AI 반도체와 주요 응용 분야에 관한 연구

(A Comprehensive Study of AI Semiconductors and Key Application Fields)

권영화*, 김보영**
(Young Hwa Kwun, Bo Young Kim)

요 약

본 논문은 최근 주목받는 AI 반도체의 종류와 주요 특징, 그리고 이를 개발하는 주요 기업들을 조사하고 AI 반도체가 사용되는 응용 분야를 분석하였다. 데이터센터에서는 GPU와 TPU가 주로 훈련 분야에 활용되며, 추론 분야에서는 GPU, NPU, ASIC이 사용된다. 스마트폰에는 GPU와 NPU가 통합된 AP가 일반적으로 사용되며, 자율주행차와 협동 로봇 역시 AP 내부에 GPU와 NPU/DLA가 결합된 형태로 사용되고 있다. 특히, 스마트폰과 자율주행차는 하이브리드 형태로 AI가 활용되지만 협동 로봇은 데이터센터 연결 없이도 독립적인 AI 연산의 수행이 가능하다. 나아가 본 논문은 AI 반도체에 대한 주요 현황과 응용 분야를 비교 분석함으로써 향후 기술 트렌드를 알 수 있어 업계와 학계에 큰 시사점을 주고 있다.

■ 중심어 : AI 반도체 ; 데이터센터 ; 스마트폰 ; 자율주행차 ; 협동 로봇

Abstract

This paper investigates the types and key characteristics of AI semiconductors that have recently gained attention, as well as the major companies developing them, and further analyzes their application areas. In data centers, GPUs and TPUs are predominantly used for training, while GPUs, NPUs, and ASICs are applied for inference. In smart phones, application processors (APs) integrating GPUs and NPUs are commonly adopted, and in autonomous vehicles and collaborative robots, APs combining GPUs with NPUs/DLAs are also utilized. Notably, while smartphones and autonomous vehicles employ AI in hybrid forms that combine on-device and cloud processing, collaborative robots are capable of performing independent AI computation without data center connectivity. Furthermore, by comparatively analyzing the current status and application domains of AI semiconductors, this paper provides valuable insights into future technology trends for both industry and academia.

■ keywords : AI semiconductor ; Data center ; Smart phone ; Autonomous vehicle ; Collaborative robot

Ⅰ. 서 론

최근 AI(Artificial Intelligence) 시장이 급격하게 커지면서 AI에 대한 관심이 증가하고 있다. AI의 출현은 다양한 복잡한 문제를 해결하는 데 큰 가능성을 보여주고 있다. [1] 특히, AI는 특정 범주의 문제들에 있어 이미 인간의 능력을 능가할 수 있다. [2] 나아가 AI는 다양한 산업 전반에 걸쳐 변혁을 일으키는 힘으로도 부상하고 있다. [3] AI 시장이 커짐에 따라 AI 반도체의 사용도 급격하게 늘어나고 있지만 아직 AI 반도체의 개발 역사가 그리 길지 않기에 AI 반도체에 대한 연구는 그리 많은 편이 아니다.

이에 따라 AI 반도체에 대한 정의는 아직 학계에서 명확히 정립되지 않았다. [4] 하지만 일반적으로 AI 반도체는 인공지능 알고리즘을 위한 연산을 수행하기 위해 설계된 칩이라 할 수 있다 [5]. 그리고 이런 AI 반도체를 사용하여 2013년부터 다양한 엣지 디바이스(Edge Device)가 출시되었다. [6] 하지만 아직까지 AI 반도체와 응용 분야에 대한 연구는 거의 전무하기에 본 연구가 업계와 학계에 큰 의미

* 정회원, 서울과학종합대학원 AI융합공학 박사과정
** 정회원, 서울과학종합대학원 교수

접수일자 : 2025년 08월 08일
수정일자 : 2025년 09월 12일
게재확정일 : 2025년 09월 12일
교신저자 : 김보영 e-mail: bykim2@assist.ac.kr

가 있을 것으로 생각한다.

본 연구는 AI 반도체의 주요 유형과 그 특성을 분석하고, 응용 분야별 차별점을 비교 분석하는 것을 목적으로 한다. 이를 위해 AI 반도체의 종류, 그리고 국내외 주요 기업들의 AI 반도체 개발의 현황을 조사해 보도록 한다. 나아가 AI 반도체가 사용되고 있는 주요 응용 분야뿐만 아니라 각 응용 분야에서 어떤 AI 반도체가 사용되고 있는지, 그리고 각 응용 분야별로 어떤 차이점이 있는지 비교하여 분석해 보기로 한다. 아래의 표 1은 본 논문의 전체적인 내용을 보여주고 있다.

표 1. 논문의 내용 요약

구분	주요 내용
AI 반도체 종류	GPU, FPGA, ASIC, NPU, 뉴로모픽 반도체
AI 반도체 시장 현황	1. 데이터센터에서 온디바이스 AI로 확장됨에 따라 GPU에서 NPU/ASIC으로 확대 2. 다양한 응용 분야에 AI 반도체 적용 추세
AI 반도체 개발 기업	1. 엔비디아 (B200) 2. AMD (CDNA4) 3. 구글 (Tensor G5) 4. 테슬라 (A15) 5. 리벨리온 (REBEL-Quad) 6. 퓨리오사AI (RNGD) 7. 딥엑스 (DX-M1)
AI 반도체의 응용 분야 및 공급기업	1. 데이터센터: 엔비디아, AMD, 브로드컴, 마벨, 리벨리온 (GPU, ASIC, NPU) 2. 스마트폰: 애플, 삼성전자, 퀄컴, 미디어텍 (AP) 3. 자율주행차: 엔비디아, 퀄컴 (AP) 4. 협동 로봇: 엔비디아, 퀄컴 (AP)
데이터센터 연결	1. 데이터센터: 필수 2. 스마트폰: 하이브리드 3. 자율주행차: 하이브리드 4. 협동 로봇: 불필요

Ⅱ. 본 론

1. AI 반도체의 현황

가. GPU

주로 게임용으로 사용되던 GPU(Graphic Processing Unit)는 원래 AI를 위해 만들어진 반도체는 아니었다. 하지만 GPU는 AI 분야에서 병렬 연산에 강점을 보여, 현재까지도 대표적인 AI 반도체로 활용되고 있다. CPU와 달리 병렬적으로 연산을 빠르게 처리할 수 있는 GPU의 기능이 AI 분야에 적합하였기 때문이다. 하지만 GPU는 메모리 반도체와 수많은 데이터를 교환할 때 전력의 소모가 너무 많다. [7] 그리고 가격도 매우 비싸다는 치명적인 단점이 존재한다. 따라서 데이터센터의 추론 분야에는 순수하게 AI 기능에만 충실한 반도체에 대한 수요가 점차적으로 커지고 있다.

나. FPGA

FPGA(Field Programmable Gate Array)는 사용자가 자신들의 필요에 맞게 설계하여 사용할 수 있도록 만든 반도체이다. 최근 FPGA는 AI 칩의 한 종류로서 더 많은 주목을 받고 있다. [8] 사용자가 필요에 맞게 설계와 재설계를 할 수 있기에 자신들이 필요한 기능만을 넣어 만들 수 있는 특징이 있기 때문이다. 특히, FPGA는 가장 유연한 재구성이 가능하다는 것이 장점이다. [9] 그리고 FPGA는 비교적 낮은 에너지 소비로 높은 성능을 제공한다. [10] 최근 대부분의 빅테크 기업은 자신들의 제품이나 서비스의 필요에 맞게 FPGA를 설계한 후 데이터센터(Data Center)에도 적극적으로 사용하고 있다.

