태로 처리할 수 있도록 설계되었다〔그림 4－128〕. 하드웨어 아키텍처는 두 연산 모듈을 공통 버퍼 구조로 통합하고, 파이프라인 기반의 two-stage 연산 경로를 도입하여 연산 중복을 최소화하였다. 이때 multiplier-free 및 LUT-free 구조를 채택함으로써 연산 단순화와 데이터 재활용을 동시에 달성하였다. 실험 결과, SOLE은 NVIDIA GeForce RTX 2080 TI GPU 대비 Softmax와 LayerNorm 각각 36.2배, 61.3배의 속도 향상과 최대 4259배 에너지 효율 향상을 보였으며, 기존 Softermax 및 NN-LUT 대비 각각 3.04, 3.86배 높은 에너지 효율을 달성하였다. 또한 8비트 정밀도 환경에서도 평균 정확도 손실이 0.5% 미만으로 유지되어, 재학습 과정 없이도 원본 모델과 거의 동일한 성능을 보였다. 결과적으로 SOLE은 Softmax와 LayerNorm을 하드웨어-소프트웨어 공동 설계 관점에서 통합적으로 최적화한 프레임워크로, LUT나 고비용 연산 없이도 Transformer의 주요 비선형 연산을 효율적으로 가속화 할 수 있는 저전력 고성능 하드웨어 솔루션이다.

〔그림 4－127〕 SOLE 하드웨어 프레임워크의 E2Softmax 연산기 구조



자료: Wang, et al.(2023), p.5

〔그림 4－128〕 SOLE 하드웨어 프레임워크의 AILayerNorm 연산기 구조



자료: Wang, et al.(2023), p.6

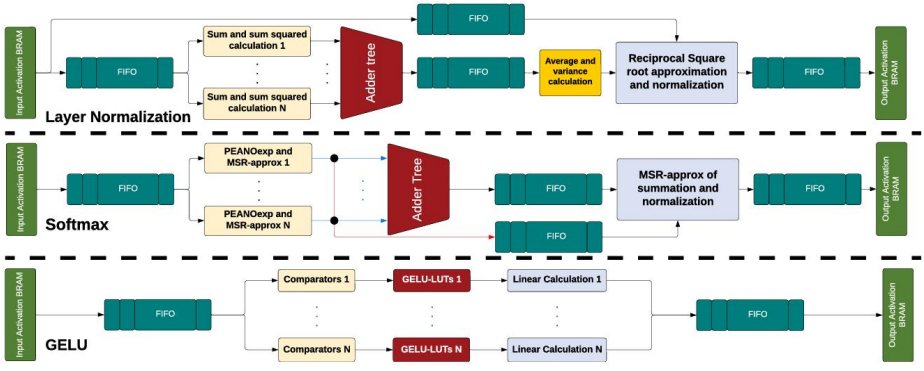
8) Peano-vit: Power-efficient approximations of non-linearities in vision transformers (Sadeghi, et al.(2024))(ISLPED'24)

Peano-vit에서는 ViT의 비선형 연산인 LayerNorm, Softmax, GELU가 FPGA 등 하드웨어 플랫폼에서 연산 복잡도와 자원 소모의 주요 병목으로 작용한다는 점을 해결하기 위해, 세 가지 비선형 연산을 모두 대상으로 한 통합 근사화 프레임워크를 제안한다. PEANO-ViT는 각 비선형 연산의 핵심 수학 연산인 제곱근, 지수, 나눗셈을 대체하기 위해 세 가지 근사화 전략을 도입한다. 먼저, LayerNorm에서는 기존의 제곱근 및 나눗셈 연산을 제거하고, 입력 분산 값의 reciprocal square root를 log domain 기반으로 근사하여 시프

트 연산만으로 구현 가능한 division-free 정규화 기법을 적용하였다. 다음으로, Softmax에서는 Pade 근사식을 이용해 지수 함수를 근사화하고, normalization 단계의 나눗셈 연산을 multi-scale reciprocal approximation 방식으로 대체하여 division-free 연산을 구현하였다. 또한 입력 범위에 따라 동적으로 scaling factor를 조정함으로써 기존 LUT 기반 reciprocal 근사 방식보다 정밀도를 높이고 메모리 사용량을 줄였다. 셋째, GELU 함수는 입력 구간별 비선형성을 최소제곱오차(MSE) 기준으로 분할한 PWL 방식을 적용하여, tanh 및 다항식 계산을 완전히 제거하였다.

FPGA 상의 하드웨어 구현에서는 각 비선형 함수가 병렬적으로 다수의 입력을 처리하도록 구성되었으며, FIFO 기반 파이프라인 구조를 통해 중복된 데이터 접근을 최소화하였다〔그림 4－129〕. 또한 LayerNorm과 Softmax의 이중 입력 읽기 문제를 완화하기 위해 병렬 FIFO 버퍼를 추가하여 전체 latency를 대폭 감소시켰다. ImageNet-1K 데이터셋을 기준으로 DeiT-S/B, Swin-B, ViT-L 모델을 평가한 결과, FP32 대비 정확도 손실은 0.5% 이내로 유지되었다. DeiT-B 기준으로 LayerNorm, Softmax, GELU 각각의 전력 효율은 기존 대비 1.91배, 1.39배, 8.01배 향상되었으며, LUT·DSP·Register 사용량은 평균적으로 60~90% 감소하였다. 결과적으로 PEANO-ViT는 ViT 구조의 핵심 비선형 연산을 모두 division-free 형태로 근사화하여, 고정소수점 환경에서도 높은 정확도를 유지하면서 FPGA 자원 소모와 전력 소비를 크게 줄인 범용 비선형 근사화 프레임워크로 평가된다.

〔그림 4－129〕 Peano-vit 하드웨어 구조

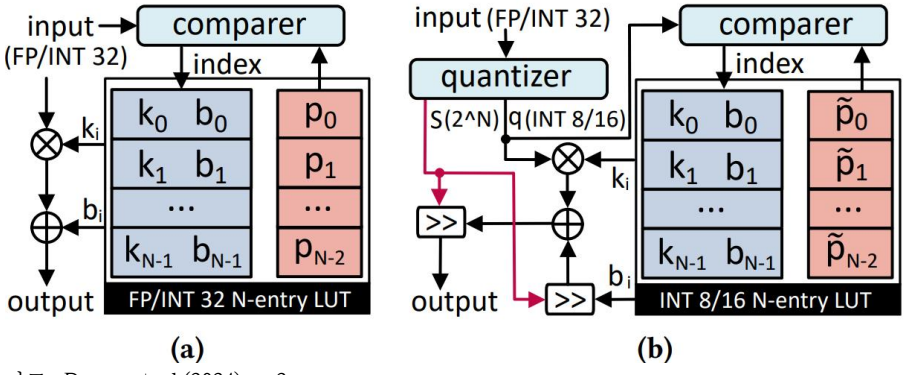


자료: Sadeghi, et al.(2024), p.5

9) Genetic quantization-aware approximation for non-linear operations in transformers (Dong, et al.(2024))(DAC'24)

해당 연구는 Transformer의 비선형 연산의 높은 연산 복잡도와 정밀도 요구로 인해 하드웨어 구현 시 과도한 자원 소모와 전력 소비를 초래하며, 특히 기존 LUT 기반 근사화 방법들이 FP32 또는 INT32 수준의 고정밀 연산에 의존해 실제 정수 양자화 환경에서는 효율성이 급격히 저하된다는 점을 지적한다. 이를 해결하기 위해, 유전 알고리즘 기반의 비선형 함수 LUT 근사화 프레임워크인 GQA-LUT를 제안한다〔그림 4－130〕. GQA-LUT은 입력 scaling factor가 근사화 정밀도에 미치는 영향을 분석하고, 이를 기반으로 LUT의 breakpoint와 slope, intercept를 자동 탐색하는 유전 알고리즘을 도입한다.

〔그림 4－130〕 GQA-LUT 근사화 방법과 FP32 LUT 근사화 방법의 비교



자료: Dong, et al.(2024), p.2

기존의 고정형 파라미터 방식과 달리, GQA-LUT은 scaling factor를 power-of-two 형태로 제한하여 시프트 연산 기반의 경량 정수 연산만으로 근사 파라미터를 계산한다. 또한 입력 범위가 넓은 연산(div, reciprocal square root 등)에 대해서는 multi-range input scaling 기법을 적용하여, 입력 구간을 여러 서브 레인지로 분할하고 각 구간별 scaling factor를 독립적으로 적용함으로써 근사 오차를 최소화한다. 특히, 기존 고정소수점 변환 과정에서 scaling factor가 커질 때 발생하는 breakpoint deviation 문제를 해결하기 위해, GQA-LUT은 rounding mutation 알고리즘을 제안하였다. 이는 breakpoint 양자화를 확률적

돌연변이로 모델링하여 다양한 스케일에서 무작위로 rounding을 적용함으로써, 대규모 스케일링 환경에서도 근사 정확도를 유지하도록 한다. 이러한 전략은 기존 NN-LUT 대비 약 13~26배 낮은 mean squared error를 달성하며, 특히 GELU, HSWISH, exp 근사에서 INT8 정밀도 하에서도 안정적인 오차 특성을 보인다.

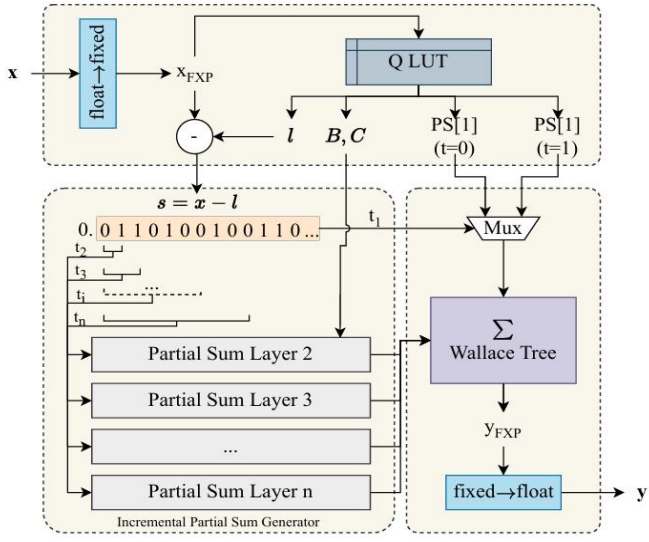
Cityscapes 데이터셋에서 Segformer-B0 및 EfficientViT-B0 모델을 대상으로 한 실험 결과, GQA-LUT은 FP32 대비 성능 손실 0.07% 이하로 유지하면서도 기존 NN-LUT 대비 최대 1.1% 향상된 fine-tuning 정확도를 달성하였다. 또한 TSMC 28nm 공정에서의 하드웨어 합성 결과, INT8 기반 GQA-LUT 유닛은 FP32 및 INT32 대비 각각 81.3~81.7%의 면적 절감과 79.3~80.2%의 전력 절감을 보였으며, 동일한 16-entry LUT 구성에서도 전력·면적 증가를 최소화하였다. 결과적으로 GQA-LUT는 LUT 기반 비선형 함수 근사화에서 최초로 유전 알고리즘과 양자화 인식 구조를 결합함으로써, integer-only 연산 환경에서도 높은 정밀도와 하드웨어 효율을 동시에 달성한 범용 근사화 프레임워크로 평가된다.

10) Algorithm-Hardware Co-design of a Unified Accelerator for Non-linear Functions in Transformers(Du, et al.(2025))(DATE'25)

해당 연구는 Transformer의 비선형 함수가 전체 연산 중 높은 비중을 차지하고, 기존의 PWL 기반 근사화 방법들이 multiply-and-add(MADD) 연산 병목으로 인한 지연과 비선형성이 높은 함수에서의 근사 한계를 가지는 문제를 지적한다. 이를 해결하기 위해, second-order Taylor expansion 기반의 quadratic fitting을 이용한 통합 근사화 및 가속기 프레임워크인 Quadratic approximation-based Unified accelerator for Non-linear Functions (QUNF)를 제안한다[그림 4－131]. QUNF는 각 비선형 함수를 입력 구간별로 분할한 뒤, 각 구간을 이차 함수로 근사하여 기존 선형 근사 대비 더 높은 근사 정밀도를 달성한다. 또한 입력값을 정규화하여 계산하는 hardware-aware approximation 방식을 통해 제곱항을 시프트 연산으로 단순화하고, quantization precision을 조절함으로써 정확도와 자원 효율 간의 균형을 세밀하게 제어할 수 있다. 이를 위해 incremental precision 설계를 도입하여, 단계별 누적 계산 구조를 통해 정밀도 증가에 따른 자원 증가를 선형적으로 제어하도록 하였다. 하드웨어 설계 측면에서는 canonical signed digit(CSD) 인코딩을 활용하여 불필요한 덧셈 항을 제거하고, segmentation factor에 따라 psum을 pruning하는 segment-

aware pruning 기법을 적용하였다. 이 방법은 각 세그먼트 내에서 불필요한 곱셈 및 누산 로직을 제거하면서도 정확도 손실이 발생하지 않도록 한다. 실험 결과, QUNF는 GELU, negative exponential, reciprocal square root 연산 모두에서 기존 RI-LUT, GQA-LUT, PLAC 대비 평균 5% 이상 높은 근사 정확도와 약 60% 낮은 area-delay product를 달성하였다. RoBERTa-large 및 T5-base 모델에 적용한 결과에서도 GLUE benchmark 기준 평균 정확도 손실이 0.72% 이내로 유지되었다. 결과적으로 QUNF는 비선형 함수 전용 unified 근사화, 가속기 구조로서, 기존 LUT 또는 PWL 기반 접근이 segmentation 중심 최적화에 머무른 한계를 넘어, 산술 논리 구조의 병목 자체를 제거한 quadratic approximation 기반 통합 가속기로 평가된다.

〔그림 4－131〕 QUNF 하드웨어 구조



자료: Du, et al.(2025), p.4

제5장 디지털 저탄소 전환을 위한 정책 및 제도적 방안

제1절 연구개발 지원 및 인센티브

1. 저탄소 ICT 연구개발 투자 확대

가. 정부 주도의 저탄소 ICT 연구개발 확대

디지털 탄소중립으로의 전환을 가속화하기 위해 정부는 저탄소 ICT 분야의 연구개발에 대한 투자를 확대하고 있다.[108)109)] 과학기술정보통신부를 중심으로 AI, 클라우드, 네트워크 기술의 에너지 효율을 높이는 기술 개발을 지원하고 있으며, 특히 디지털 인프라의 전력 소비량을 획기적으로 줄이는 데 중점을 둔다. 먼저, AI 및 클라우드 인프라의 효율화를 위해 정부는 AI 연산에 필요한 전력을 최소화하는 기술 개발에 집중 투자한다. 인공지능 모델 자체의 구조를 최적화하여 연산 효율을 높이는 스파스 모델링이나, 데이터의 정밀도를 낮춰 연산량을 줄이는 저정밀(quantization) 추론 기술 등이 주요 연구 대상이다.[110)] 이러한 기술은 AI 학습과 추론 과정에서 발생하는 막대한 전력 소비를 줄여 데이터센터의 탄소 배출량을 직접적으로 감축하는 효과가 있다. 다음으로, 저전력 네트워크 장비 개발도 핵심 과제다. 5G/6G 통신망과 같은 네트워크 인프라의 전력 소모를 줄이는 기술 개발에 집중하여 통신 네트워크의 탄소 배출량을 효과적으로 감축하고자 한다. 네트워크 장비의 효율을 높이는 기술은 디지털 전환 시대에 급증하는 데이터 트래픽으로 인한 에너지 문제를 해결하는 데 필수적이다. 마지막으로, 데이터센터 냉각 기술을 최적화하는 연구도 활발히 진행된다. 데이터센터의 냉각에 사용되는 막대한 전력을 절감하기 위해 공기 냉각 방식의 한계를 극복하는 액체 냉각(liquid cooling)과 같은 혁신적인 방안을 모색하고 있다. 이러한

108) 과학기술정보통신부, 「디지털 전환을 통한 탄소중립 촉진방안」, 2023.

109) 대한민국 정부, 「디지털 탄소중립 전담반 개최」, 2023.

110) 한국과학기술기획평가원(KISTEP), 「AI·클라우드·네트워크 저전력 기술 R&D 로드맵」, 2024.

기술은 데이터센터의 에너지 효율을 극대화하여 운영 비용을 절감하는 동시에 탄소 배출량을 대폭 줄일 수 있다. 이처럼 정부는 기술 개발에 대한 투자를 확대하고, 민간 부문과의 협력을 강화하여 디지털 기술이 탄소중립을 위한 핵심적인 해결책으로 자리매김하도록 지원하고 있다.

나. 민관 협력 기반의 저탄소 ICT 혁신 생태계 조성

저탄소 ICT 기술의 실질적 확산과 지속가능한 혁신을 위해서는 정부 주도의 연구개발을 넘어, 민간 부문과의 긴밀한 협력이 필수적이다. 정부는 공공 연구개발 성과를 민간 기업에 개방하고, 산업계의 수요를 반영한 기술개발 방향을 설정함으로써 시장 친화적인 R&D 체계를 구축해야 한다. 이를 위해 다음과 같은 전략이 필요하다

첫째, 공공-민간 공동 연구 플랫폼의 활성화이다. 공공연구기관과 민간 기업 간의 공동연구를 촉진하기 위해 개방형 테스트베드(Open Testbed) 및 실증 인프라를 확충하고, 기술 이전 및 사업화를 위한 중개 지원체계를 강화해야 한다. 예를 들어, 에너지 효율형 AI 반도체나 저전력 데이터센터 기술 개발과 같은 분야에서는 대기업·중소기업·스타트업이 함께 참여하는 연합형 R&D 구조가 효과적이다.

둘째, 탄소저감 효과를 정량화할 수 있는 민간 참여형 평가체계 구축이다. 기술개발 성과가 실제로 온실가스 감축에 얼마나 기여했는지를 객관적으로 평가하기 위해, 표준화된 저탄소 기술 인증제도 및 에너지 효율지표를 마련해야 한다. 이러한 제도는 기업이 친환경 기술 투자에 따른 인센티브를 명확히 인식하도록 돕는다.

셋째, 세제·금융 인센티브를 통한 민간 투자 촉진이다. 저탄소 ICT 기술에 투자하는 기업에 대해 연구개발비 세액공제, 녹색채권 발행 지원, 탄소감축 실적 기반의 저리 융자 프로그램 등을 도입하여 민간의 자발적 투자를 유도할 수 있다. 특히 중소기업과 스타트업이 에너지 효율 기술을 상용화할 수 있도록, 정부 보증 기반의 '녹색 R&D 펀드'를 조성하는 방안이 효과적이다.

넷째, 지속 가능한 산업 생태계 조성을 위해 기술개발 이후의 표준화·인증·유통·재활용 체계를 연계적으로 지원해야 한다. 저전력 반도체, PIM(Processing-In-Memory), 저탄소 네트워크 장비 등 핵심 기술의 국제 표준화를 선점함으로써 글로벌 경쟁력을 확보할 필요가 있다.

결국, 정부는 정책적·재정적 지원의 중심축으로서 방향성을 제시하고, 민간은 기술혁신과 시장 확산의 실행 주체로서 역할을 수행함으로써 저탄소 ICT 산업의 지속가능한 성장을 달성할 수 있다.

