AI 반도체 기술 및 산업 동향

KDB미래전략연구소 산업기술리서치센터
장은현 전임연구원(ehj@kdb.co.kr)

AI의 역할이 산업 전반으로 확장됨에 따라 고성능 반도체의 수요가 급증하고 있다. AI 반도체는 인공신경망 처리에 최적화된 반도체로, AI 반도체 시장은 AI의 발전과 함께 가파르게 성장 중이다. 세계 AI 반도체 시장은 향후 5년간 연평균 24%로 성장하여 '28년에는 1,590억 달러를 기록할 것으로 전망된다. 또한, AI 반도체의 응용 분야는 데이터센터와 스마트폰 중심에서 점차 차량, PC 등 엣지디바이스 분야로 확장될 것으로 보인다.

AI 반도체의 핵심 분야는 프로세서, 메모리, 인터페이스 등으로 구분할 수 있다. 각 분야 내에서는 AI 서비스별 최적화, 초고속·초저전력 등 연산 효율성 향상 등을 목표로 활발한 연구가 진행되고 있으며, AI GPU, NPU 및 뉴로모픽 반도체, HBM 및 PIM, CXL 등이 주요 기술로 개발되고 있다.

AI 반도체 산업은 크게 설계 분야의 팹리스, 제조 분야의 파운드리로 분업화되어 있으며, 전방 산업으로 데이터센터, 스마트폰, 차량 등이 포함된다. AI 반도체 산업은 ①기존 반도체 기업들의 사업 포트폴리오 확장, ②AI 반도체 주요 수요 기업들의 자체 칩 개발, ③팹리스 스타트업 진출 확대 등 밸류체인 내 다수의 기업들이 영역을 확장하며 생태계 패러다임의 변화를 겪고 있다.

한국은 우수한 메모리 역량, AI 반도체 팹리스 스타트업의 성장 등으로 AI 반도체 분야에서 기술격차를 좁혀가고 있으나, 아직 선도국 대비 기술수준이 열위한 편이다. 정부가 국산 AI 반도체의 중요성을 인식하고 최근 관련 프로젝트를 지속 추진 중인 가운데, AI 반도체 기술력 확보를 위해서는 하드웨어 개발 외에도 수요처와의 유기적 연계를 통한 생태계 구축이 필요할 것이다.

* 본고의 내용은 집필자 견해로 당행의 공식입장이 아님

Ⅰ. AI 반도체 개요

1. AI 반도체 정의 및 구분

❑ **AI 반도체는 인공신경망 처리에 최적화된 반도체로, 시스템 구현 측면에서는 학습용과 추론용, 서비스 플랫폼 측면에서는 데이터센터용, 엣지디바이스용으로 구분**

○ 학습(Training)은 대규모 데이터를 기반으로 AI 알고리즘을 거쳐 지식을 습득(AI 모델 구현)하는 단계이며, 추론(Inference)은 학습한 내용을 토대로 외부 명령 등에 대한 결과를 도출하는 단계

○ 학습용 AI 반도체는 시간당 많은 데이터를 처리하는 병렬 연산 처리 능력이, 추론용은 연산 가속 및 저지연(빠른 응답)·저전력 등이 주요 성능으로 요구

- 특히, 최근 AI 데이터센터 운영에 막대한 비용이 요구됨에 따라, 제품의 전력 효율 및 확장성·유연성 등의 기술적 요구가 확대 중

○ 데이터센터용은 학습용 AI 반도체 중심에서 추론용까지 확장되고 있으며, 엣지디바이스용은 추론용 AI 반도체 중심으로 구성

❑ **AI 반도체는 연산 가속 방식에 따라 통합(on-chip) AI 반도체 및 개별(off-chip) AI 반도체로 구분[1)]**

○ AI 반도체는 기존 프로세서 연산을 가속한다는 측면에서 AI 가속기에 포함되며, 프로세서와 동일한 die에 포함되어 있으면 통합 AI 반도체, 개별 die에 존재하고 PCIe, NVLink 등 인터페이스[2)]로 연결되어 있으면 개별 AI 반도체로 분류

1) Gartner('24), Omdia('23) 등

2) 인터페이스는 서로 다른 컴퓨터 시스템 구성 요소 간 정보 교환을 위한 장치로, PCIe는 인텔의 주도하에 개발된 입출력 직렬 인터페이스이며, NVLink는 엔비디아의 GPU 및 CPU 고속 연결을 위한 장치

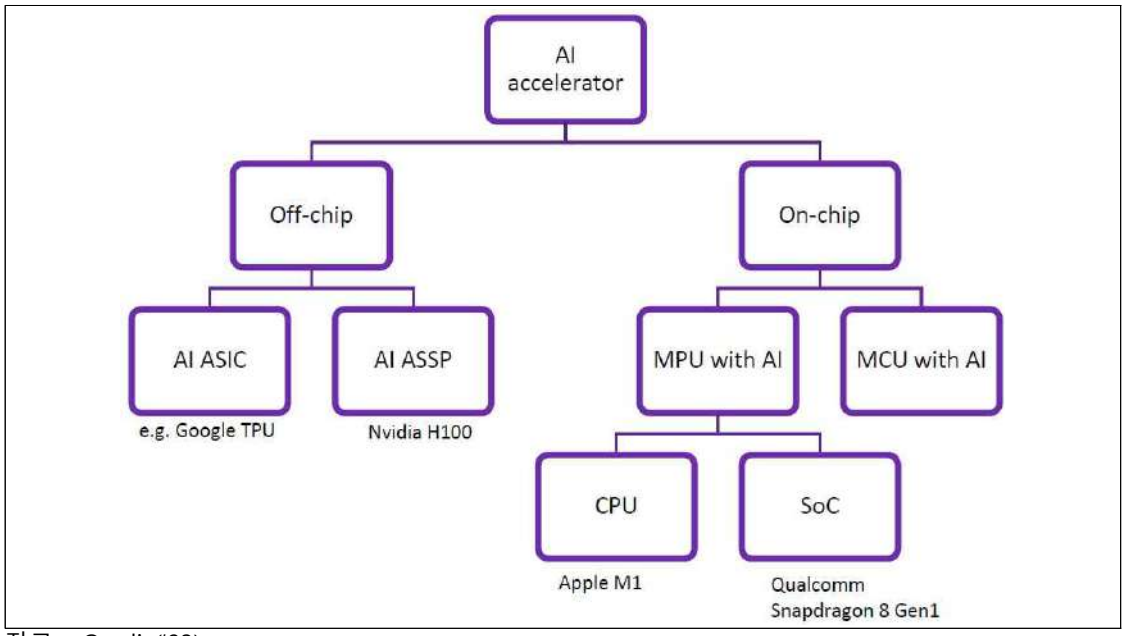
○ 개별 AI 반도체는 애플리케이션 맞춤형 반도체인 ASIC(Application Specific Integrated Circuit), ASIC의 표준형인 ASSP(Application Specific Standard Product) 방식 등으로 구분