다. ASIC

ASIC(Application Specific Integrated Circuit)은 특정한 응용 분야에 맞게 고정된 기능을 수행할 수 있게 만든 AI 반도체이다. 이런 이유로 인해 ASIC을 주문형 반도체라고 부르기도 한다. 앞서 설명한 FPGA와 달리 설계에 대한 유연성은 크게 부족하지만 대량으로 생산할 경우 단가를 크게 낮출 수 있는 장점이 있다. 나아가 ASIC은 성능과 전력 효율이 비교적 우수한 편에 속한다. 최근 데이터센터의 추론 분야에서 ASIC이 주목을 받고 있다. ASIC을 만들고 있는 유명한 반도체 기업은 브로드컴(Broadcom)과 마벨 테크놀로지(Marvell Technology) 등이 있다.

라. NPU

NPU(Neural Processing Unit)는 인공지능에 특화되어 설계된 반도체이다. 현재 NPU는 AI 반도체

의 발전에 있어 최전선에 있다. [11] NPU는 GPU와 달리 인공지능의 기능만으로 설계되었기 때문에 전력 소모가 상당히 낮고 가격이 비교적 저렴한 편이다. 그리고 NPU는 데이터센터와 온 디바이스 AI(On-device AI)에 모두 사용된다. 최근 데이터센터의 추론 부분에서 사용이 조금씩 늘어나고 있으며 국내외 많은 NPU 기업이 생겨나고 있다. 아울러 구글에서 만들고 있는 TPU(Tensor Processing Unit)도 NPU의 일종이다. 나아가 NPU는 온 디바이스 AI의 경우 주로 스마트폰에 많이 사용되고 있으며, 그 외 다른 다양한 디바이스에도 사용이 늘어나고 있는 상황이다. 특히 온 디바이스 AI에 NPU가 독립적으로 쓰이기보다 SoC(Sytem on Chip) 안에 들어가는 경우가 많다.

마. 뉴로모픽 반도체

뉴로모픽(Neuromorphic) 반도체는 인간의 두뇌를 모방하여 만든 반도체이다. AI 연구의 장기적인 목표는 인간의 뇌를 실리콘과 소프트웨어로 시뮬레이션(Simulation)하는 것이다. [12]

인간의 두뇌는 20W 정도의 적은 에너지로 각종 연산과 기억을 동시에 빠르게 처리한다. 마찬가지로 뉴로모픽 반도체도 적은 에너지로 시스템 반도체와 메모리 반도체 기능을 한 개의 칩에서 동시에 처리할 수 있도록 만든 가장 이상적인 반도체이다. 최근 데이터 양의 급격한 증가로 메모리 병목 현상이 심화되고 있다. 이와 같은 문제를 해결하기 위해 다양한 기업에서 뉴로모픽 반도체에 대한 연구를 지속적으로 진행하고 있다.

2. AI 반도체 시장

AI 반도체는 지금의 AI 시대를 가져올 수 있었던 중요한 동인이다. AI 반도체가 개발되지 않았다면 지금과 같은 AI 시대의 개막은 불가능하였을 것이기 때문이다. 그럼 기존 반도체와 AI 반도체의 차이점은 무엇인지 조사해 보도록 한다.

AI 반도체는 기존 반도체와 달리 AI 기능을 효율적으로 수행할 수 있도록 설계된 칩이라는 것이 특징이다. MAC(Multiply Accumulate Computing)이라는 데이터 처리방식으로 인공지능 연산을 빠르게 수행할 수 있다. 특히, AI 반도체는 병렬 연산에 최적화되어 있다. 그리고 AI 모델 학습과 추론은 대량의 행렬과 벡터 연산으로 이루어진다. 따라서 AI 반도체는 수천에서 수만 개의 코어를 병렬로 배치하여 기존 CPU보다 훨씬 높은 연산 밀도를 제공한다는 특징이 있다. 이에 따라 GPU, TPU(Tensor Processing Unit), NPU는 SIMD(Single Instruction, Multiple Data) 기반으로 구성되어 하나의 명령어로 여러 개의 데이터를 동시에 처리하는 구조이다.

기존 CPU와 같은 반도체는 성능은 매우 우수하지만 데이터를 직렬로 처리하기 때문에 많은 데이터를 동시에 빠르게 처리하는 데 한계가 있다. 하지만 AI 반도체는 병렬로 데이터를 처리하기 때문에 데이터를 동시에 빠르게 처리할 수 있는 특징이 있다. 그 외에도 AI 기능에 특화된 AI 반도체는 전력의 소모가 적어 지금과 같이 데이터센터의 냉각 문제가 심각해지고 있는 상황에서 대안으로 떠오르고 있다.

표 2. 기존 반도체와 AI 반도체의 차이점

구분	기존 반도체 (CPU)	AI 반도체 (NPU/TPU/ASIC)
설계 목적	범용 처리	AI 연산 특화
연산 구조	직렬·일반 연산 최적화	행렬·벡터 병렬 연산 최적화
코어 수	수 개-수십 개	수천-수만 개
메모리	DRAM/GDDR 기반	HBM (대용량 메모리)
전력 효율	낮음-보통	매우 높음 (TOPS/W 기준)
패키징	전통적 SoC	2.5D/3D, 칩렛, HBM 적층
소프트웨어	범용 OS/컴파일러	AI 프레임워크 + 전용 SDK

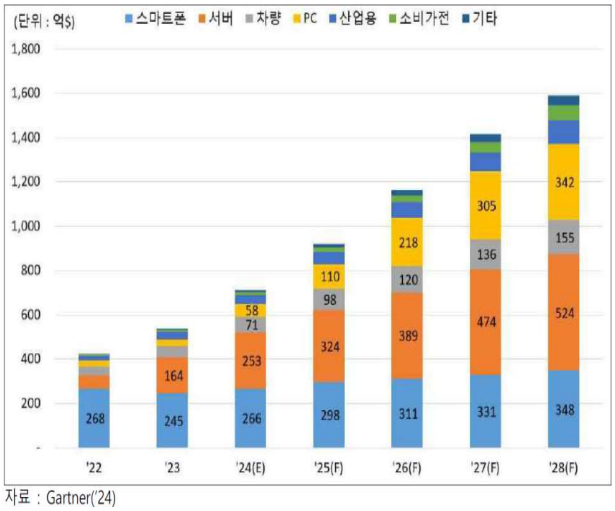
AI 반도체는 지속적으로 개발이 진행되고 있으며 기존 GPU 중심에서 점차적으로 NPU와 ASIC으로 확대되고 있는 상황이다. 그리고 인간의 두뇌와 같이 에너지 소모는 적으면서 메모리 병목을 없앨 수 있는 방향으로 AI 반도체의 개발이 진행되고 있다. 아울러 지금도 속도가 빠르고 성능이 우수한 AI 반도체를 개발하기 위해 많은 기업이 노력을 기울이고 있다. 특히, 구글(Google), 아마존(Amazon), 애플(Apple), 마이크로 소프트(Microsoft), 바이두

(Baidu), 알리바바(Alibaba) 등과 같은 빅테크(Big tech) 기업들은 자체적으로 AI 반도체를 개발하고 있다는 점에서 주목을 끌고 있다. 이들 기업이 AI 반도체를 개발하는 이유는 자사의 제품이나 서비스에 자신들이 개발한 AI 반도체를 사용함으로써 반도체 기업에의 의존을 줄이면서 가격을 낮출 수 있을 뿐만 아니라 자사의 제품이나 서비스를 최적화시킬 수 있기 때문이다. 아울러 코로나19 시기에 대두가 되었던 공급망 문제에서도 벗어날 수 있다는 장점도 있다.

