

AI 반도체 기술동향과 산업생태계

이선재 한국전자통신연구원 기술정책연구본부 선임연구원

1. 머리말

반도체는 4차 산업 혁명의 모든 분야에 필수적으로 활용되는 미래 기술의 핵심이며, 경제 및 국가 안보의 주요 자산이다. 인공지능은 생활 및 산업에 쓰이지 않는 곳이 없을 정도로 중요한 기술이다. 반도체 기술 발전은 인공지능의 성능 향상을 견인하고, 인공지능의 가능성에 대한 발견으로 많은 반도체 기업들이 인공지능 산업에 참여하게 되면서 산업 자체의 파이를 키우는 선순환 관계를 갖게 되었다.

산업이 융합하고 AI가 해결해야 하는 문제가 복잡해짐에 따라 AI 모델의 복잡도가 증가하고 대규모 연산량을 요구하게 되었으며, 고성능 저전력 인공지능 전용 반도체가 필요하게 되었다. AI 반도체란, AI의 학습, 분석, 추론 등의 서비스를 수행하기 위해 대량 연산을 목적으로 만들어진 반도체로서 생활 및 산업에 AI의 활용 범위가 넓어지면서 그 역할은 점차 확대될 것이다. AI 역량이 국가와 기업의 생존을 좌우하는 시대에서는 AI반도체와 같이 대규모 데이터 연산을 빠르고 효율적으로 실행하는 새로운 차원의 컴퓨팅 하드웨어를 통해 획기적인 AI 기술 구현이 필요하다.

또한 AI가 해결해야 하는 문제점이 산업별 특성을 내포하고 있으므로 이러한 특성을 고려한 반도체 설계의 중요성이 증가하면서 AI반도체를 설계하는 팹리스의 역할이 부각되고 있다. 각국의 반도체 산업 정책 동향[1] 및 미국의 VC 투자현황에서도 이를 확인할 수 있다[2].

본 고에서는 AI반도체 산업생태계와 주요 기업 및 팹리스 스타트업의 동향을 알아보고 AI반도체 선순환 생태계 지속가능성에 대한 시사점을 찾고자 한다.