2. 혁신기술 상용화를 위한 금융 · 세제 지원

가. 세제 지원

저탄소 디지털 전환을 가속화하기 위해, 우리는 에너지 효율을 극대화하는 저전력 반도체와 에너지 절감 소프트웨어 같은 혁신 기술의 상용화를 촉진해야 한다. 이를 위한 구체적인 다음과 같은 세제 지원 방안이 필요하다. 먼저, 기업이 저전력 반도체나 에너지 절감 소프트웨어 개발에 투자하는 비용에 대해 연구개발(R&D) 세액공제를 확대 적용한다. 이는 기술 개발에 들어가는 초기 부담을 줄여 기업의 적극적인 투자를 유도하는 중요한 정책이다[111]. 또한 해당 기술이 적용된 설비에 대해서는 신성장동력·원천기술 세액공제를 적용하여, 실제 상용화 단계에서의 기업 부담을 경감시키고 있다. 이러한 세제 혜택은 단순히 연구에 그치지 않고 시장에서 실제 제품으로 이어지도록 하는 강력한 동력이 된다[112]. 더 나아가, 저탄소 기술 개발에 필요한 핵심 장비나 원자재를 수입할 경우, 관세를 감면하거나 일시적으로 면제해주는 제도를 운영한다. 이를 통해 해외 기술 도입 및 첨단 장비 확보에 대한 비용 부담을 낮춰 국내 기술 경쟁력을 강화하고, 글로벌 기술 동향에 빠르게 대응할 수 있도록 돕는다[113]. 아울러, 저탄소 기술을 활용하여 에너지 효율을 개선한 기업에게는 지방세를 감면하는 혜택을 제공함으로써, 특히 데이터센터와 같이 대규모 시설을 운영하는 기업들이 실질적인 비용 절감 효과를 체감하고 저탄소 기술 도입을 적극적으로 고려하도록 장려해야 한다. 이러한 세제 지원 방안은 혁신적인 저탄소 기술이 시장에서 활발하게 적용될 수 있는 기반을 마련해줄 것이다.

나. 금융·투자 지원

저전력 반도체 및 에너지 절감 소프트웨어 등 혁신기술의 상용화를 촉진하기 위한 정부

111) 산업통상자원부, 「2022년도 국가표준시행계획」, 2022.
112) 과학기술정보통신부, 「디지털 기반 탄소중립 추진」, 2021.
113) 국회도서관 국가전략정보포털, 「탄소중립 녹색성장 추진전략」, 2023.

의 노력은 세제 지원에 그치지 않고, 금융 및 투자 지원으로도 이어진다.[114] 정부는 이러한 기술을 개발하고 상용화하려는 기업들이 안정적인 자금을 확보할 수 있도록 다양한 정책 금융을 제공하고 있다. 첫째, 정책자금 지원을 통해 기술 개발 초기 단계의 중소·중견기업을 돕는다.[115] 기술보증기금이나 신용보증기금 등 정책 금융기관을 활용하여 저탄소 기술 개발 기업에 대한 보증 지원을 확대하고, 저금리 대출을 제공함으로써 자금 조달의 어려움을 해소해준다.[116] 이러한 지원은 특히 자본력이 부족한 스타트업이나 벤처기업이 혁신적인 아이디어를 구체적인 기술로 발전시키는 데 큰 도움이 된다.

둘째, 민간 투자 활성화를 위한 다양한 제도적 장치를 마련하고 있다. 정부는 민간 벤처캐피탈이 저탄소 기술 기업에 투자할 경우, 투자 리스크를 줄여주기 위해 매칭 펀드를 조성하거나, 세액공제 혜택을 제공하는 방안을 추진한다. 또한 저탄소 기술 분야의 유망 기업을 발굴하고, 이들 기업에 대한 정보와 투자 기회를 제공하는 플랫폼을 구축하여 민간 자금이 자연스럽게 유입될 수 있는 환경을 조성하고 있다.

셋째, 기술사업화 지원을 강화한다. 기술 개발이 완료된 후 시장에 성공적으로 진입할 수 있도록, 시제품 제작 지원, 판로 개척, 해외 시장 진출을 위한 마케팅 지원 등을 포함한 종합적인 금융 지원 프로그램을 운영한다[117]. 이와 함께 기술가치평가를 통해 기업의 무형자산인 기술력을 금융기관이 객관적으로 평가할 수 있도록 돕는 시스템을 구축하여, 담보가 부족한 기술 기업도 원활하게 투자를 유치할 수 있도록 지원하고 있다.

3. 신기술 실증사업으로 시장 진입 가속화

가. 민관 협력 기반의 실증사업 추진

정부는 저탄소 ICT 기술의 실효성을 검증하고 산업 현장에 신속히 확산하기 위해 공공과 민간이 공동으로 참여하는 실증사업을 추진하고 있다. 이러한 실증사업은 기술 시연을 넘어 산업별 맞춤형 적용 모델을 개발하여 시장 도입과 상용화를 동시에 견인하는 것을

114) 정부 및 정책금융기관의 기후위기 대응 금융지원 확대방안(2024), greenium.kr
115) 2030년까지 기후위기대응에 452조 민관금융지원 발표(2024), 연합뉴스
116) 담보력 부족한 녹색기업 우대보증 지원 정책(2025), 환경부
117) 기후위기 대응을 위한 금융지원 관련 은행장 회의(2024), 금융감독원

목표로 한다. 특히 에너지 다소비 산업군(철강, 석유화학, 시멘트, 반도체, 디스플레이 등)을 중심으로 AI·IoT 기반 에너지 효율화 기술 실증이 진행되고 있다. 예를 들어 제조 현장에 AI 기반 에너지 관리 시스템(AI-EMS)을 도입해 공정별 전력·열·가스 사용 데이터를 실시간 수집·분석하고, 설비별 최적 운전 조건과 에너지 절감 시나리오를 자동 도출하는 프로젝트가 대표적이다.[118]. 이러한 시스템은 모니터링을 넘어 디지털 트윈과 예측 제어를 결합해 제어 효율을 높이고, 설비 이상 징후를 조기 감지해 에너지 낭비를 최소화하도록 설계된다. 실증 참여 기업은 평균 10~25%의 에너지 사용 절감 효과를 거두는 것으로 보고되고 있다. 정부는 실증사업 확산을 위해 재정·제도적 인센티브를 병행한다. 기술 개발 기업에는 현장 테스트와 장비 구축 비용을 연구개발 자금으로 지원하고, 도입 기업에는 시스템 설치·운용 비용을 보조금 또는 세액공제로 보완한다. 아울러 실증 단계에서 규제 샌드박스를 적용해 신기술 검증의 행정 부담을 완화한다[119].

실증 성과의 축적과 확산을 위해 저탄소 ICT 기술 실증 데이터 플랫폼을 구축·운영한다. 이를 통해 참여 기업뿐 아니라 타 산업군 기업도 실증 데이터를 참고해 유사 환경에서의 적용 타당성을 사전에 검토할 수 있어 기술 확산 속도가 빨라진다. 성공 사례는 정책적 표준모델로 정리되어 전국적 확산의 근거가 되고, 민간의 자발적 투자와 기술 도입을 촉진하는 시장 견인형 메커니즘으로 작동한다.

나. 기술 확산을 위한 표준화 및 정보 공유

파일럿 프로젝트와 실증사업을 통해 확보한 데이터와 노하우는 단순히 일부 기업의 성과로 한정되어서는 안 되며, 산업 전반으로 확산될 수 있는 체계적 기반으로 전환되어야 한다. 이를 위해 정부는 실증사업의 결과를 바탕으로 저탄소 ICT 기술의 표준화 작업을 적극 추진하고 있다.[120] 기기술 표준화는 산업 내 다양한 주체가 동일한 기준에 따라 기술을 개발하고 적용할 수 있게 함으로써, 기술의 신뢰성과 호환성을 높이고 시장 진입 장벽을 완화하는 핵심 수단이다. 예를 들어, 서로 다른 기업이 개발한 에너지 관리 솔루션이 동일한 데이터센터 인프라에서 원활히 연동되도록 하기 위해 정부는 공통의 데이터 전송 규격,

118) 산업통상자원부, 「저탄소 산업 전환 민관 협력 로드맵」, 2024
119) 한국에너지공단, 「산업 부문 에너지 관리 실증사업 보고서」, 2023
120) 국가표준원, 「저탄소 ICT 기술 표준화 추진계획」, 2024

통신 프로토콜, 에너지 효율 측정 기준 등을 단계적으로 마련하는 것이다[121].이러한 표준은 기술 간 호환성을 높여 중복된 개발을 줄이고, 산업 전반의 기술 도입 비용을 완화하는 효과를 가져올 수 있다.

또한 실증사업을 통해 확보된 기술적 정보와 성과 데이터를 체계적으로 정리해 산업계가 활용할 수 있도록 하는 공개 데이터베이스 구축이 요구된다. 실증 과정에서 확인된 효율성, 비용 절감 효과, 적용 환경, 운영상의 문제점 등은 산업별로 유형화하여 기록함으로써, 후속 연구개발과 상용화 단계에서 참고할 수 있는 근거 자료로 활용될 수 있다. 이때 데이터 공개는 참여 기업의 기밀과 상업적 이해를 보호하기 위해 비식별화와 보안 등급 체계를 병행할 필요가 있다. 기술 표준화와 데이터 공유는 단일 정책의 결과물이 아니라, 지속적인 협의와 검증 과정을 통해 완성된다. 산업별 협의체를 중심으로 기업, 연구기관, 인증기관이 참여하는 논의 구조를 마련하고, 표준의 적용 범위와 세부 사양을 단계적으로 조정해 나가는 것이 중요하다. 이를 통해 표준은 고정된 규격이 아니라, 새로운 기술과 시장 변화를 반영해 점진적으로 확장되는 유연한 틀로 기능할 수 있다.

이와 같은 제도적 기반이 마련되면 실증 단계에서 검증된 기술들이 다양한 산업 영역으로 이전되고, 후발 기업들도 이미 검증된 기준을 바탕으로 기술을 개발하거나 도입할 수 있게 된다. 결국 표준화와 정보 공유는 저탄소 ICT 기술이 개별 사례를 넘어 산업 전반으로 확산되는 과정에서 핵심적인 연결 고리로 작용하게 된다.

4. 저탄소 디지털 기술 기업 육성으로 신산업 창출 및 일자리 증대

가. 신산업 창출과 시장 확대

저탄소 디지털 기술은 AI 기반 에너지 관리 시스템, 저전력 반도체, 스마트 그린 데이터센터 솔루션 등 다양한 분야에서 혁신을 촉진하는 핵심 동력이다[122]. 이러한 기술은 단순히 기존 산업의 효율성을 개선하는 수준을 넘어, 산업 구조의 전환과 새로운 가치사슬 형성을 이끄는 기반 기술로 평가된다. 특히 AI와 반도체, 클라우드, 네트워크 기술이 융합되면서 에너지 절감과 탄소 감축을 동시에 달성할 수 있는 새로운 산업 영역이 빠르게 부상하

121) 탄소중립위원회, 「탄소중립 녹색성장 추진전략」, 2023

122) 산업통상자원부, 「기후테크 산업의 성장모델 창출 및 수출산업 전략 수립」, 2023

고 있다. 예를 들어, AI를 활용해 공장·빌딩·데이터센터의 전력 소비 패턴을 분석하고 최적의 운전 조건을 자동 제안하는 에너지 관리 솔루션 기업들은 기존 전력 관리 시장의 한계를 넘어, 데이터 기반의 예측형 에너지 운영 시장을 개척하고 있다. 이와 같은 기업들은 전력 인프라, 제조, 물류 등 다양한 분야로 기술 적용 범위를 확장하며 산업 간 경계를 허물고 있다. 실제 사례로, AI 기반 에너지 절약 솔루션을 개발한 스타트업들은 기존의 하드웨어 중심 효율 개선에서 벗어나, 소프트웨어와 데이터 분석을 결합한 에너지 서비스 시장을 창출하고 있다[123]. 이러한 기업들은 고성능 연산, 클라우드 인프라, 엣지 단말 등 ICT 기술을 활용해 실시간 최적화와 탄소 배출 모니터링을 동시에 수행하며, 결과적으로 새로운 형태의 '기후테크(Climate Tech)' 산업군을 형성하고 있다. 정부는 기후테크 산업 육성 전략을 통해 이와 같은 기업을 발굴하고, 기술 개발과 상용화 단계를 연계하는 지원 체계를 마련하고 있다. 연구개발 자금 지원, 실증 인프라 제공, 표준화 및 인증 지원, 해외 진출 컨설팅 등 다층적 지원이 논의되고 있으며, 이를 통해 초기 기술 보유 기업이 시장 진입에 필요한 제도적·재정적 장벽을 완화할 수 있도록 한다.

이러한 접근은 단기적인 산업 지원을 넘어, 저탄소 기술을 중심으로 한 신산업 생태계 형성의 기초를 마련하는 방향으로 이어질 수 있다. 기술 혁신을 중심으로 한 기업 성장, 시장 수요의 확대, 제도적 지원이 유기적으로 맞물릴 때 저탄소 디지털 산업은 새로운 성장축으로 자리 잡을 가능성이 높다. 따라서 저탄소 디지털 기술은 에너지 효율 향상과 탄소 감축을 위한 수단이자, 동시에 국가 산업구조의 고도화를 이끌 핵심 전략 분야로 인식될 필요가 있다.

나. 고급 인력 수요 증가와 일자리 증대

신산업의 성장은 자연스럽게 새로운 형태의 전문 인력 수요를 창출하며, 고용 구조의 변화를 동반한다. 저탄소 디지털 기술 분야에서는 인공지능 알고리즘 개발자, 에너지 효율 컨설턴트, 저전력 반도체 설계자, 클라우드 및 데이터센터 아키텍트 등 다양한 영역의 고급 기술 인력이 요구된다. 이러한 인력은 기술 개발뿐 아니라, 시스템 통합과 운영, 데이터 분석, 보안 관리 등 실제 산업 현장에서의 실무적 역량까지 포괄해야 하므로 양성과 확보가 중요한 과제로 부상하고 있다. 정부는 이와 같은 인력 수요에 대응하기 위해 산학협력

123) 엔비디아, 「AI 기반 에너지 절약 솔루션 스타트업 사례」, 2025

중심의 전문 교육 체계를 확대하고 있다.[124] 대학과 기업이 공동으로 참여하는 맞춤형 교육 과정이 개설되고 있으며, 특히 반도체·AI·에너지 융합 분야에서는 실무 중심의 프로젝트형 교육이 강조되고 있다. 또한 현장 실습과 연계된 학위·비학위 과정을 통해 산업 현장에서 즉시 투입 가능한 실무형 인재를 확보하려는 움직임이 나타나고 있다.

전문성 인증을 위한 자격제도 마련도 병행되고 있다. 정부는 저전력 설계, AI 기반 에너지 관리, 클라우드 효율화 등 세부 분야별로 직무 표준을 정립하고, 이를 기반으로 한 자격검정 제도 도입을 검토하고 있다. 이러한 제도는 산업 내 기술 수준을 객관화하고, 기업의 인재 채용 과정에서 기술 역량을 평가할 수 있는 지표로 활용될 수 있다. 한편, 기업의 인력 확보 부담을 완화하기 위해 인건비 일부를 지원하거나, 전문인력 채용 시 세제 혜택을 제공하는 방안도 시행되고 있다. 스타트업이나 중소기업이 우수한 연구개발 인력을 영입할 수 있도록 정부 보조금과 기술 인턴십 프로그램을 연계하는 사례가 늘어나고 있으며, 이는 청년층의 진입 경로를 넓히는 역할을 한다.

이러한 인재 양성 및 고용 촉진 정책은 단순한 인력 공급의 차원을 넘어, 신산업 성장의 기반을 형성하는 핵심 요소로 작용한다. 전문 기술 인력이 확보되어야 산업 내 연구개발의 연속성이 유지되고, 기업 간 기술 협력과 생태계 확장이 가능해진다. 따라서 저탄소 디지털 기술 산업의 인재 양성은 환경적 목표와 경제적 성장을 함께 달성하기 위한 구조적 과제로 인식될 필요가 있다.[125]

5. 국가 탄소중립 전략에 부합하는 R&D 투자

가. 국가 전략과 연계

저탄소 디지털 기술 개발은 정부가 제시한 '2050 탄소중립 시나리오'와 '탄소중립 기술혁신 전략'의 핵심 축을 이룬다.[126] 정부는 디지털 전환이 에너지 효율을 높이고, 산업 전반의 탄소 배출량을 줄이는 데 필수적인 요소임을 명확히 인식하고 있다. 이에 따라, 저전력 반도체, AI 기반 에너지 관리 시스템 등 디지털 기술을 활용한 탄소 감축 기술은 국가의 중

124) 고용노동부, 「디지털 혁신과 저탄소 시대 일자리 해법」, 2022
125) 국가인공지능위원회, 「미래사회를 대비한 산학협력 인력양성」, 2020
126) 2050 탄소중립 시나리오(탄소중립위원회, 2021)

점 투자 분야로 지정되어 R&D 예산이 집중적으로 투입되고 있다. 이러한 기술 개발은 녹색성장 5개년 계획에 포함되어 범부처 차원의 협력을 통해 체계적으로 추진되고 있다. 저탄소 디지털 기술은 단순히 기술적 혁신을 넘어, 국가적인 차원의 기후변화 대응 목표를 달성하기 위한 필수적인 전략 수단으로 자리매김하고 있다.

나. 국제 협력

기후변화 대응은 단일 국가의 노력만으로는 한계가 있으며, 기술 개발과 확산 과정에서의 국제 협력이 필수적이다. 저탄소 디지털 기술 또한 이러한 국제적 연대의 틀 속에서 발전해야 하는 영역으로, 각국이 보유한 기술과 경험을 공유하고 공동의 기준을 마련하는 것이 점점 중요해지고 있다. 정부는 기후변화 대응과 관련된 다양한 국제 협의체와 다자간 협력 프로그램에 참여하며, 저탄소 디지털 기술 분야의 공동 연구개발을 추진하고 있다. 이를 통해 각국이 축적한 기술적 역량을 상호 보완하고, 에너지 효율 향상과 온실가스 감축을 위한 새로운 접근법을 함께 모색하는 기반이 마련되고 있다. 특히 연구개발 초기 단계에서부터 공동 프로젝트를 구성하거나, 실증 데이터를 상호 교환하는 형태의 협력이 확대되는 추세이다. 주요 협력 대상은 유럽연합(EU), 미국, 일본 등 탄소중립 정책을 선도하는 국가 및 지역으로, 이들과의 협력을 통해 저전력 AI 반도체, 고효율 연산 알고리즘, 탄소 모니터링 플랫폼 등 핵심 기술 분야에서의 상호 발전이 기대된다. 결국 저탄소 디지털 기술의 국제 협력은 기술 교류 이상의 의미를 지닌다. 이는 기후변화 대응이라는 공통 목표 아래 각국이 기술, 제도, 산업 구조의 전환 경험을 공유하며, 장기적으로 탄소중립 사회로의 전환을 촉진하는 핵심 동력이 된다.

제 2 절 에너지 효율 기준 및 규제강화

1. ICT 인프라에 대한 에너지효율 기준 제정 및 단계적 강화

디지털 탄소중립을 실현하기 위해서는 ICT 인프라의 에너지 소비를 줄이는 것이 필수적이다. 정부는 이 문제를 해결하기 위해 데이터센터, 통신시설 등 ICT 인프라에 대한 에너지효율 기준을 제정하고 단계적으로 강화하는 정책을 추진하고 있다. 이는 ICT 부문의 자발적인 에너지 절감을 유도하고, 저전력 기술 도입을 가속화하는 핵심적인 방안이 된다.

