

# 인공지능의 에너지 효율화와 엣지 컴퓨팅

## Energy Efficiency and Edge Computing in Artificial Intelligence

김말희 (M.R. Kim, mariekim@etri.re.kr)

환경ICT연구실 책임연구원

허태욱 (T.-W. Heo, htw398@etri.re.kr)

환경ICT연구실 책임연구원/실장

이일우 (I.W. Lee, ilwoo@etri.re.kr)

산업에너지융합연구본부 책임연구원/본부장

### ABSTRACT

With the rapid advancement of artificial intelligence (AI) technology, models are becoming larger and computational demands are increasing. While this trend significantly enhances AI performance, it also raises concerns regarding massive energy consumption and greenhouse gas emissions. These challenges not only threaten corporate sustainability but also pose a critical issue for the competitiveness of the AI industry.

In this context, the emergence of DeepSeek has attracted attention as a positive example demonstrating that generative AI can be implemented and operated using energy-efficient computing infrastructures. This finding suggests the possibility of AI technology evolving in a more sustainable and eco-friendly direction.

Addressing AI's energy consumption problem requires various technological approaches. Key solutions include model compression, mixture of experts, sparse computing, on-device AI, and edge computing. Additionally, major companies such as Google are actively researching AI energy efficiency improvements by developing power-optimized AI semiconductors (e.g., neural processing units), optimizing power management in cloud and data centers, and enhancing the computational efficiency of AI algorithms.

This report focuses on analyzing AI energy efficiency technologies and industry trends while proposing strategic directions for constructing sustainable AI systems.

**KEYWORDS** 모델 경량화, 엣지 컴퓨팅, 연산 최적화, 온 디바이스 인공지능

## 1. 서론

2022년 OpenAI의 ChatGPT 등장 이후 생성형 AI 서비스는 Google의 GEMINI를 비롯해 다양한 분야

에서 빠르게 확산되고 있다. 일반 대중의 접근성이 향상되면서, AI는 단순한 질의응답을 넘어 프로그램 코딩, 문서 작성, 음악 생성, 회화 등 폭넓은 영역에서 활용되고 있다.

\* DOI: <https://doi.org/10.22648/ETRI.2025.J.400303>

\* 본 연구는 산업통상자원부(MOTIE)와 한국에너지기술연구원(KETEP)의 지원을 받아 수행한 연구 과제입니다[No. RS-2023-00237018].



본 저작물은 공공누리 제4유형

출처표시+상업적이용금지+변경금지 조건에 따라 이용할 수 있습니다.

©2025 한국전자통신연구원

하지만 AI 기술이 발전하고 사용 범위가 확대되면서 AI 모델의 연산량과 에너지 소비도 급격히 증가하고 있다. 특히 대규모 언어 모델(LLM: Large Language Model)과 같은 생성형 AI는 성능이 향상될수록 더 많은 전력을 소비하는 구조적 한계를 갖는다. 이러한 문제를 해결하기 위해 에너지 효율적인 AI 기술과 엣지 컴퓨팅이 중요한 대안으로 주목받고 있다.

본고에서는 AI의 에너지 효율화를 위한 전략과 엣지 컴퓨팅의 역할을 살펴보고, 지속 가능한 AI 시스템 구축 방안을 논의한다. II장에서는 AI의 에너지 소비 문제와 해결 필요성, III장에서는 에너지 효율화 관련 기술 동향, IV장에서는 엣지 컴퓨팅과 AI 에너지 효율화, V장에서는 주요 사례 분석, VI장에서는 지속 가능한 AI 환경 구축 전략을 기술한다.

## II. AI 에너지 소비 문제와 해결 필요성

### 1. AI 데이터센터의 전력 소비 증가

생성형 AI 및 대규모 언어 모델의 확산으로 인해 AI 시스템의 연산 요구량이 폭증하고 있으며, 이에 따라 데이터센터의 전력 소비가 급격히 증가하고 있다. 특히, AI 모델의 훈련(Training)과 추론(Inference) 과정에서 대규모 GPU(Graphic Processing Unit) 및 AI 가속기의 사용이 필수적이다. 이에 따라서 데이터센터 운영 비용과 에너지 소비 부담이 지속적으로 증가하는 추세다.

Gartner 보고서에 따르면, 2027년까지 AI 작업량을 처리하는 데이터센터의 40%가 전력 공급 부족 문제에 직면할 것으로 전망된다[1]. 또한, 데이터센터의 냉각 시스템 운영 비용이 증가하면서 추가적인 전력 소비 문제도 발생하고 있다. AI 연산을 최적화하지 않는다면, 데이터센터의 전력 소비는 앞으로 기하급수적으로 증가할 가능성이 크다.