- 구글은 AI ASIC 제품으로 자체 텐서플로우[3] 맞춤형 TPU를 개발하였으며, 엔비디아는 AI ASSP인 AI GPU(H100 등)를 고객사 앞 납품 중

○ 통합 AI 반도체는 MPU, MCU 칩 내에 NPU(Neural Processing Unit)를 포함한 SoC 형태이며, 주로 엣지디바이스용으로 사용

- 애플의 M4, 퀄컴의 스냅드래곤 8 gen3 등이 대표적인 통합 AI 반도체

〈그림 1〉 **AI 가속기의 분류**



자료 : Omdia('23)

3) '15년 오픈 소스로 공개된 구글의 소프트웨어 라이브러리로, 머신러닝 시스템으로서 사용

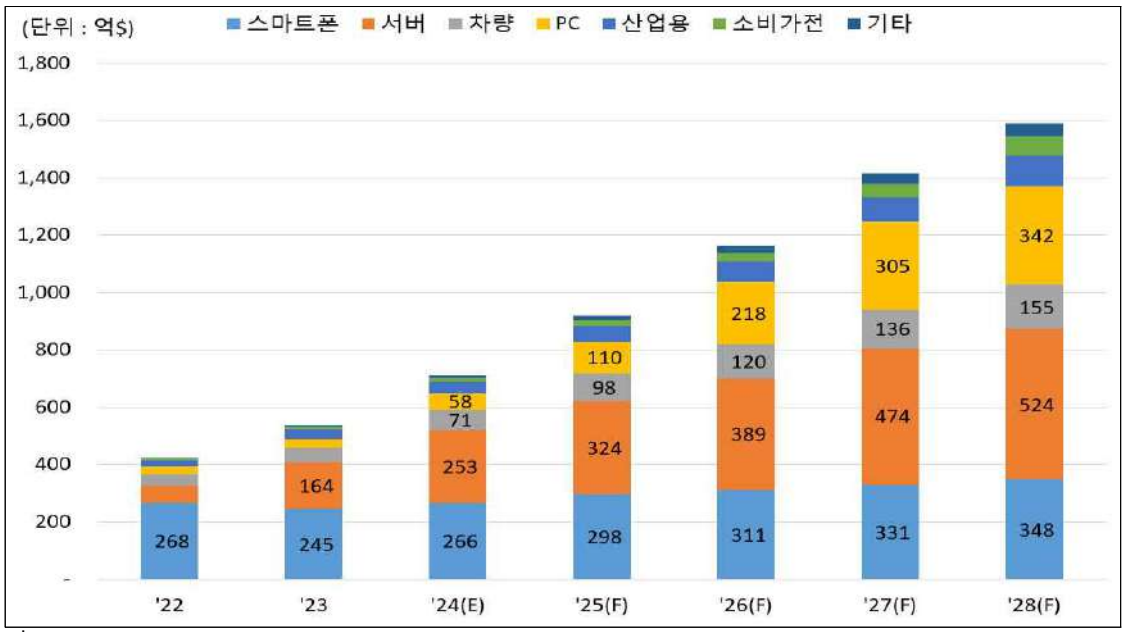
2. 시장 현황 및 전망

❑ **글로벌 AI 반도체 시장은 '23년 537억달러 규모에서 '28년 1,590억달러 규모로 향후 5년간 연평균 24% 성장할 전망(Gartner, '24.4)**

○ AI 반도체의 활용 분야는 데이터센터(서버 등) 및 스마트폰 위주에서 점차 엣지디바이스(차량, PC 등)로 확장될 전망

- 서버용 AI 반도체는 '23년 전년 대비 174.7% 급성장하였으며, '24년에도 성장세 이어져 전년 대비 54.5% 증가할 것으로 전망
- 엣지디바이스 분야에서는 PC용이 CAGR('23~'28) 62.8%로 가장 높은 성장률을 보일 것으로 전망되며, TV 등 소비가전이 41.9%로 두 번째로 높은 성장을 보일 것으로 전망

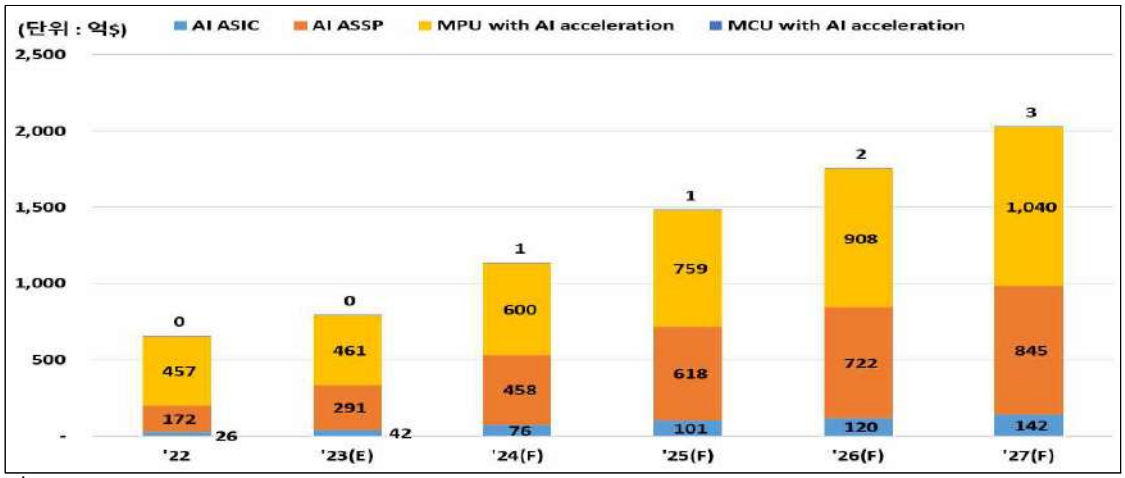
〈그림 2〉 **AI 반도체 응용 분야별 시장 전망**



자료 : Gartner('24)