가. 데이터센터 전력효율지수(PUE) 의무화 및 기준 강화

정부는 데이터센터의 에너지효율을 측정하는 가장 보편적인 지표인 PUE를 기준으로 삼아 효율 기준을 제정하고 있다. PUE는 IT 장비가 소비하는 전력 대비 전체 전력 소비량의 비율을 나타내는데, 현재 많은 데이터센터의 PUE는 1.8에서 2.0 이상인 경우가 많아 전력의 45 ~ 50%가 냉각, 조명 등 IT 장비 외적인 용도로 낭비되고 있다.[127] 이러한 비효율을 개선하기 위해 정부는 신규 데이터센터 설립 시 PUE 기준을 1.4 이하로 설정하고, 기존 시설에 대해서는 주기적인 에너지 진단과 함께 PUE 개선 목표를 부여하는 방안을 추진하고 있다[128]. 이와 더불어, 고효율 시설에 대한 인센티브를 제공하고 PUE 기준을 충족하지 못할 경우 페널티를 부과하는 등 강제성을 높여 데이터센터 운영사들이 고효율 냉각 시스템, 서버 가상화, AI 기반 에너지 관리 시스템 등 에너지 절감 기술을 적극적으로 도입하도록 촉진하는 효과를 가져올 수 있다.[129]

나. 고효율 ICT 장비 도입 의무화

정부는 데이터센터와 통신시설에 사용되는 ICT 장비 자체의 효율을 높이기 위한 규제도 강화하고 있다[130]. 고효율 서버, 스토리지, 네트워크 장비 등의 도입을 의무화하는 정책이

127) [코스콤리포트] 디지털플랫폼정부 시대와 데이터센터, 2023
128) 한국에너지공단, 「에너지다소비사업장 에너지진단 지침」, 2023
129) Cushman&Wakefield, 한국 데이터센터 산업: 전력 수급과 규제 강화 속 성장과 과제, 2024
130) 정부 통합 전산센터의 그린화 정책, Korea Science, 2009

그 예다. 이는 장비 제조사들에게 저전력 기술 개발을 요구하는 동시에, ICT 인프라 전반의 에너지 소비를 근본적으로 줄이는 효과를 낳는다. 특히, 저전력 반도체나 에너지 절감 소프트웨어가 탑재된 장비에 대해 에너지효율 등급제나 인증 제도를 도입하고, 이를 통과한 장비만 공공기관 및 민간 시장에서 사용하도록 하는 방안도 검토하고 있다.[131] 이러한 규제는 ICT 산업 전체의 기술 혁신을 유도하고, 장기적으로는 ICT 부문의 탄소 발자국을 크게 줄이는 데 기여할 것이다. 궁극적으로, 에너지효율 기준 강화는 ICT 부문의 지속가능한 디지털 사회로 나아가기 위한 필수적인 제도적 기반이 된다.

2. ICT 기업과 센터의 온실가스 배출량 모니터링

가. 온실가스 배출량 측정 및 보고 의무화

ICT 산업은 전력 집약적 특성을 지니고 있어, 디지털 전환의 확산과 함께 에너지 사용량과 탄소 배출량이 지속적으로 증가하고 있다. 이에 따라 주요 선진국을 중심으로 데이터센터, 통신 인프라, 반도체 공정 등 ICT 산업 전반의 탄소 배출을 정량적으로 관리하기 위한 제도적 장치가 강화되는 추세이다. 이러한 흐름에 발맞추어, 정부는 ICT 기업과 대규모 데이터센터를 대상으로 한 온실가스 배출량 측정 및 보고 의무화 제도를 단계적으로 도입할 필요가 있다. 이 제도는 기존의 기업 단위 배출 보고체계를 넘어, ICT 산업 특성을 반영한 세분화된 배출 관리 체계를 구축하는 것을 목표로 한다. 구체적으로는 Scope 1(직접 배출), Scope 2(간접 배출)뿐 아니라, 공급망과 서비스 운영 등에서 발생하는 Scope 3(기타 간접 배출)까지 포함하여 ICT 생태계 전반의 탄소 배출량을 포괄적으로 산정하는 방식이다. 이를 통해 기업별, 사업 유형별 배출 특성을 명확히 구분하고, 보다 정밀한 감축 정책 수립이 가능해진다. 측정과 보고의 표준화를 위해서는 통일된 산정 방식과 검증 절차가 필요하다. 정부는 국제적으로 통용되는 탄소 배출 산정 기준(GHG Protocol, ISO 14064 등)을 바탕으로 ICT 산업 특화형 가이드라인을 마련하고, 데이터 수집 주기, 단위, 검증 절차를 명시한 표준 매뉴얼을 제시할 필요가 있다. 또한 보고된 데이터의 신뢰성을 확보하기 위해 제3자 검증 기관을 통한 정기 평가 체계를 운영하고, 기업의 자발적 감축 활동을 인증·공시할 수 있는 제도적 기반을 함께 마련하는 것이 바람직하다. 온실가스 배출량의 측

131) 에너지관리공단, “효율등급제도 제도개요”, 2024

정 및 보고 의무화는 단순한 규제 수단이 아니라, 기업이 자사의 탄소 배출 구조를 객관적으로 파악하고, 감축 전략을 수립하기 위한 출발점이 된다. 이를 통해 ICT 산업 전반에서 탄소 회계(carbon accounting)의 투명성이 강화되고, 장기적으로는 기업 간 감축 성과 비교, 녹색금융 연계, ESG 평가 등 다양한 분야로 파급될 수 있는 기반이 형성된다.

나. 공시 의무화 및 정보 공개

온실가스 배출량의 측정과 보고가 제도적으로 정착되기 위해서는, 그 결과를 외부에 공개하여 사회적 검증이 가능하도록 하는 투명성 확보 장치가 병행되어야 한다. 이에 따라 ICT 기업과 대규모 데이터센터를 포함한 주요 배출원 기업을 대상으로, 보고된 온실가스 배출량 정보를 외부에 공시하도록 의무화하는 제도 도입이 필요하다. 이러한 공시 제도는 기업의 환경적 책임 이행 수준을 객관적으로 파악할 수 있는 근거를 제공하며, 이해관계자들이 기업의 지속가능 경영 수준을 평가할 수 있도록 한다. 공시 정보는 투자자, 소비자, 시민단체 등 다양한 이해관계자에게 중요한 참고 지표로 활용된다. 예를 들어, 투자자는 공시된 배출량과 감축 목표의 이행 수준을 기반으로 기업의 장기적 리스크를 평가할 수 있으며, 소비자는 친환경 경영을 실천하는 기업을 선택할 수 있는 근거를 확보하게 된다. 시민단체와 연구기관 역시 이러한 데이터를 활용해 산업 전반의 감축 추세를 분석하거나 정책 개선 방향을 제안할 수 있다. 공시된 배출 정보는 또한 기업 간 자발적인 감축 경쟁을 촉진하는 효과를 가져올 수 있다. 동일 업종 내 기업들이 공시된 데이터를 통해 서로의 감축 성과를 비교하게 되면, 자연스럽게 효율적 감축 전략의 확산과 기술 혁신이 유도될 가능성이 높다. 이러한 경쟁은 단순한 규제 이행을 넘어, 산업 내 전반적인 탄소 감축 역량 향상으로 이어질 수 있다. 이와 같은 공시 의무화와 정보 공개 체계는 단순한 행정 절차를 넘어, 기업의 ESG경영을 실질적으로 강화하는 제도적 장치로 기능할 수 있다. 투명한 정보 공개는 시장의 신뢰를 높이고, 친환경 투자를 촉진하는 유인으로 작용함으로써 저탄소 디지털 산업의 지속가능한 성장 기반을 조성하는 핵심 요소가 된다.

3. 친환경 ICT 녹색인증제 도입

가. 녹색인증제도 운영 및 범위 확장

저탄소 디지털 전환의 확산을 위해서는 기업의 탄소 감축 노력을 객관적으로 평가하고,

그 결과를 제도적으로 보상할 수 있는 인증체계의 정립이 필요하다. 이를 위해 정부는 기존의 녹색인증제도를 ICT 산업 전반으로 확대 적용하고, 데이터센터·클라우드 기업·반도체 제조업 등 주요 에너지 다소비 업종을 인증 대상에 포함하는 방안을 검토할 수 있다.

녹색인증제도는 일정 수준 이상의 온실가스 감축 실적과 에너지 효율 개선을 달성한 기업이나 시설을 공식적으로 인증하고, 해당 기업에 세제·금융·조달 등 다양한 인센티브를 부여하는 제도이다. 인증을 받은 기업은 환경경영 수준이 객관적으로 입증되어 친환경 투자 유치, 해외 시장 진출, ESG 평가 대응 등에 유리한 위치를 확보할 수 있다. 이 제도는 단순히 감축 결과를 평가하는 데 그치지 않고, 감축 목표 달성 과정에서의 기술 혁신과 운영 개선 노력까지 종합적으로 반영하는 방식으로 운영될 수 있다. 예를 들어, 데이터센터의 경우 에너지 효율지표(PUE), 재생에너지 사용 비율, 냉각 효율성, 폐열 회수 시스템 구축 여부 등을 종합적으로 평가하고, 통신기업은 네트워크 장비의 효율화, 탄소중립망 구축 계획 등을 기준으로 인증을 부여할 수 있다.

아울러 인증제도는 배출량 초과 기업에 대한 부담금 부과 정책과 연계될 수 있다. 정부가 ICT 기업과 데이터센터에 대해 연간 온실가스 감축 목표를 설정하고, 이를 초과할 경우 일정 금액의 부담금을 부과하는 방식이 도입된다면, 인증제도는 그 반대편에서 감축 목표를 초과 달성한 기업을 대상으로 인센티브를 제공하는 구조로 기능할 수 있다. 이로써 제재와 보상이 균형을 이루는 체계적 감축 관리 모델이 형성될 수 있다.

결국 녹색인증제도의 확장과 체계적 운영은 온실가스 감축 정책의 실효성을 높이는 동시에, 친환경 기술 도입을 촉진하고 기업의 지속가능성을 평가하는 제도적 기반으로 작용할 수 있다.

나. 인증 기업에 대한 인센티브 제공

녹색인증제도가 실질적인 감축 행동으로 이어지기 위해서는, 인증을 획득한 기업이 그 성과에 상응하는 혜택을 체감할 수 있도록 명확한 인센티브 체계를 마련하는 것이 중요하다. 정부는 인증 기업에 대해 세제 감면, 전력요금 할인, 정책자금 우대, 공공조달 가점 부여 등 다양한 형태의 지원을 검토할 수 있다. 이러한 인센티브는 감축 활동을 단순한 비용이 아닌 투자로 인식하게 하여, 기업의 자발적인 참여를 확대하는 데 기여한다. 예를 들어, 일정 기준 이상의 감축 실적을 달성하여 녹색인증을 받은 데이터센터는 고효율 장비 도입

이나 재생에너지 전환을 위한 설비투자비의 일부를 세액공제 형태로 환급받거나, 에너지 절감분에 비례한 전력요금 감면 혜택을 받을 수 있다. 또한 인증을 받은 ICT 기업은 정부 연구개발(R&D) 과제 선정 시 우선권을 부여받거나, 친환경 기술 관련 규제 특례(규제 샌드박스) 적용 대상에 포함될 수 있다. 이러한 방식은 기술혁신과 제도 지원을 연계하여 산업 전반의 감축 역량을 강화하는 효과를 낳는다.

이와 같은 인센티브 체계는 규제 중심의 감축 정책에서 벗어나, 시장의 자율성과 경쟁을 활용하는 지속가능한 저탄소 정책 모델로 발전할 수 있다. 결과적으로 인증과 인센티브의 조합은 기업의 감축 노력을 실질적인 경제적 가치로 전환시키는 제도적 기반이 된다.

4. 시장기반 규제를 통한 배출 감축 유도

디지털 탄소중립 목표를 효과적으로 달성하기 위해, 정부는 기업들의 자발적인 온실가스 감축 노력을 촉진하는 시장기반 규제 도입을 추진하고 있다. 이는 단순한 행정적 명령이 아니라, 경제적 유인을 통해 산업계 스스로 감축 기술을 개발하고 투자하도록 유도하는 방안이다.

가. ICT 부문 온실가스 감축 목표 설정 및 초과배출 부담금 부과

정부는 ICT 기업과 데이터센터에 대해 온실가스 감축 목표를 설정하고, 목표를 초과해 배출할 경우 부담금을 부과하는 방식을 도입할 수 있다. 예를 들어, 데이터센터에 연간 탄소 배출 허용량을 부여하고, 이를 초과하는 배출량에 대해 톤당 일정 금액의 부담금을 부과한다. 이러한 부담금은 기업에 직접적인 재정적 부담으로 작용하여, 에너지 효율 개선이나 저전력 기술 도입에 대한 투자를 유도하는 강력한 동기가 된다. 부담금으로 조성된 재원은 다시 저탄소 기술 연구개발이나 시설 개선 지원에 활용되어 선순환 구조를 만들어낼 수 있다. 이 제도는 특히 탄소 감축에 소극적인 기업에 경제적 불이익을 줌으로써 시장 전반의 감축 노력을 평준화하는 데 기여할 것이다.

나. 배출권거래제(ETS) 편입 검토

더 나아가, ICT 부문을 배출권거래제(ETS)에 편입하는 방안도 검토할 수 있다. 배출권거래제는 기업에 온실가스 배출 허용량을 할당하고, 남거나 부족한 배출권을 시장에서 거래하게 하는 제도다. 이 제도를 도입하면, 감축 기술에 적극적으로 투자하여 배출량을 줄인

기업은 남은 배출권을 팔아 수익을 창출할 수 있는 반면, 감축에 소극적인 기업은 배출권을 구매해야 하는 비용 부담이 발생한다. 이는 ICT 기업 간의 자발적인 감축 경쟁을 유도하여, 산업 전반의 저탄소 전환을 가속화하는 효과를 가져온다. 배출권거래제 편입은 ICT 부문의 탄소 감축에 대한 경제적 책임을 명확히 하고, 시장의 힘을 활용하여 가장 효율적인 감축 수단을 찾아내는 데 기여할 것이다. 이와 같은 시장기반 규제는 장기적으로 ICT 산업의 지속가능한 성장을 위한 필수적인 제도적 기반이 될 수 있다.

5. 탄소중립 목표 이행체계 구축

가. 법적 기반 마련 및 역할 구체화

탄소중립 사회로의 전환을 추진하기 위해서는 정부와 산업계가 공통된 목표 아래에서 움직일 수 있는 법적·제도적 틀을 갖추는 것이 필요하며, 현재 정부는 탄소중립 녹색성장 기본법을 통해 ICT 부문의 탄소 감축 노력을 법적 의무로 규정하고 있다.[132] 이 법은 ICT 산업을 포함한 전 산업 부문이 온실가스 감축을 법적 의무로 이행해야 함을 명확히 하고 있으며, 이를 통해 기존의 자율적 참여 중심의 구조에서 법적 책무를 수반하는 체계로 전환하고 있다. ICT 부문은 국가 전체 에너지 소비와 탄소 배출에서 점차 비중이 커지고 있는 핵심 산업으로, 법적 의무화의 대상이자 이행의 주체로서 명확한 역할이 부여된다. 법에 근거하여 정부는 ICT 부문별 온실가스 감축 목표를 설정하고, 기업 규모·사업 특성·에너지 사용 구조 등을 고려해 감축 의무를 차등화할 수 있다. 또한 감축 목표를 이행하지 못한 기업에 대해서는 행정적 제재나 경제적 부담을 부과할 수 있는 근거를 마련함으로써, 제도의 실효성을 확보할 수 있다.

결국 탄소중립 기본법에 근거한 ICT 부문의 법적 책무화는 단순한 규제 강화를 넘어, 산업계가 장기적 감축 목표를 계획적으로 이행할 수 있도록 하는 제도적 틀을 제공한다. 이를 통해 정부는 감축 목표 달성의 관리·감독 기능을 수행하고, 기업은 명확한 법적 기준 아래에서 기술 혁신과 투자를 추진할 수 있는 안정적 여건을 확보하게 된다.

나. ICT 부문 NDC 달성 기여도 제고

국가 온실가스 감축목표(NDC, Nationally Determined Contribution)는 2030년까지 국가

132) 기후위기 대응을 위한 탄소중립·녹색성장 기본법 - 국가법령정보센터

전체 온실가스 배출량을 일정 수준으로 줄이기 위한 중장기 계획으로, 모든 산업 부문이 각자의 감축 책임을 이행해야 하는 공동의 목표를 담고 있다.[133] ICT 부문은 높은 에너지 집약도를 지닌 동시에, 기술 혁신을 통해 다른 산업의 효율을 향상시킬 수 있는 이중적 특성을 지니고 있어, NDC 달성 과정에서 전략적 핵심 분야로 주목받고 있다.

ICT 기술은 에너지 효율 개선과 온실가스 감축을 동시에 달성할 수 있는 잠재력을 가지고 있다. 인공지능(AI), 사물인터넷(IoT), 클라우드, 엣지컴퓨팅 등은 산업·교통·건물 등 다양한 부문에서 에너지 사용을 최적화하는 데 활용될 수 있으며, 이러한 기술의 보급은 산업 전반의 탄소 배출을 간접적으로 감축하는 효과를 낳는다. 예를 들어, AI 기반 예측 유지보수 시스템은 제조 공정의 불필요한 에너지 소모를 줄이고, 스마트그리드 기술은 전력 수요를 실시간으로 조정함으로써 전력 생산 과정에서의 탄소 배출을 감소시킨다.

또한 ICT 기술은 타 산업의 디지털 전환을 가속화함으로써 간접적인 감축 효과를 창출할 수 있다. 예를 들어, 산업 설비의 디지털 트윈(Digital Twin) 기술을 활용한 공정 시뮬레이션은 생산 단계에서 불필요한 자원 투입을 최소화할 수 있으며, 물류·운송 부문에서는 실시간 경로 최적화를 통해 연료 소비와 배출량을 줄이는 효과를 가져온다.

정부는 이러한 ICT 부문의 직접적·간접적 감축 잠재력을 체계적으로 반영하기 위해, NDC 이행계획에 ICT 세부전략을 포함시키고, 매년 그 성과를 평가·보완하는 관리체계를 운영할 필요가 있다. 부문별 감축 실적이 명확히 기록되고 검증될 때, ICT 기술의 기여도는 정량적으로 평가될 수 있으며, 향후 감축정책 및 기술투자 방향의 조정에도 활용될 수 있다.

133) 2030 국가 온실가스 감축목표(NDC) 상향안

제 3 절 공공부문 선도 및 지원체계

1. 공공부문 저탄소 디지털 기술 선도 도입

정부는 디지털 탄소중립으로의 전환을 가속화하기 위해 공공부문이 선도적인 역할을 수행하도록 유도하고 있다. 공공기관이 친환경 데이터센터와 클라우드 등 저탄소 디지털 기술을 선제적으로 도입함으로써, 민간 부문의 참여를 촉진하고 혁신적인 모범 사례를 창출하는 것이 목표다.

가. 공공기관의 친환경 디지털 기술 도입 의무화

정부는 공공기관이 신규 또는 기존 정보 시스템을 구축하거나 운영할 때, 친환경 데이터센터와 클라우드 서비스를 우선적으로 활용하도록 의무화하고 있다.[134] 이는 공공 부문이 막대한 데이터 처리량을 바탕으로 저전력 반도체, 고효율 냉각 기술 등이 적용된 친환경 인프라의 초기 수요를 창출하는 효과를 가져온다. 구체적으로, 행정안전부는 2021년 '행정·공공기관 정보자원 클라우드 전환·통합 추진계획'을 발표하고, 2025년까지 행정기관과 공공기관이 운영 중인 모든 정보시스템 약 1만 9천 개를 클라우드로 전면 전환할 계획을 제시했다.[135] 이 중 2025년까지 공공부문 정보시스템의 약 46%가 보안 및 안정성을 인증받은 민간 클라우드를 이용할 것으로 예상된다.데이터센터의 에너지 효율을 측정하는 핵심 지표인 PUE(전력효율지수)와 관련하여, 우리나라 공공기관 데이터센터의 평균 PUE는 2.0 이상으로 알려져 있으며, 이는 전력의 절반 이상이 IT 장비가 아닌 냉각과 부대시설에 소비되고 있음을 의미한다.[136] 한국데이터센터연합회에서 시행하는 '그린데이터센터인증제'에서는 PUE 1.75 이하의 데이터센터에 인증을 부여하고 있다. 행정 서비스용 서버를 구축할 때는 에너지효율 등급이 높은 서버만을 도입하도록 규정하고 있으며, 클라우드 서비스 계약 시에는 재생에너지 사용 비율이 높은 사업자에게 가점을 부여하는 방식을 채택하고 있다. 이러한 정책은 공공기관이 비용 효율성만을 고려하는 것을 넘어, 환경적 책임

134) 모두싸인, '공공기관의 디지털 전환 촉진, 클라우드 우선 검토 법령', 2025
135) 대한민국 정책브리핑, 「행정·공공기관 정보자원 클라우드 전환·통합 추진계획」, 2021
136) 아카이브H, 「국내 데이터센터 PUE 현황과 그린데이터센터인증제」, 2024

까지 고려하도록 유도한다. 실제로 전 세계 대규모 데이터센터의 평균 PUE가 약 1.57 수준이며, 선도 기업인 구글의 경우 2021년 2분기 기준 PUE 1.1을 달성했다[137].