### 2. 대형 AI 모델 확장과 에너지 비용 증가

AI 모델의 크기가 기하급수적으로 증가하면서, 연산 요구량과 이에 따른 전력 소비 문제가 더욱 심화되고 있다. 최신 AI 모델의 파라미터 수는 수십억에서 수조 개에 이르며[2], 이러한 초거대 모델을 훈련하고 운영하는 데 막대한 연산 자원과 전력이 필요하다.

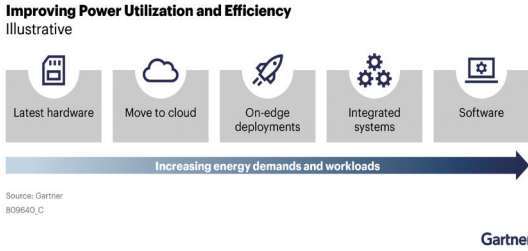
예를 들어, OpenAI의 GPT-4 모델을 훈련하는 데 수천 개의 고성능 GPU가 수주에서 수개월간 작동해야 하며, 이 과정에서 수천 MWh(Megawatt-hour)의 전력이 소비되는 것으로 추정된다[3]. AI 모델의 연산 효율성을 높이기 위한 반도체 기술이 발전하고 있지만, AI 모델의 크기 증가 속도가 하드웨어 성능 개선 속도를 초과하고 있어 근본적인 해결책이 되기 어렵다.

또한, 클라우드 기반 AI 서비스의 확산으로 인해 데이터 전송과 저장 비용도 급증하고 있으며, 이는 AI 시스템을 운영하는 기업에 상당한 부담을 초래하고 있다. 따라서, AI 모델의 성능을 유지하면서도 에너지 효율성을 극대화하기 위한 새로운 기술적 접근이 필요하다.

AI 에너지 소비 증가는 단순히 경제적인 문제뿐만 아니라 환경 문제도 초래한다. 데이터센터 운영에 필요한 전력 생산 과정에서 발생하는 탄소 배출량은 지구 온난화를 심화시키고 기후 변화를 가속화하고 있다.

### 3. AI 에너지 문제 해결 필요성

AI 에너지 문제 해결 필요성은 지속 가능한 AI 발전을 위해, 환경 보호를 위해, 그리고 기업들의 경제적 부담 완화를 위해 반드시 해결해야 할 과제이다. AI 모델 운영 비용 증가는 기업에 큰 부담으로 작용



Gartner

출처 Emerging Tech: Strategies to Achieve Energy Efficiency Goals for GenAI, Gartner, 2024.

그림의 저작권은 Gartner에 있으며, Gartner의 동의하에 사용되었습니다. 추 후 이용 시 Gartner에 문의하시기 바랍니다.

그림 1 에너지 효율화 방법

한다. 에너지 효율성을 높여 에너지 소비량을 줄이는 것은 기업들의 인공지능 기술 도입 장벽을 낮추는데 도움이 된다.

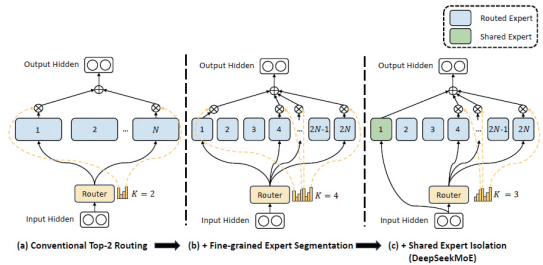
가트너는 생성형 AI 모델의 전력 사용 효율화를 위해서는 최신 하드웨어 사용, 클라우드로의 이관, on-edge 배치, 소프트웨어와 하드웨어가 공동 설계되어 높은 에너지 효율성을 제공하는 맞춤형 시스템, 소프트웨어 최적화 방법으로 AI의 에너지 효율성을 높일 수 있다고 한다(그림 1). 단기적으로는 최신의 하드웨어가 해결책이 될 수 있으나 궁극적으로는 소프트웨어 최적화가 주요 전략이 될 것으로 전망한다[4].

### III. AI 에너지 효율화 기술 동향

#### 1. AI 모델 경량화 및 최적화 기법

대형 AI 모델의 연산 부담을 줄이기 위해 모델 경량화 및 최적화 기법이 활발히 연구되고 있다. 대표적인 기법으로는 MoE 구조, 모델 프루닝(Pruning), 양자화(Quantization), 지식 증류(Knowledge Distillation) 등이 있다.