○ AI 반도체를 포함하는 AI 가속기 시장은 '23년 793억달러 규모에서 '27년 2,029억 달러 규모로 성장할 전망이며, 개별 AI 반도체인 AI ASSP과 모바일 AP 등의 MPU 통합 AI 반도체가 가장 큰 비중을 차지

〈그림 3〉 AI 가속 방식별 시장 전망

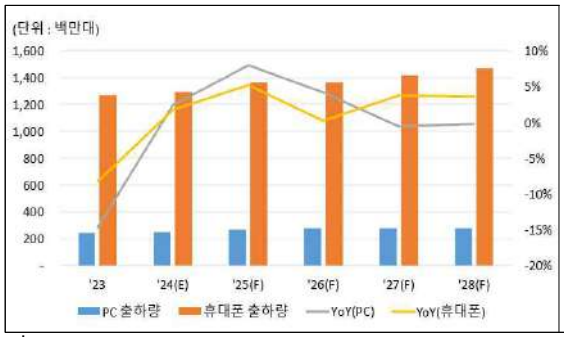


자료 : Omdia('23)

○ AI 반도체는 미래 첨단산업에 필수적인 재화이자 반도체 시장의 새로운 성장 동인이 될 것으로, 기술력과 초기시장 선점을 통한 경쟁력 확보가 필요

- 지금까지 반도체 시장의 성장을 주도했던 PC, 스마트폰의 성장률이 둔화세에 있으나 AI 분야는 고성장 추세로, '28년에는 AI 반도체가 전체 반도체 시장의 약 20%를 차지할 전망

〈그림 4〉 ICT 전방 산업 출하량 전망



〈그림 5〉 전체 반도체 시장 내 AI 반도체 비중

자료 : Gartner('24)

Ⅱ. 핵심 분야별 기술 동향

1. 프로세서 : AI GPU, NPU, 뉴로모픽 반도체

① AI GPU

❑ AI GPU는 '연산 능력(AI 학습)' 측면에서 현존하는 최고 성능의 AI 반도체로, 엔비디아(美)가 기술 선도 중

○ AI GPU의 높은 연산 능력은 기본적으로 CPU 대비 많은 병렬 연산 코어(처리 장치) 구조에 기인

- CPU는 복잡한 직렬 연산을 수행하는 수십 개 수준의 연산 코어를 보유한 반면, AI GPU는 상대적으로 단순한 병렬 연산 코어를 수천 개~만 개 이상 보유

○ 미국 엔비디아는 제한적이었던 GPU의 기능에 AI 구현에 특화된 신기술을 접목·개발하며 독보적인 AI GPU 기술로 확장

- AI 연산 특수 코어인 'Tensor Core', GPU 간 고속통신용 'NVLink', 효율적인 GPU 분할 사용을 위한 'MIG' 기술 등을 자체 개발

〈표 1〉 엔비디아 AI GPU의 주요 기술 개요

기술명	주요 내용
Tensor Core	AI 연산(행렬 연산)에 특화된 자체 개발 코어로, 최신 4세대 Tensor Core는 FP8[주)] 등 저정밀 데이터 유형까지 처리하면서도 높은 연산 정확도를 유지하도록 설계
NVLink	기존의 PCIe 등 CPU에 의존적인 주변기기 연결(인터페이스) 방식의 제한된 통신 속도, 비효율성 등을 극복하기 위해 개발한 GPU 간 직접 데이터 통신 기술. 5세대 NVLink의 경우, 최대 1,800GB/s의 대역폭 제공이 가능
MIG (Multi Instance GPUs)	하나의 GPU를 각각 메모리, 코어 등을 갖춘 독립된 Instance GPU 형태로 구분(GPU 하나로 다수의 Instance GPU를 생성)하여 GPU 활용성을 최적화하는 기술

주 : FP(Floating Point, 부동소수점), 컴퓨터 내부에서 실수를 표현하는 방식으로, FP32(2진수 32비트로 실수를 표현) FP16, FP8 등이 있으며, 실수 표현 비트수가 작아질수록 적은 메모리 사용으로 전체 연산 능력은 향상되나, 연산 정확도는 감소

자료 : 엔비디아, 업계 자료

○ 엔비디아의 AI GPU는 MLPerf[4] 벤치마크 테스트 결과 등을 감안시, 현재로서는 학습(Training) 분야에서 압도적인 성능 우위를 보유한 것으로 평가

- MLPerf 학습 분야 측정 지표 기준, 엔비디아는 H100 제품 구성으로 GPT3 모델 학습 소요 시간을 10분 이내로 달성

〈표 2〉 주요 제품별 학습 분야 MLPerf 결과(v3.1, '23.11월)

참여기업 (Organization)	제품주) (Accelerator)	제품 개수 (# of Accelerator)	학습 소요 시간(분) (GPT3 모델 기준)
엔비디아	H100(SXM-80GB)	4,096	8.57
구글	TPU-v5e	4,096	44.68
인텔	Gaudi2	384	153.58

주 : 제품은 주요 AI 가속기(AI GPU 등)만 표시하였으며, 그 외 시스템 구성을 위한 CPU 종류 등은 상이
자료 : MLCommons MLPerf 벤치마크 학습(Training) 분야

❑ AI GPU는 기존 범용 프로세서(GPU)의 변화 형태로, 다양한 AI 서비스별 최적화 제품으로서는 한계점이 존재

❍ AI GPU는 다양한 AI 서비스 구현이 가능한 '범용성'은 장점이나, 과도한 연산 성능, 높은 소비전력 등 '효율성' 측면에서는 단점을 보유