하지만 AI 반도체의 사용은 비교적 최근의 일이기 때문에 AI 반도체의 개발은 아직 초기 단계라 할 수 있다. 따라서 어느 기업이 우수한 AI 반도체를 먼저 개발하고 대량으로 양산할 수 있느냐에 따라 AI 시장의 주도권을 가져갈 수 있을 것으로 보고 있다. 특히, AI 반도체는 기존의 데이터센터 중심에서 점차적으로 온 디바이스 AI 분야로 확대되고 있기 때문에 GPU 중심에서 NPU와 ASIC으로 확대되고 있는 상황이다. 물론 아직까지 온 디바이스 AI 시장은 그리 크지 않지만 시간이 지날수록 폭발적으로 성장할 수 있을 것으로 전망하고 있다. Dhruvitkumar V. Talati도 AI 반도체의 범주를 클라우드와 네트워크 엣지로 구분하였는데 네트워크 엣지는 우리가 인터넷에 연결되어 있지 않은 상태에서도 다양한 디바이스에서 AI 기능을 사용할 수 있는 온 디바이스 AI와 같은 개념이다. [13]

한편, 앞으로 AI 반도체의 수요는 더욱 늘어날 것으로 보인다. AIoT(AI of things)와 같이 모든 사물에 AI가 적용될 것이 거의 분명하기 때문이다. 이에 따라 앞으로 10년 내 AI 반도체가 전체 반도체의 1/3일 차지하게 될 것이라는 전망도 나오고 있다.

결론적으로 현재 AI 반도체는 다양한 응용 분야로 사용이 확대되고 있는 상황이다. 데이터센터를 중심으로 스마트폰, 스마트홈, 스마트 팩토리, 디지털 헬스케어, 감시카메라, 로봇, 자율주행차 등에 적용되고 있으며 앞으로도 더욱 다양한 디바이스에 확대되어 적용될 것으로 보고 있다.



자료 : Gartner('24)

그림 1. 글로벌 AI 반도체 시장 현황 및 전망

3. 국내외 AI 반도체의 개발 상황

가. 엔비디아 B200

엔비디아가 개발을 완료한 가장 최신의 AI 반도체는 B200이다. B200은 H100보다 질의응답 테스트(LLM Q&A)에서 성능이 2.5배 높다고 알려져 있다. 이 제품은 블랙웰(Blackwell)에 탑재되고 있으며 블랙웰 울트라(Blackwell Ultra)에는 B300이 들어갈 예정이다. 이미 알려진 바와 같이 엔비디아는 AI 반도체의 개발 주기를 2년에서 1년으로 앞당기면서 시장에서 독점적인 지위를 누리고 있는 중이다. 특히, 엔비디아는 쿠다(Compute Unified Device Architecture) 라는 소프트웨어 생태계를 통해 고객이 손쉽게 AI를 개발할 수 있도록 하면서 시장을 리드해 나아가고 있다.

나. AMD CDNA4

AMD가 올해 출시한 CDNA4는 MI350X에 들어가는 GPU로 TSMC의 3nm 공정으로 제조되었다. 성능은 엔비디아 B200보다 약간 높거나 같은 수준이지만 가격은 상당히 저렴한 것으로 알려져 있다. 이 칩은 AI와 HPC(High Performance Computing)를 위해 고성능 연산과 에너지 효율에 중점을 두고 개발이 되었다. CDNA3보다 연산 성능

은 4배가 높아진 것으로 알려져 있기에 앞으로 기대가 되고 있다. 특히, 최근 AMD의 시장점유율이 조금씩 올라가고 있는데 그 이유는 AI 시장의 수요가 지속적으로 커지고 있기 때문이다.

다. 구글 Tensor G5

구글의 Tensor G5는 자사의 스마트폰인 픽셀 10 프로 폴드(Pixel 10 Pro Fold)에 들어가는 AI 반도체로 TSMC의 3nm 공정으로 제조가 되었다. Tensor 시리즈는 삼성 파운드리에서 제조되었지만 Tensor G5는 TSMC에서 제조가 되었다. 그동안 구글은 하드웨어 측면에서 발열과 전성비가 만족스럽지 않은 경우가 많아 좋은 평가를 받지 못하였다. 하지만 이번에 TSMC의 팹으로 바꾸면서 성능이 크게 향상된 것으로 보인다. 이외에도 구글은 자사 클라우드에 들어가는 AI 반도체도 자체적으로 만들어 사용하고 있는 것으로 알려져 있다.

라. 테슬라 A15

A15는 자율주행에 쓰이는 AI 반도체로 TSMC의 4nm 공정에서 생산된다. 2,500 TOPS(Tera Operation Per Second)의 성능을 보유하고 있으며 주요 목적은 테슬라의 자율주행차와 로봇에 탑재하는 것이지만 데이터센터의 추론 부분에도 사용이 가능하다. 특히, 차량에 탑재로 인해 발열이나 전력 소모를 줄일 수 있도록 최적화되어 설계된 특징이 있다. 그리고 A15는 SoC로 제작이 되고 내부에 GPU와 NNA(Neural Net Processor)가 들어간다. NNA는 레이더 신호와 카메라 이미지 등을 센서로부터 받은 후 자율주행을 위한 판단을 하고 추론을 하게 된다.

마. 리벨리온 REBEL-Quad

리벨리온은 설립한 지 약 5년 정도 된 국내 AI 반도체 스타트업(Start-up)이다. 작년 SK텔레콤의 자회사인 사피온코리아와 합병이 되었다. 설립 후 아이온(Ion)과 아톰(Atom)을 출시하였고 올해 REBEL-Quad를 개발하였다. REBEL은 삼성 파운드리의 4nm 공정으로 제조가 되고 주로 데이터센터의 추론 분야에 사용되고 있다. 특히, REBEL은 GPU 대비 전력 소모를 최대 1/5로 줄일 수 있는 것과 비용이 적게 든다는 것이 큰 장점이다. 나아가 REBEL은 칩렛(Chiplet) 기반으로 HBM3E가 탑재되는 것이 주요 특징이라 할 수 있다. [14]

최근 데이터센터에서 추론 분야의 수요가 조금씩 증가하고 있기 때문에 앞으로 REBEL의 수요도 점차적으로 늘어날 것으로 보인다.