- Mixture of Experts(MoE): 모델의 일부 계층을 여러 전문가(Experts) 그룹으로 나누고, 입력 데이터가 적절한 전문가를 선택해 연산하는 기법이

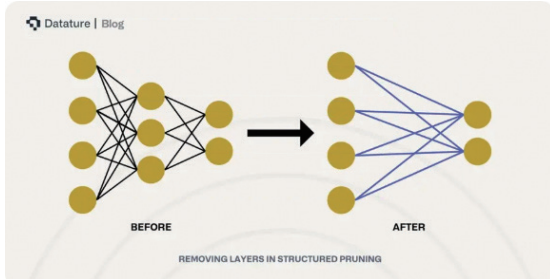


출처 D. Dai et al., "DeepSeekMoE: Towards Ultimate Expert Specialization in Mixture-of-Experts Language Models," in Proc. Annu. Meeting Assoc. Comput. Linguistics., (Bangkok, Thailand), vol. 1, Aug. 2024, pp. 1280–1297.

그림 2 DeepSeekMoE 개요

다. 여러 전문가 모델 유지를 위한 전체 파라미터 수는 증가하지만, 실제 활성화되는 파라미터가 제한되어 연산량과 메모리 사용량을 절감할 수 있다. Google의 GShard[5]는 Transformer 모델의 일부 FNN(Feed Forward Network)을 MoE 계층으로 대체하고, 각 토큰이 최대 두 개의 전문가만 선택하도록 설계해 모델을 효율적으로 분산 학습했다. DeepSeekMoE[6]는 이를 개선하여 전문가를 세분화하고, 공통 지식을 학습하는 공유 전문가와 특정 지식을 학습하는 전문가를 분리해 연산 효율성과 전문성을 더욱 향상시켰다(그림 2). MoE는 모델 크기를 확장하면서도 연산량과 메모리 사용을 줄이는 AI 모델 최적화 기술로 활용된다.

- 모델 프루닝(Pruning): 중요도가 낮은 뉴런과 연결을 제거하여 연산량을 줄이며, 이를 통해 AI 모델의 추론 과정에서 전력 효율성을 개선할 수 있다(그림 3)[7]. 일반적인 소프트 프루닝 방식에서는 뉴런과 가중치를 계산상에서만 제외하지만, 하드 프루닝을 적용하면 모델 구조 자체에서 불필요한 뉴런과 연결이 제거되어 메모리 사용량까지 감소시킬 수 있다. 최적의 에너지 효율을 위해서는 프루닝과 양자화를 함께 적용하는 것이 효과적이다.



출처 Reprinted with permission from Datature, "A comprehensive guide to neural network model pruning," Datature Blog, 2024. 2. 29. <https://www.datature.io/blog/a-comprehensive-guide-to-neural-network-model-pruning>

그림 3 모델 프루닝 개요

- 양자화(Quantization), 저정밀 연산(Reduced Precision Computing): 연산 과정에서 사용되는 숫자의 정밀도를 낮추어 AI 모델의 메모리 및 전력 소모를 절감하는 방법이다[8].
- 지식 증류(Knowledge Distillation): 대형 모델의 지식을 경량화된 모델에 전달(Transfer)하여 추론 시에 경량화된 모델을 사용하도록 한다. 기존 성능을 유지하면서도 경량 모델의 전력 소비를 최소화하는 기술이다[9,10]. 지식 증류 과정에서는 Teacher Model과 Student Model이 별도로 존재하며, 최종적으로는 Student Model만 배포되어 운영된다. 이 과정을 통해 성능을 유지하면서도 연산량과 에너지 소비를 크게 줄일 수 있다.
- 저랭크 근사(Low-Rank Approximation): 행렬의 차원을 줄여 연산량을 줄이고 모델을 경량화하는 기법이다[11].

이러한 기법들은 AI 모델의 전력 효율성을 높이는 데 중요한 역할을 하며, 엣지 디바이스와 같은 제한된 컴퓨팅 환경에서도 효과적인 AI 실행을 가능하게 한다.

## 2. AI 소프트웨어 최적화 기법

AI 모델의 전력 소비를 줄이기 위해 소프트웨어 차원의 최적화가 필수적이다. 이를 위해 희소 연산(Sparse Computing), 에너지 인식 학습(Energy-aware Training), Green Software Engineering 등의 기법이 연구되고 있다(표 1).

- 희소 연산: 필요하지 않은 연산을 건너뛰도록 하여 전력 소비를 줄이는 기법이며, 특히 AI 추론 과정에서 효과적이다[12].
- 에너지 인식 학습: AI 모델이 훈련될 때 전력 소비를 고려하여 학습하는 방식으로, 학습 과정에서 불필요한 연산을 줄이고 저비용으로 높은 성능을 유지하는 것을 목표로 한다[13].
- Green Software Engineering: AI 소프트웨어를 개발할 때 에너지 소비를 최소화하는 방법론으로, 코드 최적화, 알고리즘 개선, 클라우드 자원 최적화 등이 포함된다[14].
- 클러스터 기반 연산 최적화: 분산된 환경에서 연산을 최적화하여 불필요한 자원 사용을 줄이는 기법으로, AI 데이터센터에서 활용된다[15].