- GPU에는 AI 연산에 불필요한 그래픽 연산용 기능 등이 기본적으로 포함되어 있으며, 특히 추론 작업 시 소모 전력 대비 성능 구현의 비효율성이 큰 것으로 알려짐

❍ 다만 AI GPU는 빠른 기술 고도화, 높은 기술 성숙도 및 범용성 등으로 핵심 AI 반도체의 지위를 당분간 유지할 것으로 예상

- 주요 AI 기업들은 NPU의 필요성에 대해서는 인지하고 있으나, 단기적으로는 AI GPU 도입이 AI 서비스 선점 경쟁에 유리하다고 판단

4) '20년 결성된 산·학·연 컨소시움인 MLCommons가 주관하는 AI 시스템 성능 테스트로, 크게 학습과 추론 분야로 구분하고 다양한 기준(벤치마크)을 설정함

② NPU[5)]

❑ **NPU는 AI 연산 가속에 특화 설계된 ASIC 기반 AI 반도체로, 개별 AI 모델에 최적화되어 있어 AI GPU의 한계점 극복 가능**

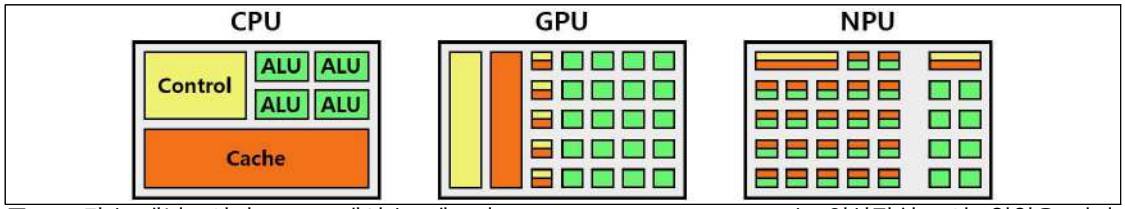
○ NPU는 설계 시점부터 특정 분야별로 특화되기 때문에, AI GPU 대비 '효율성' 측면에서 장점을 보유

- NPU는 특정 AI 연산 가속을 저지연 및 저소비전력으로 해결하여 TCO(Total Cost of Operation)를 감소시키는 것이 가장 큰 특징이자 목표
- NPU 또한 AI 학습이 가능하나, 낮은 기술 성숙도와 학습 분야에서 AI GPU의 기술적 우위 선점 등으로 주로 추론용을 중심으로 개발 중

○ NPU는 다차원 텐서 연산 특화, 온칩 메모리 방식, 저정밀 연산 등의 기술적 특징이 있으며, 기업별로 다양한 기술 및 구조의 제품을 개발 중

- NPU는 행렬 곱셈(2차원 계산), 합성곱[6)](Convolution), 활성화 함수[7)] 등 다차원(텐서) 수준의 수학적 연산에 특화
- 구조적으로는 메모리가 NPU 칩 내부에 통합되어 있는 온칩 메모리 방식을 활용하여 소형화, 저소비전력 등을 도모
- 또한 최근 인공신경망(AI 모델)의 지속적인 거대화에 대응하여 NPU용 저정밀[8)] 연산 기술 개발이 활발히 진행 중

〈그림 6〉 **CPU·GPU·NPU 구조(개념도) 비교**



주 : 단순 개념도이며, Cache(캐시)는 메모리, ALU(Arithmetic Logic Unit)는 연산장치(코어) 영역을 의미
자료 : 업계 자료

5) NPU는 AI 연산에 특화된 프로세서를 통칭하는 광의의 개념에서 시작하였으나, 최근에는 AI GPU와 대비하여 추론용 특화 AI 반도체 등 협의의 개념으로도 표현됨
6) 입력 데이터의 특징을 감지하고 추출하는데 사용되는 수학적 연산 기술
7) 인공신경망에서 입력값을 변환하는 함수로서, 활성화 여부를 조절하여 입력 처리 방향을 결정
8) 데이터 처리 시 더 적은 비트를 사용하여 연산을 수행하는 기술로, 인공신경망 경량화 기술과 함께 HW적으로는 저정밀 연산기 고집적화 기술, 메모리와 연산기의 효율적인 구조 기술 등이 연구 중임

❑ 다양한 AI 모델별 최적화된 NPU 요구 확대로 다수 업체 간 시장 선점 경쟁이 본격화될 것으로 예상

○ AI 반도체 수요는 AI 모델 학습을 위한 AI GPU 제품 중심에서 추론을 위한 NPU 중심으로 점차 변화할 것으로 예상

- 전방 산업별 AI 모델의 기술적 요구사항은 매우 다양하여, 현재 주류인 AI GPU 단순 활용은 비효율적인 요소가 다수 존재하여, 엣지디바이스용(온-디바이스(On-Device) AI[9] 등)에는 NPU가 필수적

○ NPU는 현재 경쟁우위가 존재하지 않는 新시장으로, 글로벌 대기업, 스타트업들을 포함한 다수 기업이 시장 진출에 도전

- (데이터센터용) 구글, MS 등 빅테크 기업들은 자사 AI 모델에 최적화된 NPU를 개발하며 효율적인 데이터센터 운영과 엔비디아 의존도 감소 등을 도모
- (엣지디바이스용) PC 및 모바일에 NPU 도입이 먼저 진행되면서, 기존 CPU 및 AP(Application Processor) 설계 능력을 보유한 인텔, 퀄컴, 미디어텍, 애플, 삼성전자 등이 개발을 주도 중
- 국내외 스타트업들은 상기 주요 기업 제품과 경쟁하는 동시에 중장기적으로는 자율주행, 금융, 의료, 스마트팩토리 등 타겟시장별 세분화된 제품군 개발 전략을 추진

〈그림 7〉 국내외 주요 AI 반도체 스타트업

자료 : 한국정보통신기술협회('23)