바. 퓨리오사AI RNGD

퓨리오사AI도 2017년 설립된 국내 스타트업이다. 최근 퓨리오사AI는 메타(Meta)로부터 8억 달러에 인수 제안을 받았으나 거절한 것으로 알려져 있다. 2021년 처음으로 14nm 공정으로 워보이(Warboy)를 출시한 적이 있다. 최근 출시한 칩은 RNGD(Renegade)로 TSMC의 5nm 공정으로 제조가 된다. 마찬가지로 RNGD도 데이터센터의 추론 분야에 특화한 제품이다. [15] RNGD의 클럭 속도는 1.0GHz이며 HBM3가 들어가 리벨리온보다는 성능이 떨어질 수 있지만 전력 소모는 GPU대비 2배 이상 줄일 수 있다. 최근 RNGD는 LG AI Research의 EXAONE LLM 플랫폼에 탑재되었다.

사. 딥엑스 DX-M1

딥엑스도 설립된 지 얼마 되지 않은 국내 스타트업이다. DX-M1은 삼성 파운드리의 5nm 공정으로 제조되고 있는 엣지 디바이스용의 NPU이다. 리벨리온과 퓨리오사AI와 같이 데이터센터의 추론용 NPU도 만들고 있지만 DX-M1은 엣지 디바이스 분야에 특화되어 사용된다는 것이 차이점이다. AI 성능은 25 TOPS이고 TDP(Thermal Design Power)는 3-5W 정도로 성능은 좋은 반면 전력 소모는 매우 적은 것이 특징이다. 뿐만 아니라 다른 칩

들과 비교해도 성능 측면에서 세계적인 수준이다. DX-M1이 사용될 수 있는 분야는 IoT, 스마트 카메라, 로봇 등으로 매우 다양하다.

표 3. 국내외 주요 기업들의 최근 AI 반도체 개발 현황

제품명	제조사	공정	성능	개발 상황	응용 분야	특징
B200	엔비디아	4nm	1,038 TPS/User (Llama 4 기반)	2024년 개발	데이터 센터	
CDNA4	AMD	3nm	MI355X DLC(128 GPU 랙): FP4 기준 약 2.6 ExaFLOPS	2025년 개발	데이터 센터	
Tensor G5	구글	3nm	CPU: +30~39% (전 세대 대비) GPU: +27% (전 세대 대비) TPU: +60% (Gemini Nano 2.6× 속도, 2× 효율)	2025년 개발	스마트폰	자사용
A15	테슬라	4nm	2,000-2,500 TOPS	2025년 개발	자율주행차	자사용
REBEL-Quad	리벨리온	4nm	2 PFLOPS (FP8, 4칩렛 구성)	2025년 개발	데이터 센터	추론용
RNGD	퓨리오사AI	5nm	FP8 512 TFLOPS / BF16 256 TFLOPS / INT8 512 TOPS / INT4 1024 TOPS	2024년 개발	데이터 센터	추론용
DX-M1	딥엑스	5nm	25 TOPS	2024년 개발	엣지 디바이스	

4. AI 반도체의 주요 응용 분야

위에서 살펴본 바와 같이 AI 반도체가 주로 많이 사용되고 있는 분야는 데이터센터이다. 다음은 온디바이스 AI라고도 하는 엣지 디바이스이다. 이는 AI 기능을 사용하기 위해 굳이 데이터센터를 이용할 필요 없이 자체 디바이스에서 가능한 것을 의미한다. 최근 엣지 AI 컴퓨팅 애플리케이션의 수요가 크게 증가되고 있으며 [16] 엣지 컴퓨팅은 다음 세대의 지능형 기술을 이끌 것으로 보인다. [17]

최근 스마트폰에서도 AI 기능이 가능한데 이는 온디바이스 AI가 본격적으로 시작되고 있는 것을 의미한다. 굳이 인터넷에 접속되어 있지 않더라도 NPU를 통해 AI 기능을 즐길 수 있는 것이다. 엣지 디바이스는 스마트폰뿐만 아니라 스마트 글래스, 가전제품과 로봇 등 다양한 형태가 있으며 앞으로 그 종류는 더욱 다양해질 것으로 보인다. 그럼 AI 반도체가 사용되는 주요 응용 분야에 대해 자세히 분석해 보도록 한다.

가. 데이터센터

데이터센터는 많은 서버로 이루어진 각종 데이터를 저장하는 대규모 공간이다. 과거에는 텍스트 데이터가 주류를 이루고 있었지만 최근에는 이미지, 사진, 음성과 동영상 데이터 등이 매일 엄청난 규모로 생산되고 있다. 나아가 각종 데이터의 양도 기하급수적으로 증가하고 있다. 이와 같은 상황에서 다양한 데이터를 저장할 수 있는 데이터센터의 수요가 급격하게 늘어나고 있다. 이에 따라 많은 빅테크 기업이 경쟁적으로 새로운 데이터센터의 건설과 확장에 나서고 있는 상황이다. 그리고 데이터센터에는 많은 AI 반도체가 필요하기 때문에 AI 반도체 수요도 크게 늘어나고 있는 것이다.

나아가 오픈 AI의 챗 GPT, 구글의 제미나이(Gemini)와 메타의 라마(LLaMA) 같은 LLM(Large Language Model)은 수천억에서 수조 개의 파라미터를 학습하고 추론해야 한다. 기존 CPU 기반 데이터센터로는 연산 속도와 전력 효율이 크게 부족할 수밖에 없다. 따라서 GPU와 NPU 같은 AI 반도체 없이는 대규모 AI 서비스의 운영이 불가능하다. 특히, 데이터센터에서 AI 반도체는 대규모 병렬 연산, 초고속 메모리, 전력 효율, 확장성, 맞춤형 아키텍처라는 특징을 가지고 있다. 결과적으로 AI 반도체는 기존 서버용 반도체와 달리 AI 슈퍼 컴퓨팅의 엔진 역할을 하고 있는 것이다.

한편, 데이터센터는 크게 훈련(Training) 분야와 추론(Inference) 분야로 나눌 수 있는데 현재 양쪽 분야 모두 엔비디아의 GPU가 시장의 90% 가까이 장악하고 있다. 특히, 엔비디아는 AI를 개발하기 위해 반드시 필요한 쿠다를 중심으로 AI 생태계를 장악하고 있기 때문에 향후 적어도 몇 년간 AI 시장을 리드해 나아갈 것으로 전망하고 있다. 앞으로도 많은 빅테크 기업이 경쟁적으로 데이터센터를 건설할 예정이어서 GPU 수요는 크게 늘어날 전망이다. 나아가 국가마다 소버린 AI(Sovereign AI)를 강화하려는 움직임도 확산되고 있기에 GPU 수요는 더욱 증가될 것으로 보인다.

한편, 많은 반도체 기업이 엔비디아의 데이터센터 시장을 빼앗아 오기 위해 다양한 노력을 기울이고 있다. 특히, 최근 AMD는 엔비디아의 B200에 대응할 수 있는 CDNA4를 출시하였다. 그리고 AMD는 엔비디아의 쿠다를 대체할 수 있는 ROCm(Radeon Open Compute)이라는 AI 개발을 위한 소프트웨어로 AI 생태계를 조성하고 있다. AMD의 제조공정은 3nm로 엔비디아의 4nm보다 앞서고 있기 때문에 성능 측면에서 유리하다. 특히, 가격 측면에서도 AMD는 상당한 이점이 있는 것으로 알려져 있다. 그 외 다른 기업들은 훈련 분야에서 엔비디아와 경쟁할 수 있는 기업은 아직까지 거의 없는 것으로 보인다. 하지만 구글, MS, 아마존과 같은 클라우드 빅테크 기업들은 엔비디아의 의존을 줄이기 위해 추론 분야뿐만 아니라 훈련 분야에도 TPU와 같은 AI 반도체를 자체적으로 개발하여 자사의 클라우드에 사용하고 있다. 그리고 훈련 분야에서는 엔비디아 GPU의 대안으로 사용될 수 있는 AI 반도체는 앞으로도 상당 기간 나오기 쉽지 않을 것이다. 엔비디아의 GPU가 훈련 분야에서는 매우 적합한 AI 반도체로 볼 수 있기 때문이다.