표 1 AI 소프트웨어 최적화 기법

기법	설명	주요 적용 대상
희소 연산	필요하지 않은 연산을 피해서 전력 소비 효율화	AI 추론 (Inference)
에너지 인식 학습	AI 모델 훈련 시 전력 소비를 고려하여 불필요한 연산을 최소화	AI 모델 학습 (Training)
Green Software Engineering	AI 소프트웨어 개발 과정에서 에너지 소비를 최소화	AI 소프트웨어 개발
클러스터 기반 연산 최적화	분산 환경에서 연산을 최적화하여 연산 자원 사용을 최적화	AI 데이터센터, 클라우드
런타임 최적화	AI 모델 실행 환경을 최적화하여 메모리 사용량, CPU/GPU 사용량, 전력 소비량을 줄이는 기술	AI 모델 실행 (Inference & Training)

- 런타임 최적화: AI 모델 실행 환경을 최적화하여 메모리 사용량, CPU(Central Processing Unit) 사용량, 전력 소비량 등을 줄이는 기술이다.

### 3. AI 반도체 및 하드웨어 최적화

AI 모델의 전력 소비 문제를 해결하기 위해 저전력 AI 반도체 및 특화 하드웨어가 개발되고 있다.

- AI 특화 칩셋: AI 연산을 최적화한 전용 프로세서로, 범용 CPU·GPU 대비 높은 성능과 전력 효율을 제공한다. Google TPU(Tensor Processing Unit)는 텐서 연산을 가속하여 AI 모델 학습을 최적화한다. Tesla Dojo는 자율주행 AI 학습을 위한 고성능 칩을 제공한다[16]. ARM Ethos NPU는 모바일·임베디드 기기의 저전력 AI 연산을 지원한다[17].
- 광 기반 AI 가속기(Photonic AI Accelerators): 기존 전자 회로 대신 광 기반 연산을 활용하여 초저전력 AI 연산을 가능하게 하는 기술이다. 전자식 칩보다 빠르고 에너지 효율적인 연산을 제공한다. AI 모델의 연산 속도를 향상하고 전력 소비를 절감할 수 있다. Lightelligence는 광학 기반 AI 칩을 개발하여 연산 성능을 극대화하고[18], Optalysys는 광학 컴퓨팅 기술을 활용한 고성능 AI 연산 연구를 진행하고 있다. 프랑스의 Light On과 같은 기업들이 광학 컴퓨팅 솔루션을 개발하고 있다.
- 메모리 중심 AI 연산(Memory-Centric AI Computing): 연산을 메모리 내에서 수행하여 데이터 이동으로 인한 전력 소모를 최소화하는 기술이다. NVIDIA는 HBM(High Bandwidth Memory)3 메모리를 통합한 Hopper H100 GPU로 메모리 병목을 완화시켰고, 후속으로 HBM3E도 준비 중이다. 삼성전자는 2024년 HBM3 생산에 대한 엔

비디아 품질 인증을 통과했다.

- 칩렛(Chiplet) 및 3D IC 패키징: 칩렛 기술은 하나의 칩을 여러 개의 작은 모듈(칩렛)로 분할하여 성능과 확장성을 높이는 방식이다. 3D IC 패키징은 칩을 수직으로 적층해 데이터 전송 거리를 줄이고 전력 소모를 감소시키는 기술이다. TSMC는 칩을 웨이퍼와 기판에 통합하는 기술 CoWoS(Chip on Wafer on Substrate) 및 칩을 3D로 적층하는 기술 SoIC(System on Integrated Chips)를 개발했다. 인텔은 Foveros 3D 패키징을 적용한 Meteor Lake를 출시했다.
- 뉴로모픽 컴퓨팅(Neuromorphic Computing): 인간의 뇌 신경망 구조와 동작 방식을 모방한 AI 반도체 및 하드웨어 설계 기법이다. 기존의 폰 노이만 아키텍처와 대비되는 저전력, 고효율 연산 방식을 제공한다. Intel은 Loihi 칩을 개발하여 신경망 학습과 추론을 최적화하고 있으며[19], IBM은 TrueNorth 칩을 통해 뇌과학 연구 및 AI 모델 개발을 진행하고 있다. Qualcomm과 AMD 등 다양한 기업이 뉴로모픽 반도체 및 AI 하드웨어 연구를 지속적으로 확대하고 있다.