9) 클라우드 서버를 거치지 않고 스마트 기기 내에서 자체적으로 정보를 수집 및 연산하는 기술

③ 뉴로모픽(Neuromorphic) 반도체

❑ 뉴로모픽 반도체는 AI 프로세서의 궁극적인 형태로, 생물학적 뇌의 구조(뉴런·시냅스) 및 기능적 특성(초고속·초저전력) 등을 모사

❍ 신경계 구조[10]를 모사하기 위해서는 기존 방식(연산·저장 장치 분리)이 아닌 연산·저장을 동시에 수행하는 뉴로모픽 소자 기술이 요구
- 특히, 학습된 정보를 저장하고 전달하는 시냅스를 모방하기 위해 멤리스터[11](Memristor) 특성 소자에 관한 연구가 활발

❍ 또한 초고속 병렬 연산을 위한 고집적 기술과 함께 초저전력 특성 등이 동시에 요구되는 등 다양한 기술적 난제가 존재
- 현재 인텔, IBM 등 일부 기업과 학계 중심의 연구가 진행되고 있으나, 단기간 내 상용화는 요원할 것으로 예상
- 다만, 중간단계로의 PIM(Processing In Memory) 컴퓨팅[12] 기술개발 등 뉴로모픽에 근접하기 위한 업계의 노력은 지속

〈표 3〉 인텔과 IBM의 뉴로모픽 반도체 칩

기업명	주요 제품	주요 특징
인텔	Loihi 2세대('21년)	· 1세대('17년) 제품 대비 면적을 절반 수준으로 축소 · 100만개의 뉴런과 1억 2천만개의 시냅스로 구성
IBM	TrueNorth('14년)	· 104만개의 뉴런과 2억 5천6백만개의 시냅스로 구성
	NorthPole('23년)	· TrueNorth 대비 속도 및 에너지 효율을 20배 이상 개선

자료 : 업계 자료

10) 인간의 뇌는 연산·저장을 동시 수행하는 1,000억개 이상의 뉴런(신경세포)과 100조개 이상의 병렬로 연결된 시냅스(접합부)를 통해 동시다발적으로 정보를 송·수신하며 처리함

11) 메모리(Memory)와 저항(Resistor)의 합성어로, 입력 전압에 따라 내부 저항값이 변화하며 이를 이용하여 정보를 저장하거나 처리하는 소자로 대표적으로 RRAM(Resistive RAM), PRAM(Phase-change RAM) 등이 있음

12) 기존 프로세서(CPU) 중심의 컴퓨팅 구조에서 변화하여 메모리 중심 컴퓨팅 구조의 개념이며, 궁극적으로는 프로세서와 메모리의 통합 설계가 목표

2. 메모리 : HBM, PIM

❑ AI 연산에 있어 '메모리 병목 현상'의 해결이 주요 과제로 대두되면서 AI 반도체向 메모리의 중요성이 부상

○ 메모리 병목 현상(또는 메모리 벽(Memory Wall))은 메모리에서 프로세서로 데이터를 전송할 때 많은 시간이 소요되는 현상을 말하며, 프로세서의 연산 속도 향상 대비 상대적으로 느린 메모리 성능 개선에 기인

○ HBM(고대역폭메모리)이 현재 AI 반도체용으로는 가장 적합한 메모리 방식인 것으로 알려짐
- HBM은 메모리 제품 중 유일하게 업계의 요구 성능[13]에 부합
- AI GPU에 HBM 사용은 필수적이며, NPU에서의 HBM 도입 또한 지속적으로 확대되고 있음

○ 또한 HBM에서 발전하여 메모리 내에 연산 기능을 추가한 PIM(Processing In Memory) 기술이 주목
- PIM 기술은 프로세서와 메모리 영역 간의 거리를 최대한 축소 또는 프로세서와 메모리를 통합 설계하여 저지연 및 높은 에너지 효율 등을 도모

① HBM(High Bandwidth Memory)

❑ HBM은 개별 DRAM 칩을 고밀도 적층하여 대용량·고대역폭이 특징

○ HBM은 8~12개의 DRAM 칩을 적층한 후 TSV[14] 공정 기술 등을 적용하는 방법으로 입출력(I/O) 핀 수를 극대화하여 고대역폭을 확보
- DRAM 제품군 중 GDDR(GPU용 DRAM)의 주요 특징인 '고대역폭'을 크게 개선함과 동시에 '대용량'의 장점까지 보유

13) 개별 AI 반도체당 메모리 대역폭 1TB/s 이상

14) TSV(Through Silicon Via) : 적층된 개별 칩에 수천 개의 미세한 구멍을 뚫어 상층과 하층 칩의 구멍으로 수직 관통하는 전극을 연결하는 기술

〈표 4〉 HBM 및 GDDR 성능 비교

구분		HBM3	GDDR6
용도		AI GPU, NPU	일반 GPU
대역폭		819.2GB/s (=1,024x0.8GB/s)	64GB/s (=32x2GB/s)
	입출력 Pin 수	1,024	32
	Pin 당 속도	0.8GB/s	2GB/s
용량		24GB	2GB

주 : 평균적인 성능 기준
자료 : 업계 자료

○ 12단 적층 HBM3E(5세대) 제품까지 개발 완료('24년 양산 예정)되었으며, '26년 양산 예정인 HBM4는 16단 적층, 입출력 핀 수 2,000개 이상 수준으로 구현될 것으로 예상

❑ HBM은 고가로 수요처가 제한적이었으나 AI 반도체의 핵심 제품으로 부상하며, 글로벌 DRAM 3社간 기술·양산 경쟁 격화

○ 단기적으로는 시장 선점 효과, 생산능력 등 감안시 국내 업계가 HBM 시장을 주도[15]할 것으로 예상
- HBM 기술을 선도[16]하던 SK하이닉스가 HBM3(4세대)를 엔비디아 앞 독점 공급하며 시장을 선점하였으며, 삼성전자는 '24년 내 공식 납품 예상

○ HBM 수요 급증이 전망되는 가운데, HBM3E부터 마이크론의 본격적인 참전으로 국내와의 기술 및 고객사 확보 경쟁 심화 예상
- 마이크론은 '24.2월 엔비디아 차세대 AI GPU(H200, '23.4Q 공개)용 HBM3E 제품 양산을 발표
- HBM은 제조 난이도가 높은 제품으로, 수율 개선(현재 60~70% 추정) 정도가 향후 기업별 경쟁력 결정에 핵심으로 작용할 것으로 판단