반면 추론 분야에서는 많은 기업이 NPU와 ASIC을 만들어 엔비디아의 GPU를 대체하려 노력하고 있다. 엔비디아의 GPU가 고사양이어서 가격이 너무 비쌀 뿐만 아니라 전력 소모가 매우 크다는 약점이 있기 때문이다. ASIC을 만드는 기업은 브로드컴(Broadcom)과 마벨 테크놀로지(Marvell Technology)가 대표적이다. 이들 기업은 고객이 원하는 사양의 추론용 ASIC을 만들어 주는 비즈니스로 큰 성과를 내고 있다. 그리고 NPU를 만드는 기업들도 지속적으로 늘어나고 있다. 특히, 국내에서도 NPU를 만들고 있는 몇 개의 스타트업이 있다. 이들 기업은 엔비디아의 GPU가 사용되는 데이터센터의 추론 분야를 자신들의 칩으로 대체시키기 위해 노력을 기울이고 있다. 그것이 가능한 이유는 NPU가 인공지능만을 위해 설계된 칩이기 때문에 GPU에 비해 전력 소모와 가격 측면에서 많은 이점이 있기 때문이다. 그렇다 하더라도 엔비디아는 단순히 칩만 만드는 기업이 아니기 때문에 단기간 내 이들 기업에 지위를 내줄 것 같지는 않다. 특히, 엔비디아는 GPU뿐만 아니라 쿠다, 각종 소프트웨어, NV링크, 인피니밴드(InfiniBand), DGX 서버 같은 각종 제품과 서비스를 풀스택 솔루션(Full Stack Solution)으로 제공하고 있기 때문에 고객의 입장에서는 매우 편리하다. 그렇다 보니 고객들은 다른 벤더(Vendor)로 쉽게 바꾸기 어려운 상황이다. 물론 어느 기업도 영원히 독점은 할 수 없지만 아직까지 엔비디아의 생태계는 매우 강해 보인다.

나. 스마트폰

스마트폰은 생활에 없어서는 안되는 필수품으로 우리의 생활을 매우 편리하게 만들어 주고 있다. 스마트폰만 있으면 거의 모든 것이 가능한 시대로 변화하고 있는 상황에서 스마트폰에 다양한 AI 기능이 탑재되는 것은 획기적인 변화를 가져올 수 있다. 현재 스마트폰에서 사용 가능한 AI 기능은 통번역, 음성인식, 보안, 카메라 보정 등으로 그리 많지 않지만 앞으로는 AI 기능이 매우 다양해질 것으로 보인다. 그리고 스마트폰에서 간단한 AI 기능은 인터넷이 연결되지 않아도 사용할 수 있는 특징이 있다. 이로 인해 보안 문제에서 자유로울 수 있다. 나아가 간단한 AI 기능은 데이터센터를 사용하지 않기 때문에 지연이 발생되지 않아 반응 속도가 빠르다. 하지만 온 디바이스 AI로만 AI 기능을 사용하게 되면 메모리 한계, 전력과 발열 문제 등으로 연산 능력이 제한적이다. 특히, 대형 AI 모델 처리에는 한계가 있기 때문에 하이브리드 형태로 온 디바이스 AI와 데이터센터를 동시에 사용하고 있다.

나아가 스마트폰에 사용되는 AI 반도체는 소형일 뿐만 아니라 저전력 구조로 설계되어 있다. 스마트폰은 공간적인 제약에 따라 빅칩(Big Chip)을 사용하지 못하고 휴대용으로 사용하기에 전력 소모가 적어야 하기 때문이다. 그리고 AI 반도체는 실시간 AI 연산을 수행하여 사용자 경험을 개선하는 핵심적인 역할을 수행하고 있다. 또한 데이터센터에서 사용되

는 AI 반도체가 두뇌라면 스마트폰에서 사용되는 AI 반도체는 지능을 가진 개인화된 비서의 역할을 수행한다고 볼 수 있다.

특히 스마트폰은 SoC를 사용하는 것이 일반적이고 SoC 안에 GPU와 NPU라는 AI 칩을 탑재시키고 있다. 삼성전자의 갤럭시(Galaxy) 스마트폰도 엑시노스(Exynos)라는 SoC 안에 GPU와 NPU를 탑재시켜 다양한 AI 기능을 가능하게 하고 있다. GPU와 NPU가 서로 협력하면서 AI 기능을 수행한다. GPU는 AR 렌더링, 실시간 영상, 사진합성 후처리, 게임 AI 가속 등을 담당하고 NPU는 음성인식과 이미지 분류 등의 추론을 담당한다.

구글의 픽셀폰은 SoC 안에 GPU, NPU와 TPU로 나뉘어 AI 기능을 수행한다. 작은 AI 연산을 TPU에서 돌리면 전력 소모가 커서 NPU가 대신 처리하며 반대로 대규모 텐서 연산은 NPU보다 TPU가 더 빠르고 효율적이다. 애플도 아이폰(I-phone)에 A시리즈의 SoC가 들어가고 그 안에 GPU와 뉴럴 엔진(Neural Engine)이라는 자체 제작한 NPU가 탑재되고 있다.

나아가 중국 기업들도 스마트폰을 많이 만들고 있다. 샤오미(Xiaomi)의 경우 자사의 스마트폰에 GPU와 NPU를 포함하고 있는 SoC를 사용하고 있다. 다른 기업들과 다르게 SoC를 자사에서 만들지 않고 퀄컴(Qualcomm)과 미디어텍(MediaTek)의 제품을 주로 사용하고 있다. 오포(Oppo)와 비보(Vivo) 등도 퀄컴과 미디어텍으로부터 SoC를 구입해서 사용하고 있다. 하지만 화웨이(Huawei)의 경우 하이실리콘(Hisilicon)이라는 자회사가 개발한 SoC를 사용하고 있으며 여기에도 GPU와 NPU가 들어가고 있다.

이와 같이 모든 스마트폰은 GPU와 NPU를 사용하고 있으며 공간적인 제약으로 SoC 안에 내장시키고 있다. 앞으로도 큰 변화가 없는 한 대부분의 스마트폰에 GPU와 NPU를 사용해 AI 기능을 사용하게 될 것으로 보고 있다. 그리고 복잡한 대규모의 데이터 연산이 필요할 경우 데이터센터를 통해 처리할 것으로 보인다. 최근 삼성전자의 경우 갤럭시에 제미나이(Gemini)를 탑재하였는데 다양한 AI 기능을 GPU와 NPU가 수행하게 된다. 하지만 대규모의 복잡한 데이터 처리를 위해서는 데이터센터로 자료를 보낼 수밖에 없다. 그렇게 되면 데이터센터에서 사용되고 있는 GPU, TPU와 ASIC/NPU 등이 AI 기능을 수행하게 된다.