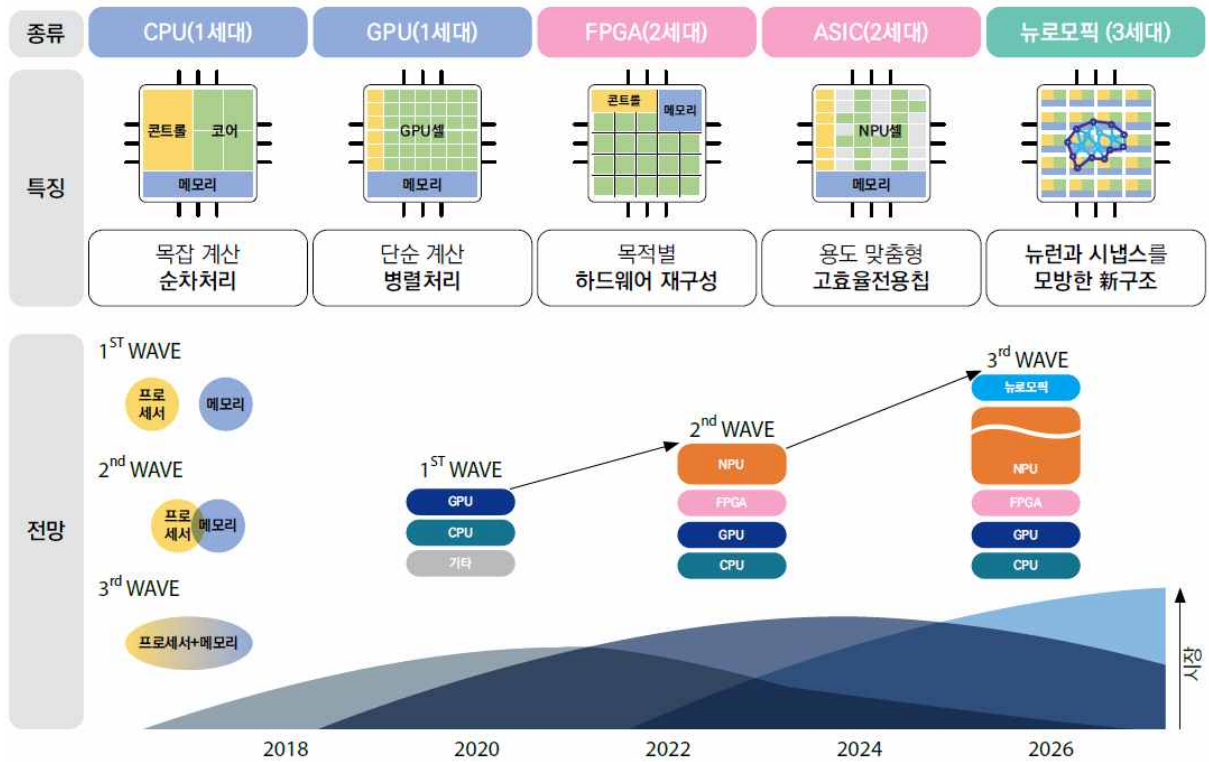
2. AI반도체 기술동향 및 산업생태계

2.1 AI반도체의 필요성, 등장 배경 및 기술진화

알렉스넷은 2012년 사물 인식 경진 대회인 이미지넷(Imagenet)에서 GPU를 활용하여 기존 프로그램이 보여주지 못한 성능을 보여주었으며, 이를 계기로 인공지능 분야에서 GPU를 활용한 AI 모델이 급속도로 발전하였다. 알렉스넷의 성공은 하드웨어가 AI의 성능 향상을 견인한 중요한 사건이며, 세계가 인공지능의 가능성을 깨닫게 된 계기가 되었다.

하드웨어를 기반으로 혁신을 이루고 있는 AI는 시각·언어·음성과 같은 단일지능에서 트랜스포머를 기반으로 한 초거대 AI로 발전하고 있다. 딥러닝 기반의 단일지능 기술들이 성숙하면서 멀티모달 및 복합지능으로 진화하여 수천억 개의 파라미터로 구성된 대규모 연산이 요구되고

있으며, 이에 따라 인공지능 연산을 위한 하드웨어인 AI 반도체는 (1세대) CPU+GPU ⇒ (2세대) NPU¹⁾ ⇒ (3세대) 뉴로모픽으로 발전하고 있다.([그림 1] 참조) 또한 메모리(기억)-프로세서(연산) 통합으로 미래 반도체 설계 패러다임을 완전히 바꿀 수 있는 신개념 PIM²⁾기술도 발전하고 있다. '3세대 반도체'인 非폰노이만 방식의 뉴로모픽이 가장 진보된 형태의 AI 반도체로 평가받고 있지만 아직 기초 연구단계 수준이며, 현재는 NPU 중심으로 상용제품이 개발되어 활용 가능성을 높여가고 있다.



[그림 1] AI반도체 기술진화[3]