표 2는 AI 하드웨어 기술과 관련 대표 기업을 정리한 것이다.

표 2 AI 하드웨어 기술 및 대표 기업

기술 분야	대표 기업 및 기술
AI 특화 칩셋	Google TPU, Tesla Dojo, ARM Ethos NPU, NVIDIA A100/H100, Grace Hopper Superchip
광 기반 AI 가속기	Lightmatter, Lightelligence, Optalysys
메모리 중심 AI 연산	NVIDIA HBM3(Hopper H100), IBM PIM, 삼성 전자·SK하이닉스 HBM4
칩렛 및 3D IC 패키징	TSMC CoWoS/SoIC, Intel Foveros 3D(Meteor Lake), 삼성전자 H-Cube/X-Cube
뉴로모픽 컴퓨팅	Intel Loihi, IBM TrueNorth, Qualcomm, AMD

## 4. 지속 가능한 에너지 활용

데이터센터의 지속 가능한 운영을 위해 태양광, 풍력 등의 신재생 에너지를 적극 활용하는 것이 바람직하다. 또한, 에너지 저장 장치(ESS: Energy Storage System)와 연계하여 전력 공급의 안정성을 확보하는 것이 중요하다. 예를 들어, Google은 데이터센터에 풍력과 태양열 발전을 도입하고[20], Amazon은 태양광 발전을 활용하여 친환경 에너지 사용을 확대하고 있다. 또한, ESS를 통해 신재생 에너지의 변동성을 조절하여 데이터센터의 안정적인 전력 공급을 지원할 수 있으며, 이를 통해 에너지 효율성을 극대화하고 지속 가능성을 높일 수 있다.

## 5. AIops 기반 에너지 효율성 향상

AIops(Artificial Intelligence for IT Operations)는 AI 인프라의 전력 소비 최적화 및 운영 자동화를 통해 에너지 절감과 비용 절감을 지원하는 핵심 기술이다. 2027년까지 AI 서버 운영에 필요한 전력이 2023년 대비 2.6배 증가(500TWh/년)할 것으로 예상되며, 40%의 데이터센터가 전력 가용성 문제로 AI 인프라 확장에 제약을 받을 것으로 전망된다. 하지만 반도체 기술 혁신만으로는 AI의 전력 문제를 완전히 해결할 수 없으며, AIops 기반의 전력 관리 최적화 및 소프트웨어 혁신이 필수적이다[21].

AIops는 실시간 전력 소비 분석, 이상 탐지, 전력 최적화 자동화, AI 클라우드와 온프레미스의 전력 효율 비교 분석 등의 기능을 제공한다. 이를 통해 기업들은 데이터센터의 에너지 소비를 효과적으로 관리하고 탄소 배출량을 줄이며, 지속 가능한 AI 인프라 구축이 가능해진다.

## IV. 엣지 컴퓨팅과 AI에너지 효율화

### 1. 엣지 컴퓨팅의 개념과 필요성

엣지 컴퓨팅은 AI 연산을 클라우드에서 실행하는 대신, 데이터가 생성되는 현장에서 처리하는 방식이다. 이를 통해 클라우드 서버로의 데이터 전송에 따른 네트워크 비용과 전력 소비를 줄일 수 있다. 자율주행, 스마트 팩토리, 헬스케어 등 실시간 데이터 처리가 필수적인 분야에서 엣지 컴퓨팅은 반응 속도 개선과 에너지 효율 향상에 기여한다. 다만, 엣지 컴퓨팅의 에너지 절감 효과를 최대화하기 위해서는 엣지 디바이스와 중앙 시스템 간의 효율적인 협업 및 전체 시스템의 통합적 최적화가 필수적이다.

### 2. 엣지 AI 최적화 기술

엣지 AI 최적화 기술은 여러 관점에서 분류할 수 있으며, 여기에서는 세 가지 주요 범주로 나누어 설명한다.

- 모델 최적화: 프루닝, 양자화, 지식 증류 등 다양한 기법을 통해 모델을 경량화해서 엣지에서 분석이 가능하도록 한다. 특히, 도메인 특화 모델은 특정 산업이나 업무에 최적화되어 최소한의 연산으로도 높은 성능을 달성할 수 있다.
- 하드웨어 최적화: 엣지 디바이스에 최적화된 전용 AI 칩(예: Google Edge TPU, ARM 기반 NPU 등)과 FPGA(Field Programmable Gate Array), ASIC(Application-Specific Integrated Circuit) 등 하드웨어 솔루션을 활용하여 에너지 효율을 개선한다.
- 소프트웨어-하드웨어 공동 최적화 및 데이터 관리: 온-디바이스 AI(On-Device AI) 방식처럼 AI 연산을 로컬에서 수행하는 동시에, 엣지와 클라우드 간의 효율적인 데이터 분산, 전송 최적화 및 시스템 통합 설계를 통해 전체 에너지

사용을 줄인다.