이와 같이 스마트폰은 하이브리드 형태로 고객은 AI 기능을 사용하게 된다. 다른 제조사 스마트폰의 경우도 하이브리드 형태로 AI 기능을 사용하게 된다.

한편, 아직까지 모든 스마트폰에서 AI 기능을 사용할 수 없지만 점차적으로 대부분의 스마트폰에서 AI 기능을 사용할 수 있게 될 것으로 전망되고 있다. 이에 따라 앞으로 스마트폰에 AI 용도로 사용되는 GPU와 NPU의 수요는 크게 증가될 것으로 보인다. 나아가 모든 스마트폰에서 대형 AI 모델을 사용하기 위해 새로운 데이터센터의 확장이 필요해질 전망이다. 이에 따라 데이터센터에 필요한 AI 반도체인 GPU, TPU와 ASIC/NPU 등의 수요도 지속적으로 늘어날 것으로 보고 있다.

다. 자율주행차

자율주행은 인간의 삶에서 다양한 문제를 해결할 미래가치가 높은 기술이다. [18] 지금 자율주행차는 아직 자율주행 5단계까지 불가능하다. 하지만 3단계까지는 어느 정도 가능한 수준에 이르고 있으며 많은 기업이 완전한 자율주행이 가능할 수 있도록 노력을 기울이고 있다.

자율주행차에서 AI 반도체는 실시간 초저지연 연산을 통해 주행에서 안전을 보장한다. AI 반도체는 카메라, 레이더, 라이다(LiDAR) 등의 다양한 센서로부터 데이터를 받아 차량 주변의 환경을 인식하게 된다. 카메라로 촬영한 영상은 빠른 속도로 처리되고 [19] 감지된 센서 정보도 마찬가지이다. 또한 AI 반도체는 제한된 차량 내 전력에서 고성능과 저전력의 기능을 수행한다. 나아가 클라우드가 아닌 차량 내에서 온 디바이스 AI 연산을 수행하면서 객체 탐

지, 경로 계획과 의사결정을 실시간으로 처리한다. 뿐만 아니라 AI 반도체는 운전자 모니터링과 음성 인식 같은 스마트 기능까지 수행이 가능하다. 결론적으로 AI 반도체는 자율주행차의 두뇌로써 안전한 주행을 위한 핵심 엔진의 역할을 수행한다.

차량이 더 진화된 자율주행 단계로 발전되면서 반도체의 수요가 급증하고 있다. AI 반도체는 주로 ECU(Electronic control unit)라고 하는 모듈에 사용되어 다양한 제어 역할을 수행하고 있다. 이런 ECU는 차량에 따라 설치되는 수량이 천차만별이다.

먼저 자율주행의 핵심은 CCU(Central Computing Unit) 부문에 사용되는 AI 반도체라 할 수 있다. AI 반도체는 차량이 주행하는 동안 일어나는 모든 일을 이해하고 판단하며 움직이게 하는 두뇌의 역할을 해야만 하기 때문이다. 특히, 최근 엔비디아와 퀄컴이 자율주행차의 AI 반도체 시장을 대부분 장악하고 있다. 이들 기업은 오랜 기간 생태계를 조성하면서 자율주행 플랫폼으로 고객을 확보해 오고 있다. 물론 테슬라도 빼놓을 수 없는 기업이지만 테슬라는 AI 반도체를 자사의 자율주행차에 탑재시키기 위한 용도로만 만들고 있다는 점에서 다른 기업들과 다르다. 이외에도 국내에서 자율주행 칩을 만드는 기업으로 텔레칩스와 넥스트칩 등이 있다.

한편, 자율주행차는 엣지 컴퓨팅을 기본으로 하지만 경우에 따라 대규모 데이터의 연산, 그리고 차량의 소프트웨어의 업데이트와 V2X(Vehicle to Everything)를 위해 데이터센터와의 통신이 필수이다. 그럼에도 불구하고 차량은 움직이는 상태에서 대부분의 정보를 파악하고 결정을 내려야하기에 차량 내에서의 엣지 컴퓨팅이 중요하다. 데이터 처리를 지연 없이 바로 처리해야만 안전하게 운행할 수 있기 때문이다.

한편, 자율주행차에 주로 사용되는 AI 반도체는 GPU와 NPU/DLA(Deep Learning Accelerator)이다. 이미 언급한 바와 같이 엣지 컴퓨팅에 최적으로 사용될 수 있는 AI 반도체는 GPU와 NPU/DLA이기 때문이다. 하지만 GPU와 NPU/DLA도 단독으로 사용되기 보다 주로 SoC의 내부에 사용되는 것이 일반적이다. 마찬가지로 엔비디아에서 만들고 있는 오린(Orin)과 토르(Thor) 칩도 SoC이다. 이 SoC 안에는 GPU와 NPU/DLA가 들어가는데 서로 협업하면서 AI 기능을 수행한다. 예를 들면 복잡한 이미지와 멀티모달(Multi Modal)과 같은 AI 연산은 GPU가 담당하는 반면 NPU/DLA는 전력 효율이 우수해 반복적이고 최적화된 추론 작업을 맡게 된다. 그리고 차량에 사용되는 SoC는 보통 AP라고 한다. 이런 AP는 차량의 각 부분에 사용되고 있다.

다음은 차량 내 인포테인먼트(Infotainment) 부문에 사용되고 있는 AI 반도체이다. 마찬가지로 칩의 형태는 SoC이고 일반적으로 AP이며 SoC 내에는 GPU와 NPU가 들어가 있다. 인포테인먼트에 사용되는 AI 반도체는 운행에는 관여하지 않지만 운전자나 다른 탑승자를 대상으로 승차하는 동안 무료함을 느끼지 않게 하기 위해 다양한 AI 기능을 실행한다. 예를 들면 음성비서, 차량 내 번역, 실시간 통화 노이즈 제거, 음성 명령 인식, 차량 내 LLM 실행 등의 역할을 하고 있다. 최근 많은 차량이 서서히 자율주행차로 바뀌면서 차량에 인포테인먼트 기능이 필수적으로 탑재되고 있는 추세이다.

마지막으로 운전자 모니터링(Driver Monitoring System)에 필요한 AI 반도체이다. 운전자가 운전하는 동안 얼굴과 시선을 추적하여 졸고 있지 않은지 감지하고 안전하게 운행할 수 있도록 알림을 주게 된다. 물론 완전 자율주행차로 바뀌면 굳이 필요하지 않게 될 수 있지만 현재로서는 운전자가 운전을 필수적으로 해야만 하기에 중요한 기능이라 할 수 있다.

참고로 V2X가 되면 차량과 차량, 차량과 신호체계, 차량과 기타 인프라 등과 교신을 해야만 하기에 고성능의 AI 반도체가 필요해질 수 있지만 완전한 V2X는 아직까지 불가능하다.

이와 같이 차량에는 다양한 AI 반도체가 필수적으로 탑재되며 완전 자율주행차로 바뀌게 되면 더 많은 다양한 AI 반도체가 필요해질 것으로 보고 있다.

라. 협동 로봇

최근 많은 공장이 스마트 팩토리로 전환되면서 자동화되고 있는 추세이다. 그리고 공장에 다수의 로봇이 투입되어 인간과 같이 협업하는 사례가 늘어나고 있다. 이와 같이 사람과 같은 작업 공간에서 안전하게 협력 작업을 수행할 수 있도록 설계된 로봇을 협동 로봇(Collaborative Robot)이라 정의한다.