나. 단계적 전환 로드맵과 클라우드 네이티브 전환

공공부문의 저탄소 디지털 전환이 체계적으로 이루어지도록 정부는 단계별 로드맵을 수립하고 있다. 행정안전부는 2021년부터 본격적으로 클라우드 전환을 추진하여 전체 1만 9천여 개 정보시스템 중 6천여 개를 전환 완료했다.[138] 그러나 기존 클라우드 전환 사업에서 사용자가 폭증할 경우 접속 지연, 서비스 장애 등으로 대민 서비스 이용에 문제가 생기는 등 시스템 안정성의 한계가 지적되었다. 이에 정부는 단순히 온프레미스를 클라우드로 바꾸는 것을 넘어 '클라우드 네이티브' 전환을 추진하고 있다. 2024년부터 2030년까지 7개년 계획으로 범정부 EA 포털 내 모든 시스템의 클라우드 네이티브 전환을 목표로 하며, 각 시스템의 서버 내용연수 기한인 7년 단위에 맞춰 전환을 추진한다.[139] 2024년에는 '국가대중교통정보' 등 21개 행정·공공기관 정보시스템을 선정해 클라우드 네이티브 전환을 우선 추진하고 있으며, 2026년까지 70% 이상의 공공 클라우드 네이티브 전환을 계획하고 있다. 정부는 표준 가이드라인을 마련하고 시범·중점 사업을 추진하며, 매년 진행 상황을 점검하고 필요시 목표를 조정하는 유연한 관리 체계를 운영하고 있다.

다. 모범 사례 발굴 및 확산

공공기관의 선제적 도입은 민간 부문에 신뢰할 수 있는 성공 사례를 제공하는 중요한 역할을 한다. 정부는 친환경 디지털 기술을 성공적으로 도입하여 탄소 감축 효과를 거둔 공공기관의 사례를 적극적으로 발굴하고 공유한다. 예를 들어, 특정 공공기관이 AI 기반 에너지 관리 시스템을 도입해 전력 사용량을 20% 절감했다면, 그 성공 노하우와 데이터를 민간 기업에 공개하는 것이다. 이러한 모범 사례는 민간 기업이 저탄소 기술 도입의 실질적인 이점을 확인하고, 투자에 대한 불확실성을 해소하는 데 큰 도움이 된다. 궁극적으로, 공공 부문의 선도적인 역할은 저탄소 기술 시장을 활성화하고, 디지털 전환을 통한 탄소 중립 사회로의 이행을 가속화하는 기반이 될 것이다. 국내 클라우드 시장 규모는 2024년

137) KT Enterprise, 「글로벌 데이터센터 PUE 동향」, 2021
138) OpenMaru, 「공공부문 클라우드 전환 현황」, 2024
139) 컴퓨터월드, 「범정부 클라우드 네이티브 전환 계획」, 2024

14조 6천억 원에서 2028년 24조 5천억 원으로 성장할 것으로 전망되며[140], 공공부문의 적극적인 참여가 이러한 시장 확대를 견인할 것으로 기대된다.

2. 그린 ICT 조달 기준 마련 및 우선 구매

공공부문이 저탄소 디지털 전환을 선도하기 위한 중요한 방안 중 하나는 그린 ICT 조달 기준을 마련하고, 에너지 효율이 우수한 제품과 서비스를 우선적으로 구매하는 것이다. 이는 공공 조달 시장이라는 거대한 수요처를 활용해 민간 기업의 저탄소 기술 개발을 촉진하고, 친환경 제품 시장을 확대하는 강력한 정책 수단이 된다.

가. 그린 ICT 조달 기준 제정 및 우선 구매제도 도입

정부는 국가 및 지방자치단체가 ICT 장비를 구매하거나 서비스를 계약할 때 적용할 수 있는 명확한 기준을 마련하고 있다.[141] 이 기준에는 서버, 스토리지, 네트워크 장비 등의 에너지 효율 등급이나, 클라우드 서비스의 전력효율지수를 포함한다. 예를 들어, 조달청을 통해 공공기관에 납품되는 모든 ICT 장비에 대해 최저 에너지효율 기준을 설정하고, 이를 충족하는 제품만 구매 목록에 올리는 방식이다. 구체적으로 서버의 경우 에너지효율 1등급 이상, 네트워크 장비의 경우 특정 전력 소비량 이하 제품만 조달 가능하도록 제한하는 방안이 검토되고 있다.또한 '그린 ICT' 인증을 받은 제품이나 서비스에 대해서는 입찰 평가 시 가점을 부여하거나 수의계약 우선권을 제공하는 등 우선 구매제도를 도입하여, 시장의 자발적인 친환경 기술 경쟁을 유도한다. 공공기관이 클라우드 서비스를 선정할 때도 CSAP(클라우드 보안인증제) 인증과 함께 데이터센터의 PUE, 재생에너지 사용 비율, 탄소 배출량 공개 여부 등을 평가 항목에 포함시켜 환경 성과가 우수한 사업자를 우대하는 방식을 적용하고 있다. 예를 들어, 클라우드 사업자가 데이터센터 전력의 일정 비율 이상을 재생에너지로 조달할 경우 입찰 평가에서 가점을 부여하거나, PUE 1.5 이하를 달성한 데이터센터를 운영하는 사업자에게 우선권을 제공하는 방식이다.이러한 조달 기준은 단순히 제품과 서비스의 구매 단계에서만 적용되는 것이 아니라, 운영 단계에서의 에너지 성과까지 고려하는 방향으로 확대되고 있다. 즉, 계약 이후에도 실제 운영 과정에서 약속된 에너

140) 서울경제, 「국내 클라우드 시장 전망」, 2024

141) 과학기술정보통신부, '소프트웨어(SW)분야 표준계약서(6종) 활용 안내', 2020

지 효율을 달성하는지 모니터링하고, 목표 미달 시 패널티를 부과하거나 차기 계약에서 불이익을 주는 방식으로 실효성을 높이고 있다.

나. 초기 시장 형성 및 기술 혁신 가속화

공공 조달 시장은 그 규모가 매우 크기 때문에, 그린 ICT 조달 기준 마련은 저탄소 기술을 보유한 기업들에게 안정적인 초기 시장을 제공하는 효과를 낳는다. 신기술을 개발한 스타트업이나 중소기업은 공공 조달 시장을 통해 기술의 상용화 가능성을 입증하고, 안정적인 수익을 창출하여 재투자의 기반을 마련할 수 있다. 이는 곧 기업의 기술 혁신 노력을 가속화하는 선순환 구조로 이어진다. 특히 저전력 반도체나 에너지 절감 소프트웨어와 같은 혁신 기술의 경우, 초기 시장 진입이 어렵고 투자 회수 기간이 길어 민간 기업의 적극적인 투자가 이루어지기 어려운 측면이 있다. 그러나 공공부문이 이러한 제품의 주요 수요처가 되어 안정적인 물량을 보장하면, 기업은 기술 개발에 대한 확신을 가지고 연구개발에 더욱 적극적으로 투자할 수 있다. 또한 공공 조달을 통해 검증된 제품은 민간 시장에서도 신뢰도가 높아져 시장 확대가 용이해진다. 또한 공공부문이 친환경 ICT 제품의 주요 수요처가 되면, 대기업들도 시장의 요구에 맞춰 저전력·고효율 제품 개발에 적극적으로 나서게 되어 산업 전반의 기술 수준을 높이는 데 기여한다. 더 나아가, 공공 조달을 통해 형성된 초기 시장은 규모의 경제를 실현하여 제품 단가를 낮추는 효과도 가져온다. 이는 민간 기업들이 동일한 친환경 제품을 보다 합리적인 가격에 구매할 수 있는 환경을 조성하여, 저탄소 기술의 전반적인 확산을 촉진하는 데 기여한다.

3. 제도적 지원을 통해 신기술 현장 적용 시 애로사항 해소 및 확산 촉진

저탄소 디지털 기술이 빠르게 현장에 적용되기 위해서는 기존의 규제와 제약으로 인한 걸림돌을 제거하는 것이 중요하다. 정부는 규제 샌드박스와 같은 제도적 지원을 통해 신기술의 현장 적용 시 발생하는 애로사항을 해소하고, 기술 확산을 촉진하는 정책을 추진하고 있다.

가. 규제 샌드박스 활용으로 신기술 실증 기회 확대

정부는 ICT 분야의 혁신 기술이 현행 규제에 막혀 시장에 진입하지 못하는 문제를 해결하기 위해 규제 샌드박스를 적극적으로 활용한다. 규제 샌드박스란 사업자가 신기술을 활

용한 새로운 제품과 서비스를 일정 조건(기간·장소·규모 제한) 하에서 시장에 우선 출시해 시험·검증할 수 있도록 현행 규제의 전부나 일부를 적용하지 않는 제도를 말한다.[142] ICT 규제 샌드박스는 2019년 1월부터 시행되었으며, ICT 기술이 결합된 서비스를 대상으로 하여 다양한 신기술·서비스의 시장출시 및 테스트가 가능하도록 일정 조건 하에 기존 규제를 면제하거나 유예하는 제도다.[143] 이 제도는 실증특례를 중심으로 임시허가와 신속확인 제도를 연계하여 운영하고 있다. 예를 들어, 데이터센터의 에너지 효율을 획기적으로 높일 수 있는 새로운 냉각 기술이 기존의 안전 규제나 시설 기준에 부합하지 않더라도, 특례를 부여하여 실제 환경에서 기술을 테스트하고 실증할 수 있는 기회를 제공하는 것이다. 이 과정에서 정부는 기술의 안전성과 효과를 면밀히 검증하고, 문제가 없을 경우 관련 규제를 완화하거나 재정비하는 과정을 거친다. 이러한 방식은 기업이 불필요한 행정 절차와 규제 리스크를 줄이고, 기술 개발에 집중할 수 있는 환경을 조성해준다. 규제 샌드박스는 신속처리, 임시허가, 실증특례의 세 가지 주요 방식으로 운영된다.[144] 신속처리는 규제 존재 여부·내용 문의 시 30일 이내 미회신 시 무규제로 간주하는 제도이며, 임시허가는 법령이 모호·불합리할 시 기존 규제 적용 없이 조기 출시를 허용하는 것이고, 실증특례는 법령이 모호·불합리·금지일 시 기존 규제 적용 없이 테스트를 허용하는 것이다. 또한 신기술 적용으로 인해 발생할 수도 있는 국민의 생명·안전 우려, 환경 훼손 등을 사전에 방지하기 위해 안전장치를 마련하고 있어, 혁신과 안전성을 동시에 추구하는 균형잡힌 제도로 운영되고 있다.

나. 신속한 제도 개선 및 정보 공유로 확산 가속화

규제 샌드박스 실증을 통해 성공 가능성이 입증된 기술에 대해서는 관련 규제를 신속하게 개선하여 기술이 시장에 안착할 수 있도록 돕는다. 정부는 규제 개선 프로세스를 간소화하고, 법령 개정을 통해 새로운 기술이 광범위하게 적용될 수 있는 법적 기반을 마련한다. 특히 2019년 시행 100일을 계기로 법령정비 요청제도를 신설하고, 동일·유사 사례 처리 패스트트랙을 도입하는 등 제도를 지속적으로 보완하고 있다.[145] 또한 시행 6개월 계기

142) 규제정보포털, 「규제샌드박스 소개」, 2025
143) 찾기쉬운 생활법령정보, 「ICT 규제샌드박스란?」, 2025
144) 정보통신산업진흥원, 「ICT 규제샌드박스 운영」, 2025

에는 특허지원 등 사업화 지원대책을 마련하여 규제 샌드박스를 통과한 기업들이 실질적인 사업화로 이어질 수 있도록 지원하고 있다. 또한 규제 샌드박스 통과 기술에 대한 정보를 적극적으로 공유하고, 공공 조달 시장 진출을 지원하는 등의 후속 조치를 통해 기술 확산을 가속화한다. 정부는 정보통신산업진흥원을 ICT 융합 분야의 접수·컨설팅 전담기관으로 지정하여, 기술력은 있으나 자본력과 행정력이 부족한 중소기업과 스타트업의 규제 샌드박스 신청을 체계적으로 지원하고 있다. 규제 샌드박스를 통과한 기업들의 성공 사례는 데이터베이스화되어 다른 기업들이 참고할 수 있도록 공개되며, 이는 유사한 기술을 개발하는 기업들에게 중요한 가이드라인을 제공한다. 이처럼 규제 샌드박스는 단순한 특례 제도를 넘어, 저탄소 디지털 기술 혁신을 위한 규제 환경을 유연하게 만들고, 기술 상용화의 속도를 높이는 중요한 제도적 지원 체계가 된다.

4. 전문인력 양성 · 컨설팅을 통한 민간 역량 강화

디지털 저탄소 전환의 성공적인 이행을 위해서는 기술 개발뿐만 아니라, 이를 현장에 적용하고 관리할 수 있는 전문 인력의 확보와 기업의 실행 역량 강화가 필수적이다. 정부는 이러한 요구에 부응하기 위해 전문 인력 양성 및 컨설팅 지원 체계를 구축하고 있다.

가. 전문인력 양성 프로그램 운영

정부는 저탄소 디지털 기술 분야의 인력난을 해소하기 위해 다양한 교육 및 양성 프로그램을 운영하고 있다[146]. 이는 대학 및 연구기관과의 협력을 통해 AI 기반 에너지 관리, 저전력 반도체 설계, 그린 클라우드 운영 등 특화된 교육 과정을 개설하는 방식이다. 또한 재직자를 대상으로 한 직무 전환 교육이나 단기 심화 과정도 운영하여, 기존 산업 인력이 저탄소 디지털 분야로 원활하게 이동할 수 있도록 지원한다. 이러한 프로그램을 통해 배출된 전문 인력은 기업의 저탄소 기술 도입과 운용에 핵심적인 역할을 수행하며, 기술의 실질적인 효과를 극대화하는 데 기여한다.

---

145) 규제정보포털, 「규제샌드박스 소개」, 2025

146) 기업마당, 『2024년 중소기업 혁신바우처 사업 지원계획 공고(탄소중립)』, 2024

나. 중소·중견기업 대상 컨설팅 지원

저탄소 기술 도입에 대한 정보와 전문성이 부족한 중소·중견기업을 위해 맞춤형 컨설팅 지원도 제공한다[147]. 정부는 전문가 그룹을 구성하여 기업의 에너지 소비 현황을 진단하고, 가장 효과적인 저탄소 디지털 솔루션을 제안하는 서비스를 운영한다. 예를 들어, 데이터센터를 운영하는 중소기업에 전문가가 방문하여 PUE(전력효율지수)를 측정하고, 저비용으로 에너지 효율을 높일 수 있는 구체적인 개선 방안을 제시하는 방식이다. 이러한 컨설팅은 기업의 의사결정 과정을 돕고, 불확실성을 줄여줌으로써 저탄소 기술 도입을 촉진하는 중요한 역할을 한다. 또한 에너지절약을 통한 기업의 경쟁력 강화를 도모하기 위하여 에너지절약형 시설 투자 시 소요자금의 일부를 장기 저리융자 및 이차보전으로 지원하고 있다.[148] 중소기업은 90%, 중견기업 및 공공기관은 70%의 지원 비율로 사업장당 최소 1천만원부터 최대 300억 원까지 융자 지원을 받을 수 있다.

다. 산학협력 강화 및 실무형 인재 육성

정부는 저탄소 디지털 기술 분야의 실무형 인재를 양성하기 위해 산학협력을 강화하고 있다. 대학과 기업이 공동으로 교육 과정을 설계하고, 현장 실습 및 인턴십 프로그램을 통해 학생들이 실제 업무 환경을 경험할 수 있도록 지원한다. 이러한 산학협력 프로그램은 기업이 필요로 하는 실무 능력을 갖춘 인력을 배출하는 동시에, 학생들에게는 취업 기회를 제공하는 선순환 구조를 만들어낸다. 또한 기업은 우수 인력을 조기에 확보할 수 있어 인력난 해소에 도움이 되며, 정부는 이러한 기업에 인건비 일부를 지원하는 등 다양한 방식으로 일자리 창출을 촉진하고 있다. 이러한 노력은 청년들에게 새로운 기술 분야에서의 도전 기회를 제공하는 동시에 미래 성장 동력을 확보하는 데 기여한다. 또한 컨설팅 결과를 바탕으로 한 정부 지원사업과 연계하여 기업의 재정적 부담까지 덜어주는 종합적인 지원 체계를 구축하고 있다. 궁극적으로 전문인력 양성과 컨설팅 지원은 저탄소 디지털 전환을 위한 기업의 실행 역량을 강화하고, 기술의 현장 적용을 가속화하는 핵심적인 정책 수단이 될 것이다.

147) 제타플랜인베스트, 『탄소중립 경영혁신 바우처사업』, 2024

148) 한국에너지공단, 「에너지절약시설투자 자금지원 및 세제지원 제도」, 2025

5. 탄소중립 녹색성장 기본법에 따른 공공부문 온실가스 감축 목표 달성

공공부문이 탄소중립으로 전환하는 과정에서 디지털 기술은 핵심적인 역할을 한다. 정부는 탄소중립 녹색성장 기본법에 따라 공공부문이 온실가스 감축 목표를 달성할 수 있도록 디지털 기술을 적극적으로 활용하는 정책을 추진하고 있다.[149]

가. 공공기관의 온실가스 배출량 보고 및 관리 의무

공공기관은 국가 전체 온실가스 감축 목표(NDC) 달성을 위한 중요한 주체이며, 법에 따라 매년 온실가스 배출량을 보고하고 감축 목표를 이행해야 한다. 온실가스·에너지 목표 관리제는 기준량 이상 온실가스를 배출하는 업체에 감축목표를 설정하고 달성하도록 관리하는 방안으로, 일정 수준 이상의 온실가스를 배출하고 에너지를 소비하는 업체 및 사업장을 관리 업체로 지정하여 온실가스 감축목표, 에너지 절약 목표를 설정하고 관리한다. 예를 들어, 공공기관이 기존의 노후화된 전산실을 고효율 친환경 데이터센터로 통합하거나, 자체 서버 대신 클라우드 시스템을 도입하면 전력 소비를 획기적으로 줄일 수 있다. 또한 AI 기반 에너지 관리 시스템을 도입하여 건물의 냉난방 및 전력 사용량을 최적화함으로써 에너지 낭비를 최소화할 수 있다. 이러한 디지털 기술 도입은 단순히 에너지 비용을 절감하는 것을 넘어, 공공부문의 온실가스 감축 목표 달성에 직접적으로 기여한다. 결론적으로, 디지털 기술을 활용한 공공부문의 저탄소 전환은 법적 의무를 이행하는 동시에, 민간 부문에 모범 사례를 제시하여 국가 전체의 탄소중립 달성을 가속화하는 중요한 동력이 된다.