15) 현재 SK하이닉스, 삼성전자, 마이크론(美)의 HBM 시장점유율은 각각 50%, 40%, 10% 수준으로 추정
16) SK하이닉스의 HBM 개발은 '08년부터 미국 AMD와의 공동개발로 시작되었으며, '13년 세계 최초로 HBM 1세대 개발에 성공

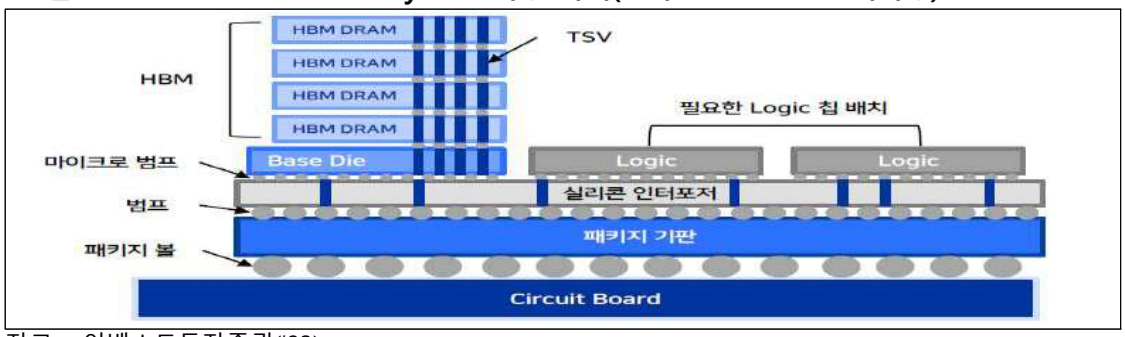
② PIM(Processing In Memory)

❑ **PIM은 메모리 내에서 연산까지 수행하는 새로운 개념의 컴퓨팅 기술로, 크게 Near-Memory PIM과 In-Memory PIM으로 구분**

○ Near-Memory PIM은 기존 연산장치와 메모리 제품 간 거리를 최대한 가깝게 배치 후 패키징하는 기술

- 해당 제품의 변화가 요구되지 않아 상용화가 확대되고 있으나, 최종적인 PIM의 형태는 아닌 전단계 수준의 기술임
- 대표적으로 인터포저[17]를 활용한 로직칩(AI GPU 등)과 HBM의 2.5D 패키징 기술 등이 해당됨

〈그림 8〉 **Near-Memory PIM 적용 사례(로직칩-HBM 2.5D 패키징)**



자료 : 이베스트투자증권('23)

○ In-Memory PIM은 최종형태의 PIM 기술이며, 세부적으로 In-Memory-Array와 In-Memory-Cell로 구분

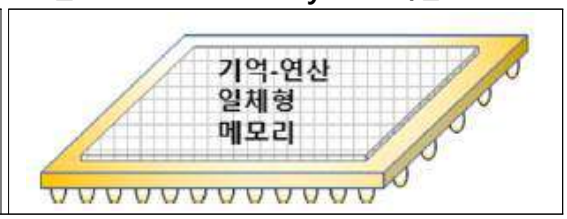
- In-Memory-Array는 메모리 영역 내에 연산장치가 배치되어 연산장치와 메모리 간 물리적 거리 축소를 심화
- In-Memory-Cell은 메모리 셀(메모리의 단위 구조) 자체에서 데이터 저장 및 연산 기능을 모두 수행하는 PIM 기술의 최종 목표로, 非폰노이만 구조[18]에 해당

17) 실리콘 등으로 제작한 고밀도 배선 층으로 칩과 패키징 기판 사이에 가교역할을 하며 연산 칩과 메모리 간의 거리 축소에도 이점이 있음

18) 폰노이만 구조는 데이터를 저장하는 메모리와 연산장치가 분리되어 있어, 데이터 연산을 위해서는 메모리의 모든 데이터가 연산장치로 이동해야 하나, 非폰노이만 구조인 PIM 기술은 메모리 내에 연산장치가 있어, 명령을 받으면 메모리 내에서 연산 후 결과값만 프로세서로 전송

〈그림 9〉 In-Memory-Array 개념도

〈그림 10〉 In-Memory-Cell 개념도



자료 : 정보통신정책연구원(KISDI)('23)

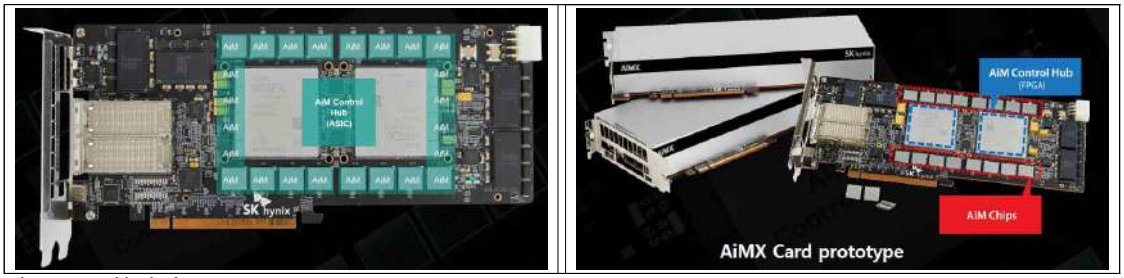
❑ In-Memory PIM 기술의 현 수준은 In-Memory-Array에 가까우며, 메모리 제조사들을 중심으로 관련 제품이 개발되고 있음

○ SK하이닉스는 '22년 'GDDR6-AiM(Accelerator-in-Memory)' 제품을 개발하였으며, '23년 상기 제품들로 구성된 'AiMX' AI 가속기카드 시제품을 공개

- 공개한 자료에 따르면, 동일 AI 모델 적용 추론 환경에서 AiMX 기반 시스템이 AI GPU 기반 대비 응답속도는 13배 빠르고 전력 소모는 17% 감소할 수 있는 것으로 제시

○ 삼성전자는 기존 대비 AI 가속기 성능은 2배 향상, 전력 소모는 50% 감소시킬 수 있는 'HBM-PIM' 제품을 개발

〈그림 11〉 In-Memory PIM 적용 사례(SK하이닉스 AiMX)