협동 로봇에 사용되는 AI 반도체는 사람과 함께 작업하는 환경에서 실시간 초저지연 연산을 하고 경로 계획과 동작 제어를 통해 안전한 동작을 가능하게 한다, 자율주행차와 마찬가지로 AI 반도체는 카메라, 3D 비전, 힘 센서 등 다양한 센서를 통해 들어온 정보를 분석하여 주변을 인식한다. 동시에 음성·제스처·시선 추적을 인식하고 인간과 로봇 간 상호작용을 가능하게 함으로써 인간과 로봇이 협력할 수 있다. 결국 AI 반도체는 협동 로봇이 사람과 안전하고 효율적으로 협력하는 데 필요한 두뇌 역할을 하는 핵심 엔진이라 할 수 있다.

나아가 인간과 안전한 상호작용을 위해 힘·속도 제한, 충돌 감지와 정밀한 제어 기능을 갖추는 것이 무엇보다 중요하다. 아울러 협동 로봇은 산업용 로봇과 비교해서 사용되는 형태가 여러 가지 면에서 차이가 있다. 표 4는 산업용 로봇과 협동 로봇의 차이점을 나타내고 있다.

표 4. 산업용 로봇과 협동 로봇의 차이점

구분	산업용 로봇	협동 로봇
작업 공간	안전펜스안에서 단독작업	사람과 같은 공간에서 협력작업
안전성	물리적 분리로 안전	힘·속도제한, 충돌감지로 안전
작업 형태	반복적 대량생산에 적합	다품종·소량 작업에 적합
프로그래밍	전문 엔지니어 필요	간단한 프로그래밍 가능
설치 공간	넓은 공간 필요	공간 절약
비용	고비용	ROI(투자 회수) 빠름

일반적으로 협동 로봇은 데이터센터와 연결하지 않은 상태에서 로봇 자체적으로 AI 기능을 활용한 작업이 가능하다. 로봇을 데이터센터와 연결하지 않은 상태에서 작업을 하는 것은 다음과 같은 두 가지 이유로 볼 수 있다. 첫째, 공장에서 인간과 같이 작업을 하기에 인간의 안전을 고려할 수밖에 없기 때문이다. 만약 데이터센터와 연결하게 되면 지연이 발생할 수 있어 작업 중 인간에게 치명적인 손상을 입힐 수 있다. 따라서 로봇 자체적으로 지연이 없이 무난하게 AI 기능을 수행할 수 있다면 굳이 데이터센터와 연결할 필요가 없게 된다. 둘째, 공장의 기술적 보안이 중요하기 때문이다. 특히, 데이터센터와 연결하여 작업을 하게 되면 외부로 기술 유출의 우려가 있기에 기술 유출을 방지하기 위해 로봇 자체적으로 AI 기능을 실행할 수밖에 없다. 이와 같은 이유로 협동 로봇은 엣지 컴퓨팅을 이용해 로봇 자체적으로 AI 기능을 활용하는 것이 일반적이다. 그럼 협동 로봇에는 어떤 AI 반도체가 사용되고 있는지 조사해 보도록 한다.

협동 로봇도 엣지 디바이스로 볼 수 있기에 보통 GPU와 NPU/DLA가 사용되고 있다. GPU는 로봇의 시각을 담당하고 NPU/DLA는 로봇의 두뇌를 담당하고 있다고 생각하게 되면 이해하기 쉽다. GPU는 CNN(Convolutional Nueral Network) 기반으로 사물을 인식할 뿐만 아니라 3D 비전 데이터 분석에 강점을 가지고 있다. NPU/DLA는 딥러닝 추론을 고속으로 실행할 수 있으며 경로를 계획하고 실시간으로 의사결정을 내릴 수 있다. 특히, NPU/DLA는 저전력과 저지연이 강점으로 로봇에 사용되기에 매우 적합하다. 마찬가지로 GPU와 NPU/DLA는 SoC 안에 탑재가 되며 SoC는 로봇에 따라 사용되는 개수가 다르다. 그리고 최근 로봇에 AI 기능이 강화되는 추세로 가고 있다. 이에 따라 AI 기능을 포함한 SoC 공급기업으로 엔비디아와 퀄컴 등이 시장에서 강자로 떠오르고 있다. 특히, 엔비디아는 AI 비전과 딥러닝 추론 최강자로서 Jetson AGX Orin이 대부분의 로봇에 표준처럼 사용되고 있다. 2022년 출시된 Jetson AGX Orin은 최대 275 TOPS 성능으로 데이터센터의 연결 없이 온디바이스 AI 연산을 가능하게 하는 핵심 SoC 플랫폼이다. 나아가 협동 로봇의 제조로 시장의 40-50%를 점유하고 있는 유명한 기업으로 덴마크

의 UR(Universal Robots)이 있다. 주로 대기업보다 중소기업의 시장을 많이 점유하고 있다. 협동 로봇 자체가 주로 중소기업에서 많이 사용하고 있기 때문이다. 참고로 협동 로봇이 사용되고 있는 장소는 주로 전자부품 조립, 포장과 라벨링, 식음료 등의 생산 공장이다.

5. AI 반도체의 응용 분야에 대한 비교 분석

지금까지 AI 반도체가 사용되는 응용 분야에 대해 알아보았다. 다시 한번 응용 분야를 정리하면 다음과 같다. 첫째, 데이터센터이다. 둘째, 스마트폰이다. 셋째, 자율주행차이다. 마지막으로 협동 로봇이다. 그럼 이들 응용 분야에 사용되고 있는 AI 반도체와 응용 분야가 서로 어떻게 다르고, 그리고 어떤 특징이 있는지를 비교하여 분석해 보도록 한다.

첫째, 데이터센터는 아직까지 엔비디아의 GPU가 거의 독점하는 시장이다. 대부분의 빅테크 기업은 엔비디아의 GPU를 사용하고 있지만 동시에 자사 제품에 적합한 AI 반도체도 자체적으로 만들고 있다. 그 종류는 주로 NPU(TPU)와 ASIC 같은 AI 반도체이다. 그리고 훈련 분야에서는 엔비디아의 GPU가 여전히 막강한 지위를 가지고 있지만 앞으로 추론 분야에서는 NPU와 ASIC이 사용될 가능성이 커지고 있다. 둘째, 스마트폰은 대부분 GPU와 NPU가 내장된 AP를 사용하는 것으로 나타났다. AI 기능이 탑재된 AP로 인해 간단한 AI 기능은 온디바이스 AI를 통해 가능하다. 그리고 복잡한 AI 기능은 데이터센터와의 연결을 통해 사용이 가능하다. 이에 따라 스마트폰은 하이브리드 형태로 볼 수 있다. 셋째, 자율주행차는 주행 중 사물에 대한 인식, 판단과 결정을 빠르게 내려야만 하기 때문에 엣지 컴퓨팅이 중요하다. 하지만 차량의 소프트웨어의 업데이트와 차량에서 복잡한 멀티 모달 기능을 사용하기 위해 데이터센터와의 연결이 필요하다. 차량에 쓰이는 AI 반도체는 주로 AP이며 AP에 GPU와 NPU/DLA가 내장되어 있다. 마지막으로 협동 로봇은 다른 응용 분야와 달리 데이터센터에 연결하지 않은 상태로 사용할 수 있는 매우 독특한 경우이다. 이에 따라 로봇 자체적으로 AI 기능을 거의 완벽히 수행할 수 있다. 협동 로봇에 사용되는 AI 반도체는 SoC로 SoC 내에는 GPU와 NPU/DLA가 내장되어 있다. GPU는 이미지와 동작을 인식하고 NPU/DLA는 경로를 계획하고 의사결정을 내릴 수 있다.