149) 기후위기 대응을 위한 탄소중립·녹색성장 기본법 - 국가법령정보센터

제 4 절 이행 거버넌스 구축

1. 디지털 · 탄소중립 연계 범부처 협업체계 구축

디지털 혁신과 탄소중립 과제를 효과적으로 연계하기 위해서는 각 주체의 역할을 명확히 하고, 정기적인 협의를 통해 진행 상황을 점검하는 체계가 필수적이다. 정부는 이러한 목표를 달성하기 위해 범부처 차원의 이행 거버넌스를 구축하고 있다.

가. 범부처 협의체 및 전담반 운영

정부는 과학기술정보통신부, 환경부, 산업통상자원부 등 관련 부처가 모두 참여하는 '디지털 탄소중립 협의회'와 같은 범부처 협의체를 구성해 운영한다. 이 협의체를 통해 부처 간 정책 목표와 추진 방향을 공유하고, 중복 투자를 방지하며, 정책의 시너지를 극대화한다. 예를 들어, 과학기술정보통신부는 저전력 AI 기술 개발을 담당하고, 산업통상자원부는 해당 기술을 산업 현장에 적용하는 실증 사업을 주관하며, 환경부는 이로 인한 온실가스 감축 효과를 측정하는 방식으로 역할 분담을 명확히 할 수 있다. 이처럼 범부처 협의체는 각 부처의 전문성을 결합하여 효율적인 정책 집행을 가능하게 한다.

나. 법적·제도적 연계 강화

범부처 협업 체계는 단순히 정책 조율에 그치지 않고, 법적·제도적 연계까지 강화한다. 탄소중립 녹색성장 기본법을 기반으로 각 부처가 추진하는 디지털 전환 관련 정책에 탄소중립 목표를 반영하도록 의무화한다. 예를 들어, 디지털 기술을 활용한 스마트시티 구축 사업을 추진할 때, 에너지 효율성 향상과 온실가스 감축을 핵심 평가 지표에 포함시키는 방식이다. 정부는 2020년부터 15개 부처가 참여하는 범정부협의체를 운영하며 탄소중립 추진전략을 마련해왔다[150]. 이러한 협력체계는 국무조정실, 환경부, 기획재정부, 과학기술정보통신부, 산업통상자원부, 외교부, 행정안전부, 농림축산식품부, 국토교통부, 해양수산부, 고용노동부, 금융위원회, 기상청, 산림청, 농촌진흥청 등이 포함되어 전 정부 차원의 협력을 이루고 있다. 이러한 제도적 연계는 각 부처의 개별 정책이 궁극적으로 국가의 탄소중립 목표에 기여하도록 유도하며, 디지털 혁신과 탄소중립이 상호 보완적인 관계 속에

150) NetZero Expo,「우리나라의 탄소중립」, 2025

서 발전할 수 있는 기반을 마련한다. 또한 범부처 협력체계를 통해 정량·객관적 이행점검, 상시 관리시스템 구축을 추진하여 정책의 실효성을 지속적으로 모니터링하고 개선할 수 있는 체계를 갖추고 있다.

2. 관계부처 역할 명확화 · 정례 협의체 운영

가. 과제별 역할 분담 및 책임 명확화

디지털 탄소중립 과제는 다양한 기술 분야와 산업에 걸쳐 있기 때문에, 각 과제를 가장 잘 수행할 수 있는 부처와 기관의 역할을 명확하게 분담한다. 2023년 11월 확정된 '디지털 전환을 통한 탄소중립 촉진방안'은 전산업 부문의 탄소감축 촉진, 디지털 부문 고효율화·저전력화, 그린 디지털 생태계 구축이라는 3대 전략을 중심으로 6개 과제를 담고 있다[151]. 구체적으로 에너지·수송·건물·농축수산·폐기물·국민생활 등 6개 분야에 AI, IoT, 디지털 트윈 등을 기반으로 하는 그린 디지털 전환 개발·도입은 과학기술정보통신부가 주도하고, 데이터센터·네트워크 등 디지털 인프라의 저전력화를 위한 기술 개발·적용은 과학기술정보통신부와 산업통상자원부가 협력하며, 디지털 부문의 탄소중립 수준 측정·평가 고도화는 환경부가 담당하는 방식이다. 탄소중립·녹색성장기본법 시행으로 환경부와 국토교통부는 공동으로 탄소중립을 공간적으로 구현하는 '탄소중립도시'를 추진하며, 지자체의 탄소중립 모델을 발굴·시행하여 전 국토 확산 기반을 마련한다. 수송부문에서는 산업통상자원부, 환경부, 국토교통부, 해양수산부 등 관계부처가 협업하여 온실가스 감축목표를 설정하고, 대중교통 활성화, 전기·수소차 전환, 철도·항공·선박의 친환경화 등 녹색교통을 활성화한다.

나. 정례 협의체를 통한 진행 점검 및 조정

각 과제의 진행 상황을 지속적으로 관리하고 부처 간 이견을 조율하기 위해 정례 협의체를 운영한다. 과학기술정보통신부는 2021년 12월 주요 ICT 기업 및 관련 협회·기관이 참석한 가운데 '디지털 탄소중립 협의회' 출범식을 개최했다. 이 협의회는 산업계와 함께 디지털 기반의 탄소중립 추진 방안을 논의하고, 기업의 애로사항을 해결하는 등 디지털

151) 2050 탄소중립녹색성장위원회, 「디지털 전환을 통한 탄소중립 촉진방안」, 2023

탄소중립 추진의 구심체 역할을 수행한다. 협의회는 분야별 구체적인 논의를 위해 하위분과인 유무선통신·디지털 플랫폼·ICT 기기제조 등을 설치해 관련 전문가들이 참여하는 논의의 장을 마련했다. 2023년에는 제3차 디지털 탄소중립 협의회를 개최하여 디지털 기술을 활용한 탄소중립 촉진방안에 대해 의견을 나누고 민간의 사례를 살펴보는 시간을 가졌다. 이 과정에서 과학기술정보통신부와 탄소중립녹색성장위원회는 최근 유럽의 탄소국경조정제도 도입 등 선진국들의 탄소중립 실현을 위한 규제가 강화되고 있어 이에 대한 대응역량을 높이기 위해서는 디지털 기술을 활용해 에너지 효율성을 증대하는 것이 중요하다고 강조했다. 또한 산업 전반의 디지털 탄소중립 촉진을 위해 인공지능(AI), 데이터 등 디지털 기술을 활용한 탄소배출 진단·분석·저감 기술을 개발해 에너지 다소비 업종 중심으로 적용하는 방안을 논의했다. '디지털 탄소중립 위원회'와 같은 정기적인 회의체를 통해 각 부처의 담당자들이 모여 추진 상황을 보고하고, 예상치 못한 문제점이나 규제 애로사항을 논의한다. 이 과정에서 발생하는 부처 간의 이견은 상급 기관인 2050 탄소중립녹색성장위원회에서 최종 조율하는 방식으로 의사결정의 신속성을 확보한다. 정례적인 진행점검은 계획 대비 실적을 파악하고, 필요시 정책 방향을 수정·보완하는 유연한 대응을 가능하게 하여, 디지털 혁신과 탄소중립 목표를 동시에 달성하는 데 중요한 역할을 한다.

3. 정책 추진 성과 모니터링 시스템을 통한 현황 파악 및 피드백 강화

디지털 탄소중립 정책의 실효성을 높이려면, 정책이 제대로 이행되고 있는지 지속적으로 확인하고 개선하는 과정이 필수적이다. 정부는 이를 위해 정책 추진 성과를 모니터링하고 평가하는 시스템을 마련하여 이행 현황을 정확히 파악하고 피드백을 강화하는 노력을 기울이고 있다.

가. 객관적 지표를 활용한 모니터링 시스템 구축

정부는 저탄소 디지털 기술 정책의 성과를 객관적으로 측정하기 위한 지표를 개발하고, 이를 바탕으로 한 모니터링 시스템을 구축한다. 예를 들어, 저전력 ICT 장비 도입으로 인한 에너지 절감량, AI 기반 솔루션을 통한 탄소 배출량 감축 효과, 그린 데이터센터의 PUE(전력효율지수) 변화 등을 핵심 지표로 설정한다. 이 시스템을 통해 각 사업별로 설정된 목표 대비 달성률을 정기적으로 파악하고, 데이터 기반의 분석을 통해 정책의 성공 여

부를 판단한다. 모니터링 시스템은 단순히 최종 결과만을 추적하는 것이 아니라, 중간 단계별 성과를 세밀하게 측정하여 조기에 문제점을 발견하고 대응할 수 있도록 설계되었다. 예를 들어, 데이터센터의 에너지 효율화 사업의 경우 분기별로 전력 사용량, 냉각 시스템 효율, 서버 가동률 등을 종합적으로 측정하여 목표 달성 여부를 실시간으로 확인한다. 이러한 모니터링은 단순히 결과만 확인하는 것이 아니라, 정책이 현장에서 어떤 방식으로 작동하고 있는지 실시간으로 파악하는 데 중요한 역할을 한다. 특히 디지털 기술의 빠른 발전 속도를 고려할 때, 정책 효과를 즉각적으로 측정하고 분석하는 것은 정책의 적시성을 확보하는 데 필수적이다.

나. 평가 결과의 피드백 및 정책 환류

모니터링을 통해 얻은 데이터와 분석 결과를 바탕으로 정책에 대한 종합적인 평가를 실시하고, 그 결과를 다시 정책에 반영하는 '환류' 과정을 거친다. 예를 들어, 특정 기술에 대한 세제 지원 효과가 기대에 미치지 못한다고 평가되면 지원 대상을 조정하거나 지원 규모를 확대하는 등 정책을 수정한다. 반대로, 성공적인 성과를 낸 사업은 다른 분야로 확장하거나 예산을 증액하여 확산을 유도한다. 이처럼 모니터링-평가-피드백의 선순환 체계는 정책의 효율성을 높이고, 한정된 자원을 가장 효과적인 곳에 투입하여 디지털 탄소중립이라는 국가적 목표를 보다 신속하게 달성하는 데 기여한다.

4. 민간 전문가 · 시민사회 참여를 통한 투명성 제고

디지털 탄소중립 정책의 성공적인 이행을 위해서는 정부의 노력만으로는 충분하지 않다. 정책의 실효성과 사회적 수용성을 높이기 위해서는 민간 전문가와 시민사회의 적극적인 참여가 필수적이다. 정부는 정책의 전 과정에 걸쳐 다양한 이해관계자들이 참여할 수 있는 채널을 구축하고, 투명한 정보 공개를 통해 정책에 대한 신뢰를 확보하고 있다.

가. 정책 수립 단계부터 민간 전문가 참여 확대

정부는 정책 수립 초기 단계부터 학계, 산업계, 시민단체 등 민간 전문가들이 참여하는 협의체를 운영하고 있다. 이들은 저탄소 디지털 기술의 발전 방향, 산업계의 현장 애로사항, 시민사회의 기대와 우려 등을 정책에 반영하는 역할을 한다. 디지털 탄소중립 협의회와 같은 민관 협력체는 정책의 초안을 검토하고 실현 가능성을 평가하는 데 중요한 기여

를 한다. 예를 들어, 데이터센터 에너지효율 기준을 마련할 때 데이터센터 운영사, 장비 제조사, 에너지 전문가 등이 참여하여 현실적이고 달성 가능한 기준을 설정하는 데 기여한다. 이러한 과정을 통해 정책이 현장의 기술 수준과 경제적 여건을 충분히 고려하여 설계될 수 있다. 또한 학계의 연구자들은 최신 기술 동향과 국제적 모범사례를 소개하며, 정책이 글로벌 표준에 부합하도록 조언한다. 환경단체와 소비자단체는 정책이 환경 보호와 국민의 권익을 제대로 반영하고 있는지 감시하고 의견을 제시한다. 이처럼 다양한 분야의 전문가들이 정책 형성 과정에 참여함으로써, 특정 집단의 이해관계에 치우치지 않는 균형 잡힌 정책을 수립할 수 있다. 이러한 사전 참여는 정책의 전문성과 실현 가능성을 높이고, 특정 집단의 이해관계에 치우치지 않는 공정한 정책을 만드는 데 필수적이다. 더 나아가 정책 수립 단계에서부터 이해관계자들의 의견을 수렴함으로써, 정책 집행 단계에서 발생할 수 있는 저항을 최소화하고 순조로운 이행을 도모할 수 있다.

나. 시민사회와의 소통 채널 구축 및 활성화

정책 집행 과정에서 발생할 수 있는 사회적 갈등을 최소화하고, 정책에 대한 국민적 공감대를 형성하기 위해 정부는 시민사회와의 소통 채널을 구축하고 활성화한다. 디지털 탄소중립과 같은 복잡한 정책은 국민들이 그 필요성과 기대 효과를 명확히 이해할 때 비로소 성공적으로 이행될 수 있다. 온라인 공론장, 정기적인 포럼, 간담회 등을 통해 디지털 탄소중립 정책의 목표, 진행 상황, 성과 등을 투명하게 공개하고 시민들의 의견을 청취한다. 정부는 정책 추진 과정에서 도출된 주요 의사결정 사항과 그 근거를 공개하여 정책의 투명성을 확보한다. 또한 시민들이 정책에 대해 궁금한 점이나 우려사항을 직접 질문하고 답변을 받을 수 있는 쌍방향 소통 채널을 마련한다. 특히 디지털 플랫폼을 활용한 소통은 시간과 공간의 제약 없이 다양한 계층의 시민들이 참여할 수 있다는 장점이 있다. 온라인 설문조사, 가상 타운홀 미팅, 소셜미디어를 통한 의견 수렴 등 다양한 방식으로 시민들의 목소리를 듣고 정책에 반영한다. 이러한 과정에서 수집된 시민 의견은 정책 평가와 개선에 중요한 참고자료로 활용된다.

5. 국가 거버넌스 연계 감축 로드맵 관리

디지털 부문의 탄소 감축 노력이 실효성을 갖추기 위해서는 국가 전체의 탄소중립 목표

와 긴밀하게 연계되어야 한다. 정부는 탄소중립위원회와 같은 국가 거버넌스를 중심으로 디지털 부문의 감축 로드맵을 수립하고 이행을 체계적으로 관리하고 있다. 개별 부문의 노력이 아무리 우수하더라도 국가 차원의 통합적 관리 없이는 목표 달성이 어렵기 때문에, 디지털 부문의 감축 계획을 국가 탄소중립 전략과 유기적으로 연결하는 것이 매우 중요하다.

가. 국가 탄소중립 목표와 연계된 감축 로드맵 수립

디지털 부문의 탄소 감축 로드맵은 국가의 장기 탄소중립 시나리오와 온실가스 감축목표에 부합하도록 수립된다. 이는 ICT 인프라의 전력 효율화, AI 기반의 산업 효율 개선, 데이터센터 녹색화 등 디지털 기술이 기여할 수 있는 잠재력을 객관적으로 평가하고, 구체적인 감축 목표를 설정하는 과정이다. 로드맵 수립 시에는 디지털 기술의 발전 속도, 관련 산업의 성장 전망, 국제적 기술 표준 동향 등을 종합적으로 고려한다. 예를 들어, 데이터센터와 네트워크 시설의 에너지 효율 개선, 저전력 반도체 기술 적용 확대, 클라우드 컴퓨팅의 최적화 등 각 분야별로 달성 가능한 목표를 단계적으로 제시한다. 이러한 목표는 기술적 실현 가능성뿐만 아니라 경제적 타당성과 사회적 수용성까지 검토하여 설정된다. 또한 디지털 기술이 다른 산업 부문의 탄소 감축에 기여하는 간접 효과도 로드맵에 반영한다. 스마트 에너지 관리 시스템, 디지털 트윈을 활용한 제조 공정 최적화, AI 기반 물류 효율화 등을 통해 전 산업에 걸쳐 탄소 배출을 줄일 수 있는 방안을 포함한다. 이처럼 디지털 부문의 직접 감축과 간접 기여를 모두 고려한 통합적 로드맵을 통해 디지털 전환과 탄소중립의 시너지를 극대화한다. 이러한 로드맵은 탄소중립위원회의 심의를 거쳐 확정되며, 국가의 최종 목표 달성에 기여하도록 설계된다. 위원회는 각 부처가 제출한 디지털 부문 감축 계획을 검토하고, 국가 전체의 감축 목표 배분 맥락에서 적정성을 평가한다. 이를 통해 디지털 부문의 감축 노력이 국가 전략의 일부로서 정합성을 유지하고, 다른 부문과의 균형을 이루도록 조정한다.

나. 정기적인 이행 점검 및 평가

수립된 로드맵의 이행 상황을 정기적으로 점검하고 평가하는 체계도 구축된다. 정부는 탄소중립위원회와 협력하여 디지털 부문 감축 로드맵의 진행 상황을 정기적으로 점검하고, 목표 대비 실적을 분석한다. 이러한 점검은 연간 또는 반기별로 실시되며, 각 세부 과제의 추진 현황, 예산 집행률, 기술 개발 진척도 등을 종합적으로 평가한다. 이행 점검 과

정에서는 양적 지표뿐만 아니라 질적 성과도 함께 고려한다. 단순히 감축량만을 평가하는 것이 아니라, 기술 혁신의 파급 효과, 산업 경쟁력 강화, 일자리 창출 등 다양한 측면에서 정책의 성과를 종합적으로 판단한다. 이를 통해 로드맵이 탄소 감축이라는 환경적 목표와 함께 경제적·사회적 가치도 창출하고 있는지 평가할 수 있다. 이 과정에서 기술 개발의 지연이나 시장 변화와 같은 예상치 못한 문제가 발생하면, 로드맵을 유연하게 수정·보완하는 절차를 거친다. 예를 들어, 특정 저전력 기술의 상용화가 예상보다 지연되는 경우 대체 기술을 검토하거나 목표 시점을 조정하고, 반대로 기술 발전이 예상보다 빠르게 진행되면 목표를 상향 조정하는 등 유연한 대응을 한다. 이러한 적응적 관리는 로드맵이 현실과 괴리되지 않고 실효성을 유지하도록 돕는다. 또한 감축 성과를 객관적으로 평가하기 위해 표준화된 모니터링 시스템을 활용하여 투명성을 확보한다. 제3자 검증 기관을 통한 독립적인 평가를 실시하고, 그 결과를 공개하여 국민과 국제사회에 대한 책임성을 강화한다. 평가 결과는 다음 단계의 로드맵 수정 및 새로운 정책 개발의 기초 자료로 활용된다.

제6장 결 론

본 보고서는 디지털 기술이 기후 위기의 주요 원인이자 동시에 가장 강력한 해결의 열쇠라는 시대적 역설을 심층적으로 분석하고, 대한민국이 나아가야 할 '디지털 저탄소 전환'의 통합적 청사진을 제시하고자 했다. 인공지능(AI)과 데이터 기반 경제의 폭발적인 성장은 전례 없는 혁신을 이끌고 있지만, 그 이면에서는 기하급수적으로 증가하는 전력 수요와 탄소 배출이라는 거대한 그림자를 드리우고 있다. 이제 고성능과 저탄소는 더 이상 선택적 가치가 아닌, 미래 디지털 시대의 국가 경쟁력과 지속가능성을 담보하는 필수 생존 조건이 되었다.

본 연구를 통해 도출된 핵심 결론은 성공적인 디지털 저탄소 전환이 '기술 혁신의 가속화'와 '지속가능한 제도적 기반 구축'이라는 두 개의 축을 중심으로 유기적으로 결합될 때 비로소 가능하다는 것이다. 이는 어느 한쪽의 노력만으로는 풀 수 없는 복합적인 문제이며, 기술과 정책, 시장이 함께 움직이는 총체적 시스템 혁신을 요구한다.