자료 : SK하이닉스

○ PIM의 최종형태인 In-Memory Cell 기술은 연구 단계 수준이며, 주로 SRAM[19]을 이용한 In-Memory Cell 연구가 활발하게 진행 중

19) SRAM(Static Random-Access Memory)은 DRAM과 같이 휘발성 메모리이나, 전원이 공급되는 한 데이터가 보존되며, DRAM보다 데이터 처리 속도가 빠름

3. 인터페이스 : CXL(Compute eXpress Link)

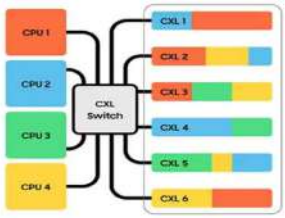
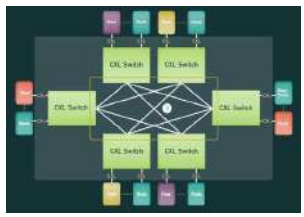
❑ CXL[20]은 프로세서, 메모리 등을 효율적으로 연결하는 차세대 인터페이스 기술이자 규격으로, 메모리 용량 확장과 공유 기술이 가장 큰 특징

○ CXL 기술은 기존 PCIe[21] 인터페이스의 확장 기술로, 대용량 데이터(메모리)의 효율적인 사용이 필요한 AI 시스템에서 중요성이 부상

- 기존 인터페이스 기반 시스템에서 대용량 메모리 시스템 구축 시, 프로세서의 대용량 메모리 수용 한계와 오버프로비저닝[22] 문제 등 메모리 사용에서의 비효율성이 심화

○ CXL 기술 표준은 '19년 1.0에서 '23.11월 3.1 버전까지 발표되었으며, 메모리 용량 확장에서 확장 및 공유까지 기술 개념이 확대

〈표 5〉 CXL 버전별 개념도 및 주요 기술 내용 변화

구분	CXL 1.1	CXL 2.0	CXL 3.0[주]
개념도			
주요 기술 내용	프로세서 영역 內에 기존 메모리 외 확장된 CXL 메모리 추가	프로세서 영역 外에 메모리 풀링(Pooling) 시스템 연결	CXL 메모리 간 네트워킹 구조 (CXL Fabric) 구현
기술의 시사점	메모리 확장 (개별 Server level)	메모리 확장 (Server Rack level)	메모리 확장 및 공유 (Rack to Rack)

주 : 3.0 버전은 '23.5월 발표되었으며, '23.11월 CXL Fabric을 개선 및 확장한 3.1 버전을 발표
자료 : 삼성전자, CXL 컨소시엄

20) '19년부터 인텔 주도로 데이터센터, 서버, 프로세서, 메모리 업체들이 참여한 CXL 컨소시엄이 공동 개발을 시작

21) PCIe(Peripheral Component Interconnect express)는 현재 상용화된 프로세서와 주변장치들을 연결하는 인터페이스 기술 또는 규격

22) 시스템 구성 요소별로 컴퓨팅 자원(메모리 등)에 대해 실제 예상 사용량보다 여유분을 미리 확보(설정)하는 기술로, 예측할 수 없는 트래픽 증가 등에 대비하는 장점도 있으나 자원 낭비와 비용 증가 등의 문제점도 존재

❑ '24년부터 CXL 시장이 본격 개화될 것으로 예상되며, CXL 메모리는 AI 반도체 向 메모리 시장에서 HBM에 이어 주요 제품으로 부상 전망

○ '24년부터 CXL 2.0 이상을 지원하는 프로세서가 출시되며 본격적인 CXL 생태계 개화 전망

- 인텔은 CXL 2.0을 지원하는 서버용 CPU(시에라포레스트 등)를 '24년 내 최초로 출시할 예정
- 엔비디아 또한 CXL 컨소시엄에 참여하고 있으며, 자체 인터페이스 기술인 'NVLink' 기술을 확장하여 CXL과 호환되도록 개발 중

○ 메모리 업계가 프로세서 대비 선제적으로 CXL 제품을 개발하며 시장 선점 준비 중

- 삼성전자는 업계 최초로 CXL 2.0을 지원하는 128GB CXL DRAM 개발 완료를 발표('23.5월)하며 '23년 말부터 양산을 시작
- 이에 대응하여 미국 마이크론은 '23.8월 관련 제품을 공개하였고, SK하이닉스는 '24년 상반기 내 고객사 인증을 완료한 후 하반기부터 양산할 것으로 알려짐

〈그림 12〉 CXL 컨소시엄 참여 주요 기업

자료 : CXL 컨소시엄

<그림 13> 삼성전자의 CXL 2.0 메모리 제품

자료 : 삼성전자

○ CXL 메모리는 AI 시스템에서 메모리 용량을 유연하게 확장할 수 있다는 강점이 있어 HBM과 함께 AI 반도체 向 주요 메모리로 부상 전망

- CXL 메모리 시장[23] 규모는 '23년 14백만달러에서 '26년 3,400백만달러까지 급성장 전망(YOLE('23.9))

23) CXL DRAM 및 CXL Memory Expander(CXL DRAM 및 컨트롤러 포함 모듈) 기준

Ⅲ. AI 반도체 산업 동향

❑ **AI 반도체 산업은 설계 위주의 팹리스, 제조 분야의 파운드리로 분업화되어 있으며, 각 분야 주요 업체들은 사업영역 확장을 통해 AI 반도체 기술 확보 경쟁에 대응 중**

○ (팹리스) 하드웨어 설계 위주였던 기존 팹리스 기업들은 자사 제품에 적합한 AI 서비스 개발 환경을 지원하거나 기업 인수·합병 등의 전략을 통해 사업 영역을 확장

- 엔비디아는 자사 제품에서만 구동되는 GPU 프로그래밍 지원 소프트웨어 'CUDA', 대규모 언어모델(Large Language Models, LLM) 구축 지원 플랫폼 'NeMo(Nvidia Enterprise Modular AI)' 등을 지원하며 락인(Lock-in) 효과[24] 극대화
- AMD는 FPGA 전문 기업 '자일링스'를 인수하며 범용 CPU, GPU 위주의 사업에서 AI 및 IoT 분야로 포트폴리오를 확장