### 3. 온-디바이스 AI

온-디바이스 AI는 데이터를 로컬 디바이스에서 직접 처리한다. 이를 통해 데이터 전송 과정에서 발생하는 전력 소모와 보안 문제를 개선한다. 스마트폰, IoT, 스마트홈, 웨어러블 등 다양한 디바이스에 적용되어 에너지 효율성을 높이고 친환경 AI 시스템 구축에 기여한다. 또한, 지속 가능한 AI 연산 모델은 AI 시스템이 실시간 전력 사용량을 모니터링하고 최적화함으로써 전체 네트워크와 디바이스 간 협업을 통해 에너지 절감 및 탄소 배출 감소를 실현하는데 중요한 역할을 한다. 온-디바이스 AI는 엣지 AI의 하위 개념으로 표 3과 같이 비교해 볼 수 있다.

## V. 주요 사례 분석

### 1. DeepSeek MoE 기반 AI 모델 경량화

DeepSeek의 접근 방식은 값비싼 하드웨어에만 의

표 3 엣지 AI vs 온-디바이스 AI

구분	엣지 AI	온-디바이스 AI
정의	데이터 생성 지점에서 가까운 엣지 서버 또는 네트워크 장치에서 AI 연산을 수행	스마트폰, 태블릿, 웨어러블 기기 등 최종 사용자 장치 자체에서 AI 연산을 수행
연산 위치	엣지 서버, 게이트웨이, 로컬 데이터센터	최종 사용자 장치 (스마트폰, IoT, 웨어러블 등)
네트워크 필요성	엣지 서버와 연결	독립적 연산 가능
응용 분야	자율주행차, 스마트 팩토리, 의료 AI, CCTV 분석	스마트폰 음성 비서, 실시간 번역, 건강관리
장점	클라우드 대비 저지연, 실시간 데이터 처리 가능	개인정보 보호 강화, 네트워크 비용 절감
단점	엣지 서버 구축 필요, 네트워크 연결이 필요한 때도 있음	디바이스의 연산 성능이 제한적

존하는 것이 아니라 혁신적인 소프트웨어 솔루션을 통해 경쟁 우위를 확보할 수 있음을 보여준다. DeepSeek은 MoE(Mixture of Experts) 기반의 AI 모델을 활용하여 AI 연산 비용을 줄이면서도 높은 성능을 유지하는 사례를 보여주었다.

### 2. Google TPU 및 Switch Transformer 기반 AI 에너지 절감

Google TPU는 AI 연산을 최적화하여 전력 효율성을 높인 대표적인 AI 하드웨어로, 6세대 TPU인 트릴리움(Trillium)은 이전 세대 대비 학습 성능이 4배, 추론 처리량이 3배 증가하면서도 전력 효율이 67% 향상되었다. 한편, 소프트웨어 측면에서는 Google의 Switch Transformer 모델[22]이 Mixture of Experts(MoE) 구조를 최적화하여 불필요한 연산을 최소화하고, 통신 및 계산 비용을 줄임으로써 전력 사용을 더욱 효율화했다.

### 3. Tesla Dojo AI Supercomputer 및 AI 최적화 사례

Tesla Dojo는 자율주행 AI 학습을 위한 고효율 슈퍼컴퓨터로, 기존 GPU 기반 데이터센터 대비 낮은 전력으로 더 높은 성능을 제공하도록 설계되었다. Dojo는 Tesla가 자체 개발한 D1 칩을 기반으로 구성되며, 고도로 병렬화된 AI 연산을 지원하여 학습 속도를 가속화하면서도 전력 소비를 최적화한다. 특히, Dojo의 AI 최적화 시스템은 자율주행 AI 학습을 빠르고 효율적으로 지원하며, 기존 대비 전력 소비를 대폭 절감하였다. 이러한 혁신적인 설계를 통해 Tesla는 데이터센터 운영 비용을 절감하는 동시에, AI 에너지 효율성을 높였다.

#### 4. ARM Ethos NPU의 초저전력 엣지 AI

ARM의 Ethos NPU는 스마트폰, IoT 기기, 웨어러블 등 엣지 디바이스에서 전력 소비를 최소화하는 AI 가속 기술이다. Ethos-U 시리즈는 초저전력 AI 연산을, Ethos-N 시리즈는 고성능 AI 애플리케이션을 지원한다. 하드웨어 가속을 통해 저전력 환경에서도 높은 연산 성능을 유지한다. 스마트폰 및 IoT 기기에서 배터리 수명을 연장하고 AI 에너지 효율성을 높인다.