첫째, 기술 혁신은 '지능적 효율성'을 향한 근본적 패러다임 전환을 요구한다. 본 보고서에서 제시된 다양한 기술적 해법들은 단순한 에너지 절감을 넘어, 컴퓨팅 아키텍처의 근본적인 재설계를 의미한다. 범용 GPU에 의존하던 '에너지 다소비' 시대에서 벗어나, AI 모델의 특정 연산(예: LLM의 prefill 및 decode 단계)에 최적화된 맞춤형 AI 가속기(ASIC/FPGA)로의 전환이 시급하다. 또한, 연산보다 데이터 이동에 더 많은 에너지를 소모하는 '메모리 병목' 현상을 극복하기 위해 메모리 내에서 직접 연산을 수행하는 메모리 중심 컴퓨팅(PIM) 아키텍처의 도입이 필수적이다. 소프트웨어 차원에서도 모델의 파라미터 정밀도를 낮추는 양자화(Quantization)나 불필요한 연결을 제거하는 프루닝(Pruning) 같은 모델 경량화 기법과, 연산 자체를 동적으로 최적화하는 에너지 지향적 알고리즘을 통해 불필요한 연산을 원천적으로 제거해야 한다. 이러한 기술들은 개별적으로 작동하는 것이 아니라, 소프트웨어의 특성을 하드웨어 설계 단계부터 반영하는 하드웨어-소프트웨어 공동설계(Co-design)라는 통합적 접근을 통해 시너지를 극대화할 때 비로소 그 잠재력을 온전히 발휘할 수 있다.이는 AI 기술 경쟁의 무대가 모델의 크기에서 '와트당 성능

(Performance per Watt)'으로 이동하고 있음을 명확히 보여준다.

둘째, 기술 혁신을 뒷받침하고 시장의 변화를 이끄는 견고한 제도적 기반이 필수적이다. 아무리 뛰어난 저탄소 기술이라도 이를 채택하고 확산시킬 시장과 제도가 없다면 사장될 수밖에 없다. 본 보고서는 정부의 세 가지 핵심 역할을 제시한다. 첫째, 미래 기술의 씨앗을 뿌리는 '전략적 투자자'로서, 저전력 반도체, PIM 등 저탄소 ICT 분야의 R&D 투자를 과감히 확대하고 민관 협력 생태계를 조성해야 한다. 둘째, 공정한 경쟁의 장을 만드는 '현명한 심판'으로서, 데이터센터의 전력효율지수(PUE) 기준을 단계적으로 강화하고, ICT 기업의 온실가스 배출량 공시를 의무화하며, '친환경 ICT 인증제'를 도입하는 등 명확하고 예측 가능한 규제와 표준을 제시하여 시장에 명확한 신호를 주어야 한다. 셋째, 초기 시장을 창출하는 '선도적 구매자'로서, 정부 및 공공기관이 앞장서 '친환경 데이터센터'와 '클라우드 네이티브' 시스템을 의무적으로 도입하여 민간의 기술 채택을 유도하고 모범 사례를 확산시켜야 한다. 이러한 제도적 틀 안에서 기업들은 탄소 감축을 비용이 아닌 새로운 성장의 기회로 인식하게 될 것이다.

결론적으로, 디지털 저탄소 전환은 기술, 산업, 정책이 긴밀하게 연계된 총체적 시스템 혁신이다. 본 보고서가 제시한 다층적 해법 스택—효율적인 하드웨어, 지능적인 소프트웨어, 최적화된 인프라, 그리고 이 모든 것을 뒷받침하는 정책 생태계—은 어느 하나도 소홀히 할 수 없는 상호의존적 관계에 있다. 이제 대한민국은 '2050 탄소중립 시나리오'와 국가 온실가스 감축목표(NDC) 달성을 위해 ICT 부문의 법적 역할을 명확히 하고, 디지털 강국의 위상을 넘어 '지속가능한 디지털 선도국'이라는 새로운 비전을 향해 나아가야 한다. 이를 위해 정부, 산업계, 학계가 모두 참여하는 국가적 거버넌스를 구축하고, 본 보고서에서 제시한 기술적·제도적 방안들을 속도감 있게 실행에 옮겨야 할 때이다. 고성능과 저탄소의 두 마리 토끼를 잡는 것은 더 이상 불가능한 꿈이 아니라, 치열한 글로벌 기술 패권 경쟁 속에서 대한민국의 미래를 확보하기 위한 가장 확실한 전략이다.

참 고 문 헌

[국내 문헌]

산업통상자원부 국가기술표준원(2023),『국가표준시행계획 수립』.

한국지능정보사회진흥원(2025), [Data Brief 2024-5] 디지털 분야의 탄소배출과 대응 노력

Greenpeace(2025), 인공지능(AI) 시대의 그림자: 반도체 제조산업의 전력 소비량과 온실가스 배출량 분석.

한국경제인협회,「미국 청정경쟁법(CCA)의 국내 파급효과 및 시사점」, 2024.

네이버(2024), 데이터센터 각 세종까지 LEED 플래티넘 획득… 친환경 분야도 글로벌 최고 수준 입증 https://www.navercorp.com/media/pressReleasesDetail?seq=31837.

포스코그룹(2021), 포스코, 제철소 '디지털 트윈' PosPLOT 개발…시뮬레이션으로 연원료의 최적 배합 찾는다. https://newsroom.posco.com/kr/%ED%8F%AC%EC%8A%A4%EC%BD%94-%EC%A0%9C%EC%B2%A0%EC%86%8C-%EB%94%94%EC%A7%80%ED%84%B8-%ED%8A%B8%EC%9C%88-posplot-%EA%B0%9C%EB%B0%9C%EC%8B%9C%EB%AE%AC%EB%A0%88%EC%9D%B4%EC%85%98%EC%9C%BC/.

과학기술정보통신부(2023a),『디지털 권리 장전』.

_____(2023b),『2023년 디지털정보격차 실태조사』.

김정언 외(2023),『디지털 대전환 메가트렌드 연구 총괄보고서 Ⅲ』, 정보통신정책연구원. 정책연구 23-30-01.

윤민우(2023), "영국의 사이버 안보전략, 대응정책 그리고 사이버 안보법,"《한국경찰연구》, 22(3), pp. 161-182.

행정안전부(2023), "2023년 행정안전부 정책 방향 업무보고".

정보통신정책연구원(2025), AI 기반 소프트웨어 정의 공장을 활용한 제조 혁신. https://library.kisdi.re.kr/$/10160/contents/4333446?checkinId=3148735&articleId=1819

036.

삼표(2023), 삼표시멘트 ESG 보고서.
https://sampyo.co.kr/media/%EC%82%BC%ED%91%9C%EC%8B%9C%EB%A9%98%ED%8A%B8-2023-esg-%EB%B3%B4%EA%B3%A0%EC%84%9C-%EB%B0%9C%EA%B0%84%EA%B8%B0%ED%9B%84%EB%B3%80%ED%99%94-%EB%8C%80%EC%9D%91-%EA%B0%95%ED%99%94/?.

효성중공업(2021), Power & Motion Magazine Vol.1.
https://www.hyosungheavyindustries.com/resources/files/PMvol1KOR21m.pdf.

_____(2022), Power & Motion Magazine Vol.2.
https://www.hyosungheavyindustries.com/resources/files/PMvol2KOR22.pdf.

_____(2025), https://www.hyosungheavyindustries.com/en/sustainability/environmental/climate-change.

현대자동차그룹 뉴스룸(2025), https://www.hyundai.co.kr/news/CONT0000000000107747.

현대자동차(2025), 2025 현대자동차 지속가능성 보고서.
https://www.hyundai.com/kr/ko/sustain-manage/manage-system/sustainability-report.

정보통신산업진흥원(2017), 스마트 그리드(에너지) 분야 도입사례 분석집.
https://www.nipa.kr/home/2-7-1-1/5054.

KT(2024), KT ESG Report 2024.
https://m.corp.kt.com/archive/ipgrpt/attach/2024/2024_KOR_Archive.pdf.

현대자동차(2025), 2025 현대자동차 지속가능성 보고서.
https://www.hyundai.com/kr/ko/sustain-manage/manage-system/sustainability-report.

한국행정연구원(2021), 스마트모빌리티 서비스 활성화를 위한 OPL 결과보고서.
https://sky.kipa.re.kr/%24/10240/contents/7136504.

인천광역시(2025), 인천시, 스마트 물류체계 혁신으로 미래기술 선도 나선다 !
https://www.incheon.go.kr/IC010205/view?repSeq=DOM_0000000011962178&curPage=123&beginDt=&srchRepTitle=&srchRepContents=&endDt=&srchMainManagerDeptNm=.

포스코(2019), 포스코, ESG, 디지털 물류 발전을 위한 세미나 개최.

https://newsroom.posco.com/kr/%ED%8F%AC%EC%8A%A4%EC%BD%94-esg%E2%80%A2%EB%94%94%EC%A7%80%ED%84%B8-%EB%AC%BC%EB%A5%98-%EB%B0%9C%EC%A0%84%EC%9D%84-%EC%9C%84%ED%95%9C-%EC%84%B8%EB%AF%B8%EB%82%98-%EA%B0%9C%EC%B5%9C/.

김점산, et al. "스마트모빌리티 서비스의 현황 및 발전방안 연구." 정책연구(2020): 1-184.

NAVER(2021), NAVER 2021 TCFD 보고서.

한국산업단지공단(2024), 스마트그린산단 CEMS 우수사례.

한국전력 및 전력그룹사(2024), 지속가능경영보고서.

스마트그린산단 통합플랫폼 - 스마트에너지플랫폼, https://www.kicox.or.kr/kfactory/business/smart-energy.do.

NAVER(2024). 2023 통합보고서(Integrated Report 2023). 금융감독원 전자공시시스템(KIND). https://kind.krx.co.kr/external/2024/06/26/000486/20240625001176/NAVER_Integrated_Report_2023_ver3.pdf.

과학기술정보통신부,「디지털 전환을 통한 탄소중립 촉진방안」, 2023.

대한민국 정부,「디지털 탄소중립 전담반 개최」, 2023.

한국과학기술기획평가원(KISTEP),「AI·클라우드·네트워크 저전력 기술 R&D 로드맵」, 2024.

산업통상자원부,「2022년도 국가표준시행계획」, 2022.

과학기술정보통신부,「디지털 기반 탄소중립 추진」, 2021.

국회도서관 국가전략정보포털,「탄소중립 녹색성장 추진전략」, 2023.

산업통상자원부,「저탄소 산업 전환 민관 협력 로드맵」, 2024.

한국에너지공단,「산업 부문 에너지 관리 실증사업 보고서」, 2023.

국가표준원,「저탄소 ICT 기술 표준화 추진계획」, 2024.

탄소중립위원회,「탄소중립 녹색성장 추진전략」, 2023.

산업통상자원부,「기후테크 산업의 성장모델 창출 및 수출산업 전략 수립」, 2023.

엔비디아,「AI 기반 에너지 절약 솔루션 스타트업 사례」, 2025.

고용노동부,「디지털 혁신과 저탄소 시대 일자리 해법」, 2022.

국가인공지능위원회,「미래사회를 대비한 산학협력 인력양성」, 2020.

2050 탄소중립 시나리오(탄소중립위원회, 2021).
[코스콤리포트] 디지털플랫폼정부 시대와 데이터센터, 2023.
한국에너지공단,「에너지다소비사업장 에너지진단 지침」, 2023.
Cushman & Wakefield,「한국 데이터센터 산업: 전력 수급과 규제 강화 속 성장과 과제」, 2024.
정부 통합 전산센터의 그린화 정책, Korea Science, 2009.
에너지관리공단,「효율등급제도 제도개요」, 2024.
2030 국가 온실가스 감축목표(NDC) 상향안.
모두싸인, '공공기관의 디지털 전환 촉진, 클라우드 우선 검토 법령', 2025.
대한민국 정책브리핑,「행정·공공기관 정보자원 클라우드 전환·통합 추진계획」, 2021.
아카이브H,「국내 데이터센터 PUE 현황과 그린데이터센터인증제」, 2024.
KT Enterprise,「글로벌 데이터센터 PUE 동향」, 2021.
OpenMaru,「공공부문 클라우드 전환 현황」, 2024.
컴퓨터월드,「범정부 클라우드 네이티브 전환 계획」, 2024.
서울경제,「국내 클라우드 시장 전망」, 2024.
과학기술정보통신부, '소프트웨어(SW)분야 표준계약서(6종) 활용 안내', 2020.
규제정보포털,「규제샌드박스 소개」, 2025.
찾기쉬운 생활법령정보,「ICT 규제샌드박스란?」, 2025.
정보통신산업진흥원,「ICT 규제샌드박스 운영」, 2025.
규제정보포털,「규제샌드박스 소개」, 2025.
기업마당,『2024년 중소기업 혁신바우처 사업 지원계획 공고(탄소중립)』, 2024.
제타플랜인베스트,『탄소중립 경영혁신 바우처사업』, 2024.
한국에너지공단,「에너지절약시설투자 자금지원 및 세제지원 제도」, 2025.
기후위기 대응을 위한 탄소중립·녹색성장 기본법 - 국가법령정보센터.
NetZero Expo,「우리나라의 탄소중립」, 2025.
2050 탄소중립녹색성장위원회,「디지털 전환을 통한 탄소중립 촉진방안」, 2023.

[국외 문헌]

Uptime Institute(2024). Direct Liquid cooling Results.

FCST(2024). 5G Energy Saving And Consumption Reduction.

Climate Transparency(2022). Report 2022: Comparing G20 Climate action towards Net Zero.

OECD(2020). Digital Economy Outlook 2020.

ITU and WBA(2025), Greening Digital Companies Report.

IEA(2025), Energy and AI.

Stanford HAI(2023), Artificial Intelligence Index Report 2023.

ITU(2024), Facts and Figures 2024.

Boston Consulting Group(2021), Putting Sustainability at the Top of the Telco Agenda https://web-assets.bcg.com/d1/bb/09fa1876412d8725e80c83d1cf5b/bcg-putting-sustainability-at-the-top-of-the-telco-agenda-jun-2021.pdf.

Apple(2023), Product Environmental Report-Watch Series9 https://www.apple.com/environment/pdf/products/watch/Apple_Watch_Series_9_PER_Sept2023.pdf.

Siemens(2024), World Economic Forum: Siemens factory in Erlangen named Digital Lighthouse Factory. https://press.siemens.com/global/en/pressrelease/siemens-electronics-factory-erlangen-awarded-digital-lighthouse-world-economic-forum.

Siemens(2024), Getting the best out of production with the Digital Twin. https://assets.new.siemens.com/siemens/assets/api/uuid:cda2b14c-23bf-4685-8fdf-adb7fd25e3a7/Whitepaper-Digital-Twin-EN.pdf.

Bosch(2020), The Bosch plant in Feuerbach – where tradition meets high-tech. https://www.bosch-presse.de/pressportal/de/en/bosch-in-feuerbach-tradition-meets-modernity-177029.html.

_____(2025), Nexeed — welcome to the smart factory. https://www.bosch.com/stories/nexeed-smart-factory/

DigitalDefynd(2025), 5 Ways Toyota is Using AI [Case Study]. https://digitaldefynd.com/IQ/toyota-using-ai-case-study/.

Rockwell Automation(2025), How AI Boosts Energy Efficiency in Wastewater Treatment. https://www.rockwellautomation.com/en-us/company/news/the-journal/how-ai-boosts-energy-efficiency-in-wastewater-treatment.html.

Rockwell Automation(2025), Rockwell Automation AI's Power for Sustainable Manufacturing. https://sustainabilitymag.com/sustainability/rockwell-automation-ais-power-for-sustainable-manufacturin.

WIRED(2018), Google's DeepMind trains AI to cut its energy bills by 40%. https://www.wired.com/story/google-deepmind-data-centres-efficiency/.

National Energy System Operator(NESO)(2023), Demand Flexibility Service delivers electricity to power 10 million households. https://www.neso.energy/news/demand-flexibility-service-delivers-electricity-power-10-million-households.

TIME(2024), How AI Is Making Buildings More Energy-Efficient. https://time.com/7201501/ai-buildings-energy-efficiency/.

Clean Energy Ministerial(2022), Global Energy Management System Implementation: Case Study – LG Electronics Gasan R&D Campus. https://www.cleanenergyministerial.org/content/uploads/2022/03/cem-em-casestudy-lg-gasan-korea.pdf.

SCATS(2022), Sydney Coordinated Adaptive Traffic System(SCATS). https://www.transport.nsw.gov.au/system/files/media/documents/2022/SCATS-Core-brochure-Final-web-spreads_0.pdf.

BSR(2016), Looking Under the Hood: ORION Technology Adoption at UPS. https://www.bsr.org/en/case-studies/center-for-technology-and-sustainability-orion-technology-ups.

Ko, Wonsuk, et al. "Implementation of a demand-side management solution for South

Korea's demand response program." applied sciences 10.5(2020): 1751.

Smith, Stephen, et al. "Smart urban signal networks: Initial application of the surtrac adaptive traffic signal control system." Proceedings of the International Conference on Automated Planning and Scheduling. Vol. 23. 2013.

IMDA(2024), Charting green growth pathways at scale for data centres in Singapore

U.S. Department of Energy(2025), Unleashing American Energy: Executive Order Implementation Report.

European commission(2023), Energy Efficiency Directive.

Microsoft(2020), The carbon benefits of cloud computin: A study on the Microsoft Cloud in partnership with WSP.

AWS(2024), How moving onto the AWS cloud reduces carbon emissions.

____(2022), 3개 회사가 AWS를 통해 클라우드를 도입하여 비용을 절약한 방법.

ITU(2010), Recommendation ITU-T L.1470.

Krizhevsky, Alex, et al. "Imagenet classification with deep convolutional neural networks." Advances in neural information processing systems 25(2012).

Brown, Tom, et al. "Language models are few-shot learners." Advances in neural information processing systems 33(2020): 1877-1901.

Achiam, Josh, et al. "Gpt-4 technical report." arXiv preprint arXiv:2303.08774(2023).

Touvron, Hugo, et al. "Llama: Open and efficient foundation language models." arXiv preprint arXiv:2302.13971(2023).

Zhao, Yiren, et al. "Focused quantization for sparse CNNs." Advances in Neural Information Processing Systems 32(2019).

Wu, Hao, et al. "Integer quantization for deep learning inference: Principles and empirical evaluation." arXiv preprint arXiv:2004.09602(2020).

Wang, Kuan, et al. "Haq: Hardware-aware automated quantization with mixed precision." Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2019.

Dong, Zhen, et al. "Hawq: Hessian aware quantization of neural networks with

mixed-precision." Proceedings of the IEEE/CVF international conference on computer vision. 2019.

Nagel, Markus, et al. "Up or down? adaptive rounding for post-training quantization." International conference on machine learning. PMLR, 2020.

Li, Yuhang, et al. "Brecq: Pushing the limit of post-training quantization by block reconstruction." arXiv preprint arXiv:2102.05426(2021).

Faraone, Julian, et al. "Syq: Learning symmetric quantization for efficient deep neural networks." Proceedings of the IEEE conference on computer vision and pattern recognition. 2018.

Choi, Jungwook, et al. "Pact: Parameterized clipping activation for quantized neural networks." arXiv preprint arXiv:1805.06085(2018).

Zhang, Dongqing, et al. "Lq-nets: Learned quantization for highly accurate and compact deep neural networks." Proceedings of the European conference on computer vision(ECCV). 2018.

Zhuang, Bohan, et al. "Towards effective low-bitwidth convolutional neural networks." Proceedings of the IEEE conference on computer vision and pattern recognition. 2018.

Kim, Jangho, et al. "Qkd: Quantization-aware knowledge distillation." arXiv preprint arXiv:1911.12491(2019).

Wang, Kuan, et al. "Haq: Hardware-aware automated quantization with mixed precision." Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2019.