지금까지 AI 반도체의 응용 분야에 대해 살펴보았다. 마지막으로 각 AI 반도체가 사용되는 응용 분야를 서로 비교하여 분석하면 아래의 표 5와 같다.

표 5. AI 반도체의 각 응용 분야별 비교

구분	데이터센터	스마트폰	자율주행차	협동 로봇
필요 AI 반도체	학습: GPU, TPU 추론: NPU, ASIC	AP(GPU, NPU/TPU)	AP(GPU, NPU/DLA)	AP(GPU, NPU/DLA)
AI 반도체 주요 기업	학습: 엔비디아, AMD 추론: 엔비디아, 브로드컴, 마벨 테크놀로지, 리벨리온	애플, 삼성전자, 퀄컴, 미디어텍	엔비디아, 퀄컴	엔비디아, 퀄컴
주요 특징	빅테크 자체 설계도 있음	애플과 삼성전자는 자사용	플랫폼으로 제공	플랫폼으로 제공
데이터센터 연결	필수	하이브리드	하이브리드	불필요

III. 결 론

본 논문에서는 AI 반도체의 다양한 종류와 특징에 대해서 알아보고 AI 반도체를 개발하고 있는 다양한 기업들에 대해서도 조사해 보았다. 나아가 AI 반도체가 사용되는 주요 응용 분야가 무엇인지 알아보고 AI 반도체가 각 응용 분야에 어떻게 활용되고 있는지, 그리고 각 응용 분야별 차이점은 무엇인지 비교하여 분석하였다. 본 연구 결과를 통해 AI 반도체는 사용 환경과 목적에 따라 연산 방식, 전력 효율, 반응 속도 등의 요구사항이 크게 다르며, 이에 따라 칩의 아키텍처와 기업 전략 또한 다르게 설계되고 있음을 확인할 수 있었다.

앞으로 AI 반도체는 데이터센터의 확장뿐만 아니라 다양한 엣지 디바이스의 출현으로 수요가 늘어날 수밖에 없다. 이와 같은 상황에서 AI 반도체에 대한 주요 현황과 응용 분야를 비교하여 분석한 것은 향후 기술 트렌드를 알 수 있어 업계와 학계에 큰 시사점을 줄 수 있다. 하지만 본 연구가 기업들의 제품에 대한 구체적인 기술적 분석에서 다소 미흡한 점이 있었다. 차후 과제는 앞으로 AI 반도체가 어떻게 발전하게 되고, 나아가 어떤 새로운 분야에 활용될 수 있는지 더욱 심도 있게 연구가 진행될 필요가 있을 것으로 보인다.

REFERENCES

[1] Er-Ping Li et al., "An Electromagnetic Perspective of Artificial Intelligence Neuromorphic Chips," *Electromagnetic Science,* vol. 1, no. 3, Sep. 2023.

[2] Wenqiang Zhang et al., "Neuro-inspired computing chips," *Nature Electronics,* vol. 3, pp. 371 - 382, Jul. 2020.

[3] Prashis Raghuwanshi, "Revolutionizing Semiconductor Design and Manufacturing with AI," *Journal of Knowledge Learning and Science Technology,* vol. 3, no. 3, Sep. 2024.

[4] Sorna Mugi Viswanathan, "AI Chips: New Semiconductor Era," *International Journal of Advanced Research in Science, Engineering,* Technology, vol. 7, no. 8, pp. 14687-14694, Aug. 2020.

[5] 김현지, 윤세영, 서화정, "국내외 인공지능 반도체에 대한 연구 동향," *스마트미디어저널,* 제13권, 제3호, 36-44쪽, 2024년 3월

[6] Hiroshi Momose, Tatsuya Kaneko, and Tetsuya Asai, "Systems and circuits for AI chips and their trends," *Japanese Journal of Applied Physics,* 59, 2020.

[7] S. Ambrogio et al., "An analog-AI chip for energy-efficient speech recognition and transcription," *Nature,* vol. 620, 24 Aug. 2023.

[8] Heyang Xu, "FPGA: The super chip in the age of artificial intelligence," *Journal of Physics: Conference Series,* 2023.

[9] Wei Gao, and Pingqiang Zhou, "Customized High Performance and Energy Efficient Communication Networks for AI Chips," *IEEE,* vol. 7, 2019.

[10] Lennart P. L. Landsmeer et al., "Tricking AI chips into simulating the human brain: A detailed performance analysis," *Neurocomputing,* vol. 598 Sep. 2024.

[11] Q. Zhang, H. Deng, and K. Song, "Latest VLSI Techniques for 3nm Technology for Building Efficient AI Chips," *Fusion of Multidisciplinary Research, An International Journal (FMR),* vol. 5, no. 2, 2024.

[12] P. Ebby Darney, "A Review on Artificial Intelligence Chip," *Recent Research Reviews Journal,* vol. 1, no. 1, pp. 99-109, Dec. 2022.

[13] Dhruvitkumar V. Talati, "Silicon minds: The rise of AI-powered chips," *International Journal of Science and Research Archive,* vol. 01, no. 02, pp. 097-108, 2021.

[14] Rebellions (2025), https://rebellions.ai/ko/rebellions-product/rebel-quad/, (accessed Aug., 10, 2025).

[15] FuriosaAI (2025), https://www.furiosa.ai, (accessed Aug., 10, 2025).

[16] Olga Krestinskaya, Khaled N. Salama, and Alex P. James, "Automating Analogue AI Chip Design with Genetic Search," *Advanced Intelligent Systems,* 2020.

[17] Manjunath Chandrashekaraiah, "Novel Semiconductor Chip-Based Smart Sensor Technology Design and Integration of IoT in AI-ML Hyperscale Infrastructure," *International Journal of Image Processing and Smart Sensors,* vol. 1, no. 1, Jan. - Jun. pp. 1-21, 2025.

[18] 윤재웅, 이주홍, "안전하고 효과적인 자율주행을 위한 불확실성 순차 모델링," *스마트미디어저널,* 제11권, 제9호, 9-20쪽, 2022년 10월

[19] 신윤선, 서주현, 이민영, 김인중, "SoC 환경에서 TIDL NPU를 활용한 딥러닝 기반 도로 영상 인식 기술," *스마트미디어저널,* 제11권, 제11호, 25-31쪽, 2022년 12월

저 자 소 개

권영화(정회원)

1994년 광운대학교 영문학과 학사 졸업.
1998년 성균관대학교 무역학과 석사 졸업.
1999년 한양대학교 일본학과 석사 졸업.
2013년 서울과학종합대학원 경영학과 박사 졸업.

<주관심분야 : AI 반도체, AI 디바이스, 자율주행차>

김보영(정회원)

1996년 이화여자대학교 정보디자인학과 학사 졸업.
2000년 이화여자대학교 정보디자인학과 석사 졸업.
2006년 브루넬 대학 공학 박사 졸업

<주관심분야 : 미디어 처리, 상황 인지>