### VI. 지속 가능한 AI 환경 구축을 위한 전략

#### 1. AI 연산의 에너지 효율 최적화

AI 연산의 에너지 효율을 극대화하려면 모델 경량화, 저전력 하드웨어 활용, 엣지 및 온-디바이스 AI 활성화가 필수적이다. 생성형 AI의 연산량과 전력 소비 증가로 운영 비용 부담이 커지면서, 소프트웨어 최적화가 가장 중요한 해결책으로 부상하고 있다.

이를 위해 양자화, 프루닝, 지식 증류 등의 기법을 활용하여 모델 크기를 줄이고 연산량을 최적화할 수 있다. 또한, 희소연산, MoE 구조 등을 통해서 불필요한 연산을 최소화할 수 있다.

하드웨어 측면에서는 TPU, Dojo, ARM Ethos NPU, NVIDIA Grace Hopper 등 저전력 AI 가속기 및 PIM(Processing In Memory) 기술을 활용해 전력 소모를 줄일 수 있다. 광 기반 AI 가속기(Photonic AI Accelerators)는 기존 반도체보다 낮은 전력으로 고성능 연산이 가능하다.

또한, AI 연산을 클라우드에서 엣지 및 온-디바이스로 이동하면 네트워크 에너지 절감과 실시간 AI 처리가 가능해진다. 온-디바이스 AI는 대규모 데이

터센터 의존도를 줄이고 전력 소비를 감소시켜 지속 가능한 AI 운영을 가능하게 한다.

결과적으로, AI의 에너지 최적화를 위해서는 모델 경량화, 연산 최적화, 저전력 하드웨어 활용, 엣지 및 온-디바이스 AI 활성화가 핵심 전략이 될 것이다.

#### 2. 친환경 AI 인프라 및 데이터센터 구축

친환경 AI 인프라 구축을 위해서는 재생 에너지 활용, 탄소 배출 절감 기술, 고효율 냉각 시스템 도입 적용이 필수적이다. AI 학습과 추론을 위한 데이터센터는 높은 에너지 소비와 탄소 배출을 초래하므로 태양광, 풍력 등 친환경 에너지원으로 전력 공급을 전환하는 것이 필요하다. 또한, AI 시스템의 탄소 발자국을 실시간으로 모니터링하고 최적화하는 솔루션을 적용하여 지속 가능한 AI 인프라를 구축해야 한다.

### VII. 결론

본고에서는 생성형 AI의 확산과 함께 증가하는 에너지 소비 문제를 분석하고, 이를 해결하기 위한 하드웨어 및 소프트웨어 최적화 기술을 살펴보았다. AI 모델이 점점 더 대형화되고 연산량이 증가함에 따라 AI 에너지 소비 최적화는 AI 산업의 지속 가능성을 결정짓는 핵심 과제가 되고 있다.

AI 기술의 발전이 가져오는 혁신적인 변화는 앞으로도 가속화될 것이며, AI가 인류와 공존하기 위해서는 효율적인 에너지 관리와 친환경적인 AI 인프라 구축이 필수적이다. AI 시스템이 환경과 조화를 이루며 지속 가능성을 확보하기 위해서는 기술 개발뿐만 아니라 정책적 지원과 산업 전반의 협력이 필수적이다. AI의 발전과 지속 가능성 간의 균형

을 맞추는 것이 미래 AI 산업의 경쟁력을 결정짓는 중요한 요소가 될 것이다.

약어 정리

AI	Artificial Intelligence
AIOps	Artificial Intelligence for IT Operations
ASIC	Application-Specific Integrated Circuit
CoWoS	Chip on Wafer on Substrate
CPU	Central Processing Unit
ESS	Energy Storage System
FFN	Feed Forward Network
FPGA	Field Programmable Gate Array
GPU	Graphics Processing Unit
HBM	High Bandwidth Memory
HPC	High-Performance Computing
LLM	Large Language Model
MoE	Mixture of Experts
MWh	Megawatt-hour
NPU	Neural Processing Unit
PIM	Processing In Memory
RLE	Research Laboratory of Electronics
SoIC	System on Integrated Chips
TPU	Tensor Processing Unit

용어해설

**Training, Inference** ML/DL 기반 인공지능 기술은 훈련(Training)과 추론(Inference) 두 과정으로 이루어짐. 훈련은 사전에 수집된 데이터를 학습하여 모델의 로직을 생성하는 과정이며, 추론은 학습된 모델을 활용하여 새로운 입력 데이터를 처리하고 예측이나 판단을 수행하는 과정임

**AIOps** AI 기술을 활용하여 IT 운영을 자동화하고 최적화하는 기술임. 대량의 운영 데이터를 실시간으로 분석하여 이상 탐지, 문제 해결, 성능 최적화 등을 수행함으로써 IT 시스템의 안정성과 효율성을 향상시킴

참고문헌

[1] Gartner, "Emerging Tech: Top Trends in Energy-Efficient Generative AI Compute Systems," Gartner Report, ID

G00807709, 2024, pp. 1-12.