Dong, Zhen, et al. "Hawq-v2: Hessian aware trace-weighted quantization of neural networks." Advances in neural information processing systems 33(2020): 18518-18529.

Lin, Yang, et al. "Fq-vit: Post-training quantization for fully quantized vision transformer." arXiv preprint arXiv:2111.13824(2021).

Yuan, Zhihang, et al. "Ptq4vit: Post-training quantization for vision transformers with twin

uniform quantization." European conference on computer vision. Cham: Springer Nature Switzerland, 2022.

Ma, Yuexiao, et al. "Outlier-aware slicing for post-training quantization in vision transformer." Forty-first International Conference on Machine Learning. 2024.

Li, Zhikai, et al. "Repq-vit: Scale reparameterization for post-training quantization of vision transformers." Proceedings of the IEEE/CVF International Conference on Computer Vision. 2023.

Zhong, Yunshan, et al. "I&s-vit: An inclusive & stable method for pushing the limit of post-training vits quantization." arXiv preprint arXiv:2311.10126(2023).

Yang, Lianwei, et al. "DopQ-ViT: Towards Distribution-Friendly and Outlier-Awa.

Wei, Xiuying, et al. "Qdrop: Randomly dropping quantization for extremely low-bit post-training quantization." arXiv preprint arXiv:2203.05740(2022).

Wu, Zhuguanyu, et al. "APHQ-ViT: Post-Training Quantization with Average Perturbation Hessian Based Reconstruction for Vision Transformers." Proceedings of the Computer Vision and Pattern Recognition Conference. 2025.

Wu, Zhuguanyu, et al. "FIMA-Q: Post-Training Quantization for Vision Transformers by Fisher Information Matrix Approximation." Proceedings of the Computer Vision and Pattern Recognition Conference. 2025.

Tseng, Albert, et al. "Quip#: Even better llm quantization with hadamard incoherence and lattice codebooks." arXiv preprint arXiv:2402.04396(2024).

Chee, Jerry, et al. "Quip: 2-bit quantization of large language models with guarantees." Advances in Neural Information Processing Systems 36(2023): 4396-4429.

Ashkboos, Saleh, et al. "Quarot: Outlier-free 4-bit inference in rotated llms." Advances in Neural Information Processing Systems 37(2024): 100213-100240.

Lin, Ji, et al. "Awq: Activation-aware weight quantization for on-device llm compression and acceleration." Proceedings of machine learning and systems 6(2024): 87-100.

Zheng, Xingyu, et al. "First-Order Error Matters: Accurate Compensation for Quantized Large Language Models." arXiv preprint arXiv:2507.11017(2025).

Liu, Zechun, et al. "Spinquant: Llm quantization with learned rotations." arXiv preprint arXiv:2405.16406(2024).

Lin, Yujun, et al. "Qserve: W4a8kv4 quantization and system co-design for efficient llm serving." arXiv preprint arXiv:2405.04532(2024).

Shao, Wenqi, et al. "Omniquant: Omnidirectionally calibrated quantization for large language models." arXiv preprint arXiv:2308.13137(2023).

Li, Xinlin, et al. "ICQuant: Index Coding enables Low-bit LLM Quantization." arXiv preprint arXiv:2505.00850(2025).

Dosovitskiy, Alexey, et al. "An image is worth 16x16 words: Transformers for image recognition at scale." arXiv preprint arXiv:2010.11929(2020).

Liu, Ze, et al. "Swin transformer: Hierarchical vision transformer using shifted windows." Proceedings of the IEEE/CVF international conference on computer vision. 2021.

Frantar, Elias, et al. "Gptq: Accurate post-training quantization for generative pre-trained transformers." arXiv preprint arXiv:2210.17323(2022).Frantar, Elias, and Dan Alistarh. "Optimal brain compression: A framework for accurate post-training quantization and pruning." Advances in Neural Information Processing Systems 35(2022): 4475-4488.

Hooper, Coleman, et al. "Kvquant: Towards 10 million context length llm inference with kv cache quantization." Advances in Neural Information Processing Systems 37(2024): 1270-1303.

Liu, Zirui, et al. "Kivi: A tuning-free asymmetric 2bit quantization for kv cache." arXiv preprint arXiv:2402.02750(2024).

Choi, Dahun, Juntae Park, and Hyun Kim. "HLQ: Hardware-Friendly Logarithmic Quantization Aware Training for Power-Efficient Low-Precision CNN Models." IEEE Access(2024).

Jeong, Sangbeom, Seungil Lee, and Hyun Kim. "LowGradQ: Adaptive Gradient Quantization for Low-Bit CNN Training via Kernel Density Estimation-Guided Thresholding and Hardware-Efficient Stochastic Rounding Unit." 2025 Design, Automation & Test in

Europe Conference(DATE). IEEE, 2025.

Kim, Nam Joon, Jongho Lee, and Hyun Kim. "Hyq: Hardware-friendly post-training quantization for cnn-transformer hybrid networks." Proceedings of the Thirty-Third International Joint Conference on Artificial Intelligence, IJCAI-24. International Joint Conferences on Artificial Intelligence Organization. Vol. 8. 2024.

Choi, Dahun, and Hyun Kim. "GradQ-ViT: Robust and Efficient Gradient Quantization for Vision Transformers." Proceedings of the AAAI Conference on Artificial Intelligence. Vol. 39. No. 15. 2025.

Kim, Sungrae, and Hyun Kim. "Zero-centered fixed-point quantization with iterative retraining for deep convolutional neural network-based object detectors." IEEE Access 9(2021): 20828-20839.

Kang, Beom Jin, et al. "A survey of FPGA and ASIC designs for transformer inference acceleration and optimization." Journal of Systems Architecture 155(2024): 103247.

Li, Shiyao, et al. "Mbq: Modality-balanced quantization for large vision-language models." Proceedings of the Computer Vision and Pattern Recognition Conference. 2025.

Wang, Changyuan, et al. "Q-vlm: Post-training quantization for large vision-language models." Advances in Neural Information Processing Systems 37(2024): 114553-114573.

Yu, JiangYong, et al. "Mquant: Unleashing the inference potential of multimodal large language models via full static quantization." arXiv preprint arXiv:2502.00425(2025).

Xue, Yufei, et al. "VLMQ: Efficient Post-Training Quantization for Large Vision-Language Models via Hessian Augmentation." arXiv preprint arXiv:2508.03351(2025).

Molchanov, Pavlo, et al. "Importance estimation for neural network pruning." Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2019.

Wang, Zi, Chengcheng Li, and Xiangyang Wang. "Convolutional neural network pruning with structural redundancy reduction." Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2021.

Sui, Yang, et al. "Chip: Channel independence-based pruning for compact neural

networks." Advances in Neural Information Processing Systems 34(2021): 24604-24616.

Frantar, Elias, and Dan Alistarh. "Optimal brain compression: A framework for accurate post-training quantization and pruning." Advances in Neural Information Processing Systems 35(2022): 4475-4488.

Frantar, Elias, and Dan Alistarh. "Sparsegpt: Massive language models can be accurately pruned in one-shot." International conference on machine learning. PMLR, 2023.

Xia, Mengzhou, et al. "Sheared llama: Accelerating language model pre-training via structured pruning." arXiv preprint arXiv:2310.06694(2023).

Kim, Nam Joon, and Hyun Kim. "Fp-agl: Filter pruning with adaptive gradient learning for accelerating deep convolutional neural networks." IEEE Transactions on Multimedia 25(2022): 5279-5290.

Koo, Kwanghyun, and Hyun Kim. "V-skp: Vectorized kernel-based structured kernel pruning for accelerating deep convolutional neural networks." IEEE Access 11(2023): 118547-118557.

Kim, Nam Joon, and Hyun Kim. "Trunk pruning: Highly compatible channel pruning for convolutional neural networks without fine-tuning." IEEE Transactions on Multimedia 26(2023): 5588-5599.

Chen, Liang, et al. "An image is worth 1/2 tokens after layer 2: Plug-and-play inference acceleration for large vision-language models." European Conference on Computer Vision. Cham: Springer Nature Switzerland, 2024.

Xing, Long, et al. "Pyramiddrop: Accelerating your large vision-language models via pyramid visual redundancy reduction." arXiv preprint arXiv:2410.17247(2024).

Zhang, Yuan, et al. "Sparsevlm: Visual token sparsification for efficient vision-language model inference." arXiv preprint arXiv:2410.04417(2024).

Yang, Senqiao, et al. "Visionzip: Longer is better but not necessary in vision language models." Proceedings of the Computer Vision and Pattern Recognition Conference. 2025.

Zhang, Yuxin, et al. "Cam: Cache merging for memory-efficient llms inference." Forty-first international conference on machine learning. 2024.

Cao, Qingqing, Bhargavi Paranjape, and Hannaneh Hajishirzi. "PuMer: Pruning and merging tokens for efficient vision language models." arXiv preprint arXiv:2305.17530(2023).

Yang, Yifei, Zouying Cao, and Hai Zhao. "Laco: Large language model pruning via layer collapse." arXiv preprint arXiv:2402.11187(2024).

Wang, Dong, et al. "Forget the data and fine-tuning! just fold the network to compress." arXiv preprint arXiv:2502.10216(2025).

Huang, Chenyu, et al. "Emr-merging: Tuning-free high-performance model merging." Advances in Neural Information Processing Systems 37(2024): 122741-122769.

Horoi, Stefan, et al. "Harmony in diversity: Merging neural networks with canonical correlation analysis." arXiv preprint arXiv:2407.05385(2024).

Beltagy, Iz, Matthew E. Peters, and Arman Cohan. "Longformer: The long-document transformer." arXiv preprint arXiv:2004.05150(2020).

Wang, Sinong, et al. "Linformer: Self-attention with linear complexity." arXiv preprint arXiv:2006.04768(2020).

Choromanski, Krzysztof, et al. "Rethinking attention with performers." arXiv preprint arXiv:2009.14794(2020).

Zheng, Chuanyang. "The Linear Attention Resurrection in Vision Transformer." arXiv preprint arXiv:2501.16182(2025).

Han, Dongchen, et al. "Flatten transformer: Vision transformer using focused linear attention." Proceedings of the IEEE/CVF international conference on computer vision. 2023.

Katharopoulos, Angelos, et al. "Transformers are rnns: Fast autoregressive transformers with linear attention." International conference on machine learning. PMLR, 2020.

Cai, Han, et al. "Efficientvit: Lightweight multi-scale attention for high-resolution dense prediction." Proceedings of the IEEE/CVF international conference on computer

vision. 2023.

Guo, Jialong, et al. "Slab: Efficient transformers with simplified linear attention and progressive re-parameterized batch normalization." arXiv preprint arXiv:2405.11582(2024).

He, Chenhang, et al. "Scatterformer: Efficient voxel transformer with scattered linear attention." European Conference on Computer Vision. Cham: Springer Nature Switzerland, 2024.

Han, Dongchen, et al. "Agent attention: On the integration of softmax and linear attention." European conference on computer vision. Cham: Springer Nature Switzerland, 2024.

Tan, Mingxing, et al. "Mnasnet: Platform-aware neural architecture search for mobile." Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2019.

Cai, Han, Ligeng Zhu, and Song Han. "Proxylessnas: Direct neural architecture search on target task and hardware." arXiv preprint arXiv:1812.00332(2018).

Stamoulis, Dimitrios, et al. "Single-path nas: Designing hardware-efficient convnets in less than 4 hours." Joint European Conference on Machine Learning and Knowledge Discovery in Databases. Cham: Springer International Publishing, 2019.

Wu, Bichen, et al. "Fbnet: Hardware-aware efficient convnet design via differentiable neural architecture search." Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2019.

Bakhtiarifard, Pedram, Christian Igel, and Raghavendra Selvan. "EC-NAS: Energy consumption aware tabular benchmarks for neural architecture search." ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing(ICASSP). IEEE, 2024.

Zhao, Yiyang, and Tian Guo. "Carbon-efficient neural architecture search." Proceedings of the 2nd Workshop on Sustainable Computer Systems. 2023.

Zhao, Yiyang, et al. "CE-NAS: An end-to-end carbon-efficient neural architecture search

framework." Advances in Neural Information Processing Systems 37(2024): 82673-82704.

Xu, Jingjing, et al. "KNAS: green neural architecture search." International Conference on Machine Learning. PMLR, 2021.

Xia, Heming, et al. "Speculative decoding: Exploiting speculative execution for accelerating seq2seq generation." arXiv preprint arXiv:2203.16487(2022).

Leviathan, Yaniv, et al. "Fast inference from transformers via speculative decoding." International Conference on Machine Learning. PMLR, 2023.

Xin, Ji, et al. "BERxiT: Early exiting for BERT with better fine-tuning and extension to regression." Proceedings of the 16th conference of the European chapter of the association for computational linguistics: Main Volume. 2021.

Schuster, Tal, et al. "Confident adaptive language modeling." Advances in Neural Information Processing Systems 35(2022): 17456-17472.

Chen, Yanxi, et al. "EE-LLM: Large-Scale Training and Inference of Early-Exit Large Language Models with 3D Parallelism." International Conference on Machine Learning. PMLR, 2024.

Elhoushi, Mostafa, et al. "LayerSkip: Enabling early exit inference and self-speculative decoding." arXiv preprint arXiv:2404.16710(2024).

Xu, Jiaming, et al. "Specee: Accelerating large language model inference with speculative early exiting." Proceedings of the 52nd Annual International Symposium on Computer Architecture. 2025.

Fan, Siqi, et al. "Not all layers of llms are necessary during inference." arXiv preprint arXiv:2403.02181(2024).

Cai, Tianle, et al. "Medusa: Simple llm inference acceleration framework with multiple decoding heads." arXiv preprint arXiv:2401.10774(2024).

PCB Hero(2024) - "CPU vs. GPU vs. FPGA vs. ASIC". https://www.pcb-hero.com/blogs/lickys-column/cpu-vs-gpu-vs-fpga-vs-asic.

A. Agrawal, N. Kedia, A. Panwar, J. Mohan, N. Kwatra, B. Gulavani, A. Tumanov, R.

Ramjee, "Taming {Throughput-Latency} tradeoff in {LLM} inference with {Sarathi-Serve}", in 18th USENIX Symposium on Operating Systems Design and Implementation(OSDI 24), 2024, pp. 117-134.

S. Hur, S. Na, D. Kwon, J. Kim, A. Boutros, E. Nurvitadhi, and J. Kim, "A Fast and Flexible FPGA-Based Accelerator for Natural Language Processing Neural Networks," ACM Trans. Archit. Code Optim., vol. 20, no. 1, Feb 2023.

Mingqiang Huang, Ao Shen, Kai Li, Haoxiang Peng, Boyu Li, Yupeng Su, and Hao Yu. "Edgellm: A highly efficient cpu-fpga heterogeneous edge accelerator for large language models". IEEE Transactions on Circuits and Systems I: Regular Papers, 2025.

Tiandong Zhao, Shaoqiang Lu, Chen Wu, and Lei He. "Chatopu: An fpga-based overlay processor for large language models with unstructured sparsity." In Proceedings of the 43rd IEEE/ACM International Conference on Computer-Aided Design, pages 1-9, 2024.

Seungjae Moon, Jung-Hoon Kim, Junsoo Kim, Seongmin Hong, Junseo Cha, Minsu Kim, Sukbin Lim, Gyubin Choi, Dongjin Seo, Jongho Kim, et al. "Lpu: A latency-optimized and highly scalable processor for larg langeuage model inference". IEEE Micro, 2024.

Seongmin Hong, Seungjae Moon, Junsoo Kim, Sungjae Lee, Minsub Kim, Dongsoo Lee, and Joo-Young Kim. "Dfx: A low-latency multi-fpga appliance for accelerating transformer-based text generation". In 2022 55th IEEE/ACM International Symposium on Microarchitecture(MICRO), pages 616-630. IEEE, 2022.

Gholami, A., Yao, Z., Kim, S., Hooper, C., Mahoney, M. W., and Keutzer, K. "Ai and memory wall". IEEE Micro, pp. 1-5, 2024.

Minxuan Zhou, Weihong Xu, Jaeyoung Kang, and Tajana Rosing. "Transpim: A memory-based acceleration via software-hardware co-design for transformer." In 2022 IEEE International Symposium on High-Performance Computer Architecture (HPCA), pages 1071-1085. IEEE, 2022.

Hyeoncheol Kim, Taehoon Kim, Taehyeong Park, Donghyeon Kim, Yongseung Yu, Hanjun Kim, and Yongjun Park, "Accelerating LLMs using an Efficient GEMM Library and Target-Aware Optimizations on Real-World PIM Devices.". In Proceedings of the 23rd ACM/IEEE International Symposium on Code Generation and Optimization(CGO '25). Association for Computing Machinery, New York, NY, USA, 225–240.

Hyojung Lee, Daehyeon Baek, Jimyoung Son, Jieun Choi, Kihyo Moon, and Minsung Jang. "Paise: Pimaccelerated inference scheduling engine for transformer-based llm.", In 2025 IEEE International Symposium on High Performance Computer Architecture(HPCA), pages 1707–1719. IEEE, 2025.

Minseok Seo, Xuan Truong Nguyen, Seok Joong Hwang, Yongkee Kwon, Guhyun Kim, Chanwook Park, Ilkon Kim, Jaehan Park, Jeongbin Kim, Woojae Shin, et al, "Ianus: Integrated accelerator based on npu-pim unified memory system.", In Proceedings of the 29th ACM International Conference on Architectural Support for Programming Languages and Operating Systems(ASPLOS), Volume 3, pages 545–560, 2024.

Jungi Lee, Wonbeom Lee, and Jaewoong Sim, "Tender: Accelerating large language models via tensor decomposition and runtime requantization" In 2024 ACM/IEEE 51st Annual International Symposium on Computer Architecture(ISCA), pages 1048–1062, 2024.

Cong Guo, Jiaming Tang, Weiming Hu, Jingwen Leng, Chen Zhang, Fan Yang, Yunxin Liu, Minyi Guo, and Yuhao Zhu. "Olive: Accelerating large language models via hardware-friendly outlier-victim pair quantization." in Proceedings of the 50th Annual International Symposium on Computer Architecture(ISCA), pages 1–15, 2023.

Chenyang Yin, Zhenyu Bai, Pranav Venkatram, Shivam Aggarwal, Zhaoying Li, and Tulika Mitra. "Tereffic: Highly efficient ternary llm inference on fpga.", arXiv preprint arXiv:2502.16473, 2025.

Yubin Qin, Yang Wang, Zhiren Zhao, Xiaolong Yang, Yang Zhou, Shaojun Wei, Yang Hu,

and Shouyi Yin. "Mecla: Memory-compute-efficient llm accelerator with scaling sub-matrix partition.", In 2024 ACM/IEEE 51st Annual International Symposium on Computer Architecture(ISCA), pages 1032-1047. IEEE, 2024.

Hanrui Wang, Zhekai Zhang, and Song Han. "Spatten: Efficient sparse attention architecture with cascade token and head pruning." In 2021 IEEE International Symposium on High-Performance Computer Architecture(HPCA), pages 97-110. IEEE, 2021.

Huizheng Wang, Jiahao Fang, Xinru Tang, Zhiheng Yue, Jinxi Li, Yubin Qin, Sihan Guan, Qinze Yang, Yang Wang, Chao Li, et al. "Sofa: A compute-memory optimized sparsity accelerator via cross-stage coordinated tiling." in 2024 57th IEEE/ACM International Symposium on Microarchitecture(MICRO), pages 1247-1263. IEEE, 2024.

M. Horowitz, "Computing's energy problem(and what we can do about it)", in International Solid-State Circuits Conference(ISSCC), 2014.

Zhang, Shijin, et al. "Cambricon-X: An accelerator for sparse neural networks." 2016 49th Annual IEEE/ACM International Symposium on Microarchitecture(MICRO). IEEE, 2016.