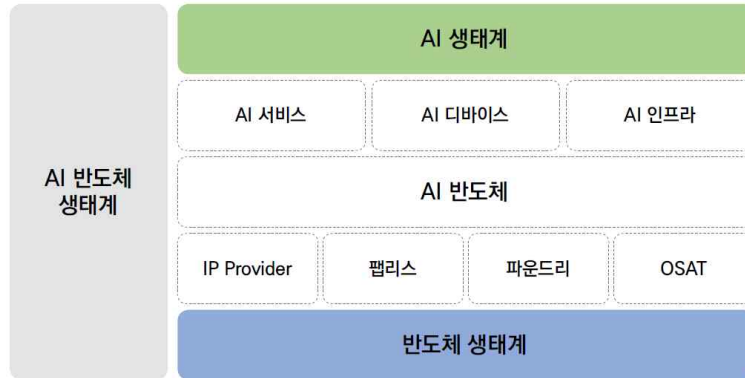
2.2 AI반도체 시장 및 생태계 변화

AI반도체 생태계는 기존 시스템 반도체 생태계와 인공지능 생태계가 융합된 새로운 산업구조로 바라볼 수 있다.([그림 2] 참조) 반도체와 인공지능 생태계가 융합된 특성으로 인해 주요 기업들은 기존의 사업 영역을 유지하면서 융합시장 방향으로 생태계를 확장하고자 하고 있다. 엔비디아, 인텔, 퀄컴 등 기존 반도체 기업들은 칩의 성능을 지속적으로 향상시키면서 호환성을 장점으로 다양한 산업 영역에서 칩 기반의 AI서비스를 제공한다. 구글, 메타와 같은 빅테크 기업들은 초기에는 호환성 측면에서 반도체 기업들의 제품을 활용하였지만, 자사 AI 알고리즘 수행에 최적화된 반도체의 필요성으로 인해 자체 칩을 제작하여 적용하고 있다.

인공지능의 성능은 하드웨어+시스템 SW+응용 SW의 최적화로 결정되기 때문에 하드웨어뿐만 아니라 이를 운영하고 응용 소프트웨어를 지원하는 시스템 소프트웨어, AI 모델, 서비스까지 모두 중요하다. AI반도체 관련 기업들의 개발 동향 및 전략 발표를 종합하여 살펴보면, 하드

1) NPU : Neural Processing Unit
2) PIM : Processing-In-Memory

웨어부터 소프트웨어까지 모든 단계의 제품 및 서비스를 제공하는 형태인 풀스택(Full-stack)³⁾을 갖추고 AI반도체 성능을 효과적으로 구현하기 위한 생태계를 조성하기 위해 노력 중이다. 팹리스를 포함한 반도체 기업부터 빅테크 기업까지 AI반도체 관련 기업들의 확장 시작점은 하드웨어나 서비스로 각기 다르지만, 풀스택을 확보하고자 하는 목표는 동일하다.



[그림 2] AI반도체 생태계[4]

2.3 주요 AI반도체 기업의 생태계 확장 전략

젠슨 황 엔비디아 CEO는 엔비디아는 반도체회사가 아니고 소프트웨어 회사라 말한 바 있다. 엔비디아 GPU는 '쿠다(CUDA)⁴⁾'라는 플랫폼을 갖고 있어 가치가 있다고도 언급했다. 이러한 사례에서 확인할 수 있듯 AI반도체 기업들의 생태계 확장 전략에서 중요한 요소 중 하나는 자사 칩 기반의 SW 플랫폼이다. 구글의 텐서플로(TensorFlow)의 사용자가 많은 것은 구글 반도체의 성능뿐만 아니라 SW 환경이 주요 요인이기도 하다. 기업이 성장하기 위해서는 소비자가 기대하는 칩의 성능 향상은 기본이며, 칩에 최적화된 인공지능 모델과 AI 컴파일러⁵⁾가 없으면 시장에서 외면받을 가능성이 생긴다.

따라서 AI반도체 기업들은 <표 1>과 같이 자사 칩에 맞는 시스템 소프트웨어를 개발하고 사용자에게 활용하도록 함으로써 소프트웨어 개발자들을 자사 생태계에 묶어두는 락인 효과⁶⁾를 기반으로 생태계 확장 및 활성화를 기대하고 있다.

<표 1> AI반도체 SW 스택[5]

AI SW	AI 서비스 및 앱
	AI 플랫폼
	AI frameworks, AI 라이브러리
	시스템 SW (AI 컴파일러 등)
AI HW	AI 반도체






3) 주로 SW와 HW 전반에 대한 이해도가 높은 개발자를 뜻하는 용어로 사용하지만, 본 고에서는 하드웨어부터 소프트웨어까지 모든 단계의 제품 및 서비스를 제공하는 형태를 의미함