[2] J. Howarth, "Number of Parameters in GPT-4 (Latest Data)," Exploding Topics, 2025. 2. 24. <https://explodingtopics.com/blog/gpt-parameters>.

[3] TRG Datacenters, "AI Chatbots: Energy usage of 2023's most popular chatbots(so far)," TRG Datacenters, 2023. <https://www.trgdatacenters.com/resource/ai-chatbots-energy-usage-of-2023s-most-popular-chatbots-so-far/>

[4] Gartner, "Emerging Tech: Strategies to Achieve Energy Efficiency Goals for GenAI," Gartner Report, ID G00809640, 2024, pp. 1-7.

[5] D. Lepikhin et al., "GShard: Scaling Giant Models with Conditional Computation and Automatic Sharding," arXiv preprint, 2020. doi: 10.48550/arXiv.2006.16668

[6] D. Dai et al., "DeepSeekMoE: Towards Ultimate Expert Specialization in Mixture-of-Experts Language Models," in Proc. Annu. Meeting Assoc. Comput. Linguistics., (Bangkok, Thailand), vol. 1, Aug. 2024, pp. 1280-1297.

[7] Datature, "A comprehensive guide to neural network model pruning," Datature Blog, 2024. 2. 29. <https://www.datature.io/blog/a-comprehensive-guide-to-neural-network-model-pruning>.

[8] B. Clark, "What is Quantization?," IBM Think Blog, 2024. 7. 29. <https://www.ibm.com/think/topics/quantization>.

[9] J. Gou et al., "Knowledge Distillation: A Survey," arXiv preprint, 2020. doi: 10.48550/arXiv.2006.05525

[10] G. Hinton et al., "Distilling the Knowledge in a Neural Network," arXiv preprint, 2015. doi: 10.48550/arXiv.1503.02531

[11] B.-S. Chu and C.-R. Lee, "Low-rank Tensor Decomposition for Compression of Convolutional Neural Networks Using Funnel Regularization," arXiv preprint, 2021. doi: 10.48550/arXiv.2112.03690

[12] D. S. Hochbaum and P. Baumann, "Sparse Computation for Large-Scale Data Mining," IEEE Trans. Big Data, vol. 2, no. 2, 2014, pp. 151-174.

[13] A. Dhasade et al., "Energy-Aware Decentralized Learning with Intermittent Model Training," arXiv preprint, 2024. doi: 10.48550/arXiv.2407.01283

[14] Gartner, "Hype Cycle for Open-Source Software, 2024," Gartner Report, ID G00811366, 2024, pp. 1-67.

[15] S. Huang and W. Feng, "Energy-Efficient Cluster Computing via Accurate Workload Characterization," in Proc. IEEE/ACM Int. Symp. Cluster Comput. Grid, (Shanghai, China), May. 2009, pp. 68-75.

[16] B. Wang, "Tesla Dojo Supercomputers and AI," Next BigFuture, 2021. 9. 17. <https://www.nextbigfuture.com/2021/09/tesla-dojosupercomputers-and-ai.html>.

- [17] ARM, "Ethos-U85," <https://www.arm.com/products/silicon-ip-cpu/ethos/ethos-u85>
- [18] Lightelligence, "PACE: Photonic Arithmetic Computing Engine," <https://www.lightelligence.ai/index.php/product/index/2.html>.
- [19] Intel, "Neuromorphic Computing," <https://www.intel.co.kr/content/www/kr/ko/research/neuromorphic-computing.html>.
- [20] U. Hölzle, "Google achieves four consecutive years of 100 percent renewable energy," Google Cloud Blog, 2021. 4. 21. <https://cloud.google.com/blog/ko/topics/inside-google-cloud/google-achieves-four-consecutive-years-of-100-percent-renewable-energy>
- [21] Gartner, "Emerging Tech: Harness Infrastructure Analytics to Rethink the GenAI Power Problem," Gartner Report, ID G00807608, 2024, pp. 1-11.
- [22] W. Fedus et al., "Switch Transformers: Scaling to Trillion Parameter Models with Simple and Efficient Sparsity," arXiv preprint, 2021. doi: 10.48550/arXiv.2101.03961