Chen, Yu-Hsin, Joel Emer, and Vivienne Sze. "Eyeriss: A spatial architecture for energy-efficient dataflow for convolutional neural networks." ACM SIGARCH computer architecture news 44.3(2016): 367-379.

Qin, Yubin, et al. "Fact: Ffn-attention co-optimized transformer architecture with eager correlation prediction." Proceedings of the 50th Annual International Symposium on Computer Architecture. 2023.

Guo, Cong, et al. "Olive: Accelerating large language models via hardware-friendly outlier-victim pair quantization." Proceedings of the 50th Annual International Symposium on Computer Architecture. 2023.

Zadeh, Ali Hadi, et al. "Gobo: Quantizing attention-based nlp models for low latency and energy efficient inference." 2020 53rd Annual IEEE/ACM International Symposium

on Microarchitecture(MICRO). IEEE, 2020.

Park, Eunhyeok, Dongyoung Kim, and Sungjoo Yoo. "Energy-efficient neural network accelerator based on outlier-aware low-precision computation." 2018 ACM/IEEE 45th Annual International Symposium on Computer Architecture(ISCA). IEEE, 2018.

Ham, Tae Jun, et al. "ELSA: Hardware-software co-design for efficient, lightweight self-attention mechanism in neural networks." 2021 ACM/IEEE 48th Annual International Symposium on Computer Architecture(ISCA). IEEE, 2021.

Lu, Liqiang, et al. "Sanger: A co-design framework for enabling sparse attention using reconfigurable architecture." MICRO-54: 54th Annual IEEE/ACM International Symposium on Microarchitecture. 2021.

Chen, Yunji, et al. "Dadiannao: A machine-learning supercomputer." 2014 47th Annual IEEE/ACM International Symposium on Microarchitecture. IEEE, 2014.

Chen, Tianshi, et al. "Diannao: A small-footprint high-throughput accelerator for ubiquitous machine-learning." ACM SIGARCH Computer Architecture News 42.1(2014): 269-284.

Jintao Zhang, Zhuo Wang, and Naveen Verma. "In-Memory Computation of a Machine-Learning Classifier in a Standard 6T SRAM Array." in IEEE Journal of Solid-State Circuits, vol. 52, no. 4, pp. 915-924, 2017.

Vivek Seshadri, Yoongu Kim, Chris Fallin, Donghyuk Lee, Rachata Ausavarungnirun, Gennady Pekhimenko, Onur Mutlu, and Thomas C. Mowry. "RowClone: Fast and Energy-Efficient In-DRAM Bulk Data Copy and Initialization." in Proceedings of the 46th Annual IEEE/ACM International Symposium on Microarchitecture(MICRO), pp. 185-197, 2013.

Vivek Seshadri, Donghyuk Lee, Thomas Mullins, Hasan Hassan, Amirali Boroumand, Jongmoo Kim, Onur Mutlu, and Thomas C. Mowry. "Ambit: In-Memory Accelerator for Bulk Bitwise Operations Using Commodity DRAM Technology." in Proceedings of the 50th Annual IEEE/ACM International Symposium on Microarchitecture (MICRO), pp. 273-287, 2017.

Mingxuan He, Chanho Song, Insoo Kim, Cheolmin Jeong, Seongmin Kim, Inyup Park, and T. N. Vijaykumar. "Newton: A DRAM-Maker's Accelerator-in-Memory(AiM) Architecture for Machine Learning." in Proceedings of the 53rd Annual IEEE/ACM International Symposium on Microarchitecture(MICRO), pp. 372-385, 2020.

Sukhan Lee, Seok-Hee Kang, Jaeyoung Lee, Hyunsu Kim, Eunji Lee, Seongsik Seo, and Nam Sung Kim. "Hardware Architecture and Software Stack for PIM Based on Commercial DRAM Technology: Industrial Product." in Proceedings of the 48th Annual ACM/IEEE International Symposium on Computer Architecture(ISCA), pp. 43-56, 2021.

Guseul Heo, Seokjin Lee, Jaehun Cho, Hyungjun Choi, Seongjun Lee, Hyunwoo Ham, and Jaejin Park. "NeuPIMs: NPU-PIM Heterogeneous Acceleration for Batched LLM Inferencing." in Proceedings of the 29th ACM International Conference on Architectural Support for Programming Languages and Operating Systems(ASPLOS), Volume 3, pp. 722-737, 2024.

Ali Shafiee, Anirban Nag, Narayanan Muralimanohar, Rajeev Balasubramonian, John Paul Strachan, Matthew Hu, R. Stanley Williams, and Vivek Srikumar. "ISAAC: A Convolutional Neural Network Accelerator with In-Situ Analog Arithmetic in Crossbars." in ACM SIGARCH Computer Architecture News, vol. 44, no. 3, pp. 14-26, 2016.

Shubham Jain, Ankit Ranjan, Kaushik Roy, and Anand Raghunathan. "Computing in Memory with Spin-Transfer Torque Magnetic RAM." in IEEE Transactions on Very Large Scale Integration(VLSI) Systems, vol. 26, no. 3, pp. 470-483, 2017.

Kim, Yulhwa, et al. "Winning both the accuracy of floating point activation and the simplicity of integer arithmetic." The Eleventh International Conference on Learning Representations. 2023.

Jang, Jaeyong, et al. "Figna: Integer unit-based accelerator design for fp-int gemm preserving numerical accuracy." 2024 IEEE International Symposium on High-Performance Computer Architecture(HPCA). IEEE, 2024.

Park, Gunho, et al. "FIGLUT: An Energy-Efficient Accelerator Design for FP-INT GEMM Using Look-Up Tables." 2025 IEEE International Symposium on High Performance Computer Architecture(HPCA). IEEE, 2025.

You, Haoran, et al. "Shiftaddllm: Accelerating pretrained llms via post-training multiplication-less reparameterization." Advances in Neural Information Processing Systems 37(2024): 24822-24848.

Junki Park, Heesu Yoon, Dongkyun Ahn, Jinsu Choi, and Jae-Joon Kim. "OPTIMUS: OPTImized Matrix Multiplication Structure for Transformer Neural Network Accelerator." in Proceedings of Machine Learning and Systems(MLSys), vol. 2, pp. 363-378, 2020.

Zhe Zhou, Junjie Liu, Zhe Gu, and Guangming Sun. "Energon: Toward Efficient Acceleration of Transformers Using Dynamic Sparse Attention." in IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems, vol. 42, no. 1, pp. 136-149, 2022.

Qu, Z., Liu, L., Tu, F., Chen, Z., Ding, Y., & Xie, Y.(2022, February). Dota: detect and omit weak attentions for scalable transformer acceleration. In Proceedings of the 27th ACM International Conference on Architectural Support for Programming Languages and Operating Systems(pp. 14-26).

Tao Yang, Fan Ma, Xiang Li, Fangxin Liu, Yibo Zhao, Zhiwei He, and Li Jiang. "DTATrans: Leveraging Dynamic Token-Based Quantization with Accuracy Compensation Mechanism for Efficient Transformer Architecture." in IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems, vol. 42, no. 2, pp. 509-520, 2022.

Shikhar Tuli and Niraj K. Jha. "AccelTran: A Sparsity-Aware Accelerator for Dynamic Inference with Transformers." in IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems, vol. 42, no. 11, pp. 4038-4051, 2023.

Yubin Qin, Yang Wang, Dongsheng Deng, Zheyuan Zhao, Xinyu Yang, Lei Liu, and Shouyi Yin. "FACT: FFN-Attention Co-Optimized Transformer Architecture with Eager

Correlation Prediction." in Proceedings of the 50th Annual International Symposium on Computer Architecture(ISCA), pp. 1-14, 2023.

Dong, H., Wang, M., Luo, Y., Zheng, M., An, M., Ha, Y., & Pan, H.(2020). PLAC: Piecewise linear approximation computation for all nonlinear unary functions. IEEE Transactions on Very Large Scale Integration(VLSI) Systems, 28(9), 2014-2027.

Stevens, J. R., Venkatesan, R., Dai, S., Khailany, B., & Raghunathan, A.(2021, December). Softermax: Hardware/software co-design of an efficient softmax for transformers. In 2021 58th ACM/IEEE Design Automation Conference(DAC)(pp. 469-474). IEEE.

Kim, S., Gholami, A., Yao, Z., Mahoney, M. W., & Keutzer, K.(2021, July). I-bert: Integer-only bert quantization. In International conference on machine learning(pp. 5506-5518). PMLR.

Yu, J., Park, J., Park, S., Kim, M., Lee, S., Lee, D. H., & Choi, J.(2022, July). NN-LUT: Neural approximation of non-linear operations for efficient transformer inference. In Proceedings of the 59th ACM/IEEE Design Automation Conference(pp. 577-582).

LI, Zhikai; GU, Qingyi. I-vit: Integer-only quantization for efficient vision transformer inference. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. 2023. p. 17065-17075.

Kim, Janghyeon, et al. "Range-invariant approximation of non-linear operations for efficient BERT fine-tuning." 2023 60th ACM/IEEE Design Automation Conference(DAC). IEEE, 2023.

Wang, W., Zhou, S., Sun, W., Sun, P., & Liu, Y.(2023, October). Sole: Hardware-software co-design of softmax and layernorm for efficient transformer inference. In 2023 IEEE/ACM International Conference on Computer Aided Design(ICCAD)(pp. 1-9). IEEE.

Sadeghi, M. E., Fayyazi, A., Azizi, S., & Pedram, M.(2024, August). Peano-vit: Power-efficient approximations of non-linearities in vision transformers. In Proceedings of the 29th ACM/IEEE International Symposium on Low Power Electronics and Design(pp. 1-6).

Dong, P., Tan, Y., Zhang, D., Ni, T., Liu, X., Liu, Y., ... & Cheng, K. T.(2024, June). Genetic quantization-aware approximation for non-linear operations in transformers. In Proceedings of the 61st ACM/IEEE Design Automation Conference(pp. 1-6).

Du, H., Wen, C., Chen, Z., Zhang, L., Sun, Q., Yan, Z., & Zhuo, C.(2025, March). Algorithm-Hardware Co-Design of a Unified Accelerator for Non-Linear Functions in Transformers. In 2025 Design, Automation & Test in Europe Conference (DATE)(pp. 1-7). IEEE.

[기타 자료]

SK telecom(2023. 11. 14). "SKT, AI서버 액침냉각으로 전력사용 37% 절감 확인", https://news.sktelecom.com/199628.

연합뉴스(2023. 11. 16). "사피온, 데이터 센터용 '4배 빠른' AI 반도체 X330 출시", https://www.yna.co.kr/view/AKR20231116048500017.

IEA(2023). "Data Centres and Data Transmission Network", https://www.iea.org/energy-system/buildings/data-centres-and-data-transmission-networks.

Microsoft(2024. 6. 2). "Microsoft's Datacenter Community Pledge", https://blogs.microsoft.com/blog/2024/06/02/.

Businesswire(2025. 4. 29). "Korea Data Center Market Investment Analysis Report", https://www.businesswire.com/news/home/20250429109262/en/.

EEJournal(2025. 7. 22). "LG AI Research taps FuriosaAI to achieve 2.25x better LLM inference performance vs. GPUs", https://www.eejournal.com/industry_news/lg-ai-research-taps-furiosaai-to-achieve-2-25x-better-llm-inference-performance-vs-gpus/.

The guardian(2025). "Energy demands from AI datacentres to quadruple by 2030, says report". https://www.theguardian.com/technology/2025/apr/10/energy-demands-from-ai-data

centres-to-quadruple-by-2030-says-report.

탄소중립녹색성장위원회(2025. 6. 10.) “온실가스 배출량 경계 Scope 1, 2, 3”, https://www.2050cnc.go.kr/base/board/read?boardManagementNo=69&boardNo=5415&searchCategory=&page=1&searchType=&searchWord=&menuLevel=2&menuNo=97.

ADaSci(2024), How Much Energy Do LLMs Consume? Unveiling the Power Behind AI. https://adasci.org/how-much-energy-do-llms-consume-unveiling-the-power-behind-ai/.

EPOCH AI(2024), How Much energy does ChatGPT use? Do LLMs Consume? Unveiling the Power Behind AI. https://epoch.ai/gradient-updates/how-much-energy-does-chatgpt-use.

EMERGING TECHNOLOGIES(2024), How to manage AI's energy demand-todeay, tomorrow and in the future https://www.weforum.org/stories/2024/04/how-to-manage-ais-energy-demand-today-tomorrow-and-in-the-future/.

Hutchinson(2024), How optical fibre will help create a sustainable future for data centreshttps://www.intelligentdatacentres.com/2024/01/03/how-optical-fibre-will-help-create-a-sustainable-future-for-data-centres/.

Le Monde(2023), Smartphones' carbon footprints are largely underestimated. https://www.lemonde.fr/en/pixels/article/2023/04/16/smartphones-carbon-footprints-are-largely-underestimated_6023093_13.html.

canalys(2025), Worldwide PC shipments up 9% in Q1 2025 but tariffs threaten future market performance. https://canalys.com/newsroom/worldwide-pc-shipments-q1-2025.

디지털데일리(2022), [데이터센터 퀀텀점프] 우후죽순 늘어나는 데이터센터, '지속가능성'이 화두. https://www.ddaily.co.kr/page/view/20220629150130165OO.

그린데이터센터, https://kdcc.or.kr/green/pageview.do?url=green01&keyvalue=green

KOTRA(2022. 3. 22.), “中 메가급 프로젝트 '동수서산'(東數西算)”,

https://dream.kotra.or.kr/dream/kotra/actionKotraShortUrl/aAs2Jlw4NPr6.do.

SK 에코플랜트, https://news.skecoplant.com/plant-tomorrow/18677.

Microsoft(2024), Sustainable by design: Next-generation datacenters consume zero water for cooling.

전기신문(2025. 4. 28.), “입찰연계형 PPA 2차 시범사업, 용량 완화·계약 유연화로 재시동”, https://www.electimes.com/news/articleView.html?idxno=353993.

_____(2025. 4. 28.), 동아일보(2023. 7. 10), “편리한 클라우드 서비스, 에너지 절약에도 ‘효자’”, https://www.donga.com/news/Economy/article/all/20230709/120151296/1.

“입찰연계형 PPA 2차 시범사업, 용량 완화·계약 유연화로 재시동”, https://www.electimes.com/news/articleView.html?idxno=353993.

ZDNET Korea(2024. 6. 16.), “제조업 설계도 클라우드로...KT, 비용 60% 절감”, https://zdnet.co.kr/view/?no=20240616072215.

전자신문(2024. 8. 12.), 정부, “2030년까지 기존 시스템 90% 클라우드 네이티브로 전환”, https://www.etnews.com/20240812000245.

Deloitte Insights(2024), Dialing down the carbon: Telco sustainability surges on the back of four new trends.

Global Market Insights(2024), Software Defined Networking Market.

전자신문(2024. 4. 15). “정부·통신·장비사, 네트워크 저전력화 머리 맞댄다”, https://www.etnews.com/20240415000319.

전자신문(2025. 3. 16.), “韓, AI 내재화 통신 연구 주도…6G 표준화 앞장선다”, https://www.etnews.com/20250316000024.

정부 및 정책금융기관의 기후위기 대응 금융지원 확대방안(2024), https://greenium.kr.

2030년까지 기후위기대응에 452조 민관금융지원 발표(2024), https://www.yna.co.kr.

담보력 부족한 녹색기업 우대보증 지원 정책(2025), https://www.me.go.kr.

기후위기 대응을 위한 금융지원 관련 은행장 회의(2024), https://www.fss.or.kr.

정책·정보 ¦ 규제정보 ¦ 규제샌드박스 — 국무조정실(2024), https://www.opm.go.kr.

기후위기 대응을 위한 탄소중립·녹색성장 기본법 — https://www.law.go.kr/법령/기후위기대응을위한탄소중립·녹색성장기본법.

정부 및 정책금융기관의 기후위기 대응 금융지원 확대방안(2024), https://greenium.kr.
2030년까지 기후위기대응에 452조 민관금융지원 발표(2024), https://www.yna.co.kr.
담보력 부족한 녹색기업 우대보증 지원 정책(2025), https://www.me.go.kr.

디지털 대전환 메가트렌드 연구 시리즈

25-28-01 디지털 대전환 메가트렌드 연구 총괄보고서Ⅴ: 2035 AI·디지털 대전환 메가트렌드 (문아람 외, 정보통신정책연구원)

25-28-02 2035 AI · 디지털 대전환 미래전략 (문아람 외, 정보통신정책연구원)

25-28-03 디지털 전환 기반 구축을 위한 차세대 통신기술 및 양자기술 활용 방안 (이인규 외, 한국통신학회)

25-28-04 디지털 자산 시장 발전에 따른 혁신과 공정의 기업 생태계 조성 (양성병 외, 한국경영학회)

25-28-05 디지털 전환기 기업 동태 분석 기반 혁신성장 정책 방향 (이경원 외, 정보통신정책학회)

25-28-06 AI·디지털 전환기 노동 구조의 재편과 노동권 보호를 위한 안전망 설계 (임운택 외, 한국사회학회)

25-28-07 디지털 전환기 민주주의 회복력 강화를 위한 정치제도의 변화 (김범수 외, 한국정치학회)

25-28-08 디지털 전환에 대응하는 AI·데이터 기반 디지털 복지서비스 개발 연구 (윤건 외, 한국행정학회)

25-28-09 AI·디지털 전환에 따른 공공부문 데이터 거버넌스 전략 (남태우 외, 한국정책학회)

25-28-10 디지털 저탄소 전환을 위한 기술적·제도적 방안 (김현 외, 대한전자공학회)

25-28-11 디지털 전환기 AGI 기술 경쟁력 확보를 위한 중장기 전략 (송길태 외, 한국정보과학회)

● 저 자 소 개 ●

김 현

- 서울대학교 전기컴퓨터공학 석사
- 서울대학교 전기컴퓨터공학 박사
- 현 서울과학기술대학교 교수

김 남 준

- 서울과학기술대학교 전기정보공학과 석사
- 현 서울과학기술대학교 전기정보공학과 박사과정

구 광 현

- 서울과학기술대학교 전기정보공학과 석사
- 현 서울과학기술대학교 전기정보공학과 박사과정

장 지 훈

- 서울과학기술대학교 전기정보공학과 석사
- 현 서울과학기술대학교 전기정보공학과 박사과정

황 인 성

- 서울과학기술대학교 전기정보공학과 석사
- 현 서울과학기술대학교 전기정보공학과 박사과정

강 범 진

- 서울과학기술대학교 전기정보공학과 석사
- 현 서울과학기술대학교 전기정보공학과 박사과정

이 승 일

- 서울과학기술대학교 전기정보공학과 석사
- 현 서울과학기술대학교 전기정보공학과 박사과정

신 진

- 서울과학기술대학교 전기정보공학과 석사
- 현 서울과학기술대학교 전기정보공학과 박사과정

최 다 훈

- 서울과학기술대학교 전기정보공학과 석사
- 현 서울과학기술대학교 전기정보공학과 박사과정

이 길 하

- 현 서울과학기술대학교 전기정보공학과 박사과정

정 상 범

- 서울과학기술대학교 전기정보공학과 석사
- 현 서울과학기술대학교 전기정보공학과 박사과정

ICT 진흥 및 혁신기반 조성사업 RS-2025-02173110
정책연구 25-28-10

디지털 저탄소 전환을 위한 기술적·제도적 방안 (A Study on Technological and Institutional Measures for Digital Low-Carbon Transformation)

2025년 12월 일 인쇄
2025년 12월 일 발행

발행인 과학기술정보통신부 장관
발행처 과학기술정보통신부
세종 갈매로 477 정부세종청사 4동
Homepage: www.msit.go.kr
인 쇄 인 성 문 화