4) CUDA : Computed Unified Device Architecture

5) 딥러닝 모델을 타겟 디바이스에서 최적의 속도와 정확도를 낼 수 있는 머신 코드로 자동 변환 작업을 지원하는 도구

6) 특정 서비스를 한 번 이용하면 다른 서비스를 소비하기 어려워져 기존의 것을 계속 이용하는 효과

<표 2> AI반도체 기업별 시스템 소프트웨어

기업명	엔비디아	인텔	AMD	구글	메타
시스템 소프트웨어					
AI 하드웨어 (주요 제품)	GPU (H100)	GPU (ARC A350M)	GPU (Radeon RX 7900)	TPU V4	오쿨러스용 칩

2.4 주요 AI반도체 기업 동향

2.4.1 엔비디아

하드웨어인 GPU의 성능 개선 및 데이터센터용 CPU 개발을 진행 중이며, GPU 및 시스템 소프트웨어(CUDA)를 기반으로 다양한 AI 서비스를 제공하고 있다. CUDA는 GPU 컴퓨팅에서 일종의 컴파일러 역할을 하는 도구로, 엔비디아 GPU가 인공지능(AI) 혁명을 견인한 원동력이자 고성능 컴퓨팅(HPC) 분야의 필수 솔루션으로 빠르게 확산된 주요 요인이라 할 수 있다.

2.4.2 인텔

하드웨어 기반 소프트웨어 역량 강화 추진 전략으로서 슈퍼컴퓨팅 환경을 구축하고 서비스형 소프트웨어(SaaS, Software as a Services)를 제공하고 있다. 인텔은 2020년 12월 oneAPI 툴킷 출시를 발표했으며, 이 oneAPI 툴킷을 이용하여 개발자는 CPU, GPU, FPGA(통칭 XPU)를 활용한 고성능 교차 아키텍처 애플리케이션을 개발할 수 있다.

2.4.3 AMD

2022년 3월 새로운 AMD 인스팅트 MI210(AMD Instinct MI210) 액셀러레이터와 ROCm 5 소프트웨어를 발표하며 AMD 인스팅트 생태계를 다시 한번 확장하고자 하고 있다.

2.4.4 퀄컴

2022년 6월 퀄컴은 제조사와 개발자에게 엔드-투-엔드 AI 소프트웨어 및 단일 통합 소프트웨어 스택으로 퀄컴의 AI 소프트웨어를 제공하는 퀄컴 AI 스택을 공개하였다. 기존 AP 칩을 기반으로 소프트웨어 역량을 강화하면서 다양한 산업 분야에서 경쟁력을 확보하고자 하고 있다.

2.4.5 구글

‘구글I/O 2022’에서 스마트워치부터 태블릿 PC, 무선이어폰에 이르기까지 다양한 신제품들을 선보이며 하드웨어 시장 진출을 공식화함으로써 기존 소프트웨어 시장을 넘어 본격적으로 하드웨어 시장에 진출하고 있다.

2.5 국내외 팹리스 스타트업 동향

현재 인공지능 시장은 엔비디아의 GPGPU 기반의 시스템이다. 하지만 1세대 AI반도체인 GPU

는 AI를 처리할 수 있는 성능은 갖췄어도 당초 개발 목적이 AI용이 아니므로 AI연산 외 부분에서 성능이 낭비되고, 비용이나 전력 소모 등의 측면에서 비효율적 부분이 발생하는 문제점이 있다. AI반도체 팹리스 스타트업들은 이러한 약점을 파고들며 NPU를 개발하고 서버 및 엣지 시장에 진출하고자 노력하고 있다.

해외 주요 기업으로는 세레브라스, 그래프코어, 그로크, 하일로 등이 있으며, 국내 기업으로는 퓨리오사, 리벨리온, 딥엑스 등이 GPU 대비 우수한 성능 및 가격, 저전력을 무기로 자율주행 및 메타버스와 같은 미래 신산업을 선점하기 위해 치열하게 경쟁하고 있다.([그림 3] 참조)



[그림 3] AI반도체 스타트업의 시장 공략

국내 팹리스 스타트업 제품의 성능을 Mlperf 기준으로 측정한 결과, 퓨리오사의 워보이와 리벨리온의 아이온 모두 동급 성능의 GPU 대비 높은 성능을 기록했다. 또한 엣지용 칩에 주력하고 있는 딥엑스는 벤치마크의 모바일넷 알고리즘에서 최상위 결과를 확인했다고 발표하였다. (<표 3>참조)

3. 맺음말

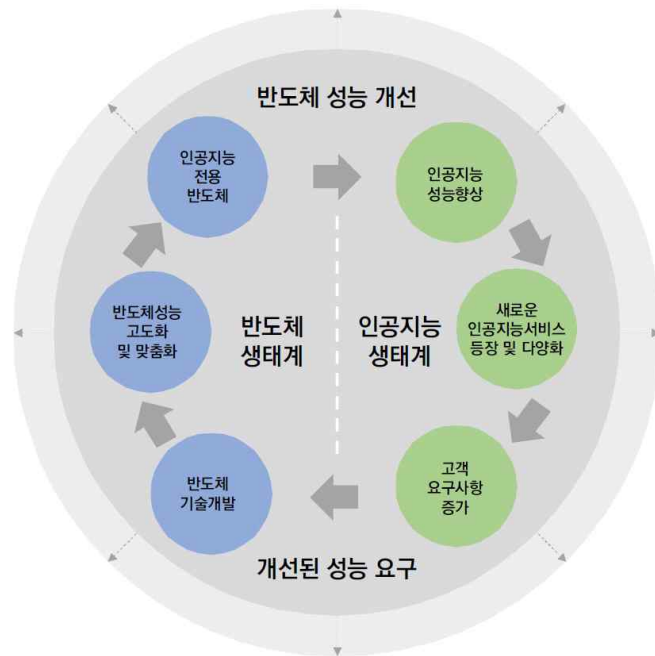
AI반도체 시장은 아직 지배적 강자가 존재하지 않는 초기 단계로, 지금부터의 국가적 대응 노력이 글로벌 주도권 경쟁의 성패를 좌우한다. 국내 팹리스 스타트업들이 타겟 시장에서 성공을 거둔다면, 해당 산업 분야에서는 인텔이나 엔비디아와 같이 시장을 선도하는 기업으로 성장할 가능성이 존재하기 때문에 팹리스 스타트업의 중요성을 인식하고 육성할 필요성이 있다.

NPU 반도체를 오픈소스로 공유해 개발하고 SW 컴파일러, SW SDK까지 공유하는 현실을 감안할 때, 비즈니스 모델 관점에서 국내 AI반도체 산업이 경쟁력을 확보하려면 관련 기업들이 다양한 인공지능 반도체 솔루션에서 역량을 키울 방안을 모색해야 한다.

생태계 관점에서 국내 팹리스 스타트업이 대기업의 수직계열화, 대기업과의 연합모델이 아닌 독자 생존을 모색하려면 엔비디아 GPU의 벽을 넘어야 하며, 칩이 경쟁력을 갖기 위해서는

GUDA 기반의 락인 효과를 극복할 방안을 찾아야 한다.

국내 AI반도체 생태계의 지속가능성을 위해서는 AI반도체 팹리스의 성장을 기반으로 [그림 4]와 같이 반도체는 AI반도체 팹리스를 중심으로 AI서비스 구현을 위한 반도체 성능 개선과 기술개발을 수행하며, 인공지능은 향상된 반도체 성능을 기반으로 고객 만족을 위한 다양한 AI서비스를 제공하는 산업 간 선순환 구조를 구축해야 할 것이다.



[그림 4] AI반도체 선순환 생태계

[참고문헌]

- [1] ICT Brief(2022-05), 인공지능 반도체 특집호 (IITP, 2022)
- [2] PCAST Semiconductors Report (PCAST, 2022)
- [3] 인공지능(AI) 반도체의 산업경쟁력 (특허청, 2022)
- [4] 인공지능 반도체 산업 발전전략 (관계부처 합동, 2020)
- [5] 인공지능 발전에 따른 지능형 반도체의 등장과 반도체 생태계 변화에 관한 연구(KISDI, 2019)

※ 출처: TTA 저널 제205호