〈표 6〉 **주요 반도체 기업의 AI 기업 인수·합병 현황**

기업명	인수 기업	시기
엔비디아	런AI(Run:ai, 이스라엘) - AI 및 클라우드 인프라 관리를 위한 SW 플랫폼 구축	'24.4 인수 계약
AMD	자일링스(Xillinx, 미국) - 재프로그래밍이 가능한 FPGA 및 적응형 SoC, AI 추론 엔진 및 소프트웨어 전문 기업	'22.2
	노드닷AI(Nod.ai, 미국) - 오픈소스 AI 소프트웨어	'23.10
인텔	하바나랩스(Habana Labs, 이스라엘) - 데이터센터용 딥러닝 가속기 개발	'19.12
	시그옵트(SigOpt, 미국) - AI 모델링 및 시뮬레이션 지원	'20.11

자료 : 업계 및 언론 자료 종합

24) 고객을 묶어둔다는 의미로, 특정 재화나 서비스를 선택 후 기존 선택을 바꿀 때 발생하는 전환 비용이 클 때 작용

○ (파운드리) 글로벌 파운드리 Big3(TSMC(대만)·삼성전자·인텔(美))는 전공정[25] 위주의 사업에서 첨단 패키징 기술 등 AI 반도체 제조에 요구되는 후공정 분야로 투자를 확대

- TSMC는 첨단 패키징 기술인 CoWoS[26]를 토대로 시장을 선점하였으며, '23년 기준 월 1.5만장 수준이던 패키징 공급 능력을 '24년에는 2배, '25년에는 최대 5.5만장 수준으로 확대하는 등 투자를 지속할 계획
- 삼성전자는 '23년 어드밴스드패키징(AVP) 사업팀을 신설하였으며, 첨단 패키징 기술 개발에 18억 달러 규모를 투입
- 인텔은 '23년 47억 달러 이상을 첨단 패키징 설비에 투자하였으며, '24년 미국 뉴멕시코주에 3D 패키징 기술이 포함된 '팹9(Fab9)'을 오픈

❑ 클라우드 업체 등 AI 반도체의 주요 수요 기업들은 AI 반도체 수급 안정, 서버 운영 비용 절감, 자사 서비스 최적화 등을 위해 자체 AI 반도체를 개발하는 추세

○ (클라우드) 글로벌 클라우드 Big3(아마존(美)·마이크로소프트(美)·구글(美))는 엔비디아 의존도를 낮추고 연산 효율성을 높이기 위해 자체 칩 개발

- 아마존은 '15년 이스라엘 반도체 기업 '안나푸르나 랩스'를 인수하며 본격적으로 자체 반도체를 개발해왔으며, 데이터센터와 AI 음성인식 서비스에 자체 반도체를 적용 중
- 마이크로소프트는 오픈AI社와 협력하여 데이터센터 워크로드 최적화[27]를 통해 연산 및 에너지 효율을 향상시킬 수 있는 자체 AI 반도체를 개발
- 구글은 알파고에 탑재된 TPU(Tensor Processing Unit) v1('15년 도입)을 시작으로 텐서플로우 연산에 최적화된 TPU를 개발 중

○ (엣지디바이스) 테슬라(美), 삼성전자 등 디바이스 기업에서는 온-디바이스AI 시장 개화에 대응하여 자사 제품에 적합한 AI 반도체를 개발

25) 반도체 웨이퍼에 회로를 증착하는 과정

26) CoWoS(Chip on Wafer on Substrate)는 2.5D 패키징 기술로 인터포저 위에 반도체 다이(CPU, GPU, I/O, HBM 등)를 수평 배치하는 기술

27) HW, 네트워크 및 애플리케이션 성능을 극대화하기 위해 컴퓨팅 리소스를 할당하는 것으로, HW, SW 및 운영 체제 업그레이드, 데이터 분배 및 스토리지 관리, 백업 프로세스 등이 포함

- 테슬라는 차량 내부에 FSD(Full-self Driving) 컴퓨터, GPU 및 자체 개발한 NPU를 SoC 형태로 탑재하였으며, 완전자율주행 소프트웨어 구축을 위해 슈퍼컴퓨터 '도조(Dojo)'에 자체 AI 반도체 'D1' 탑재
- 삼성전자는 '24.1월 자체 개발한 모바일 AP '엑시노스 2400'이 탑재[28]된 AI 스마트폰 '갤럭시S24 시리즈'를 출시하였으며, 실시간 통화 번역 등 온-디바이스 기반 생성형 AI 기능을 탑재

〈표 7〉 **클라우드 Big3의 AI 반도체 개발 현황**

기업명	주요 제품	공개 시기
아마존	인퍼랜시아(Inferentia) 2세대(추론용)	'22.12
	트랜티움(Traintium) 2세대(학습용)	'23.11
마이크로소프트	마이아(Maia) 100(GPU) - AI 학습 및 추론 워크로드용	'23.11
	코발트(Cobalt) 100(CPU) - 클라우드 워크로드용	'23.11
구글	TPU(Tensor Processing Unit) v6	'24.5

자료 : 업계 및 언론 자료 종합

❑ NPU 등 차세대 AI 반도체 분야는 현재 경쟁 우위가 존재하지 않는 시장으로, 국내외 팹리스 스타트업의 진출이 확대 중

○ 팹리스 스타트업은 전력 소모, 연산 효율성 등 기존 범용 AI 가속기의 단점을 보완하고 新시장을 개척하기 위해 저전력·고성능, 엣지디바이스 AI 등에 특화된 AI 반도체를 개발 중

○ 대표적인 기업으로는 Cerebras systems(美), Sambanova systems(美), Graph Core(英) 등이 있으며, AI 엔진 시스템, 자동화 시스템 및 신약 개발 등에 적합한 반도체를 개발 중
- 한국에서는 리벨리온이 금융 거래용 NPU '아이온', 모빌린트가 영상처리에 특화된 NPU '에리스'를 개발

28) 갤럭시S24 일반 및 플러스 모델에는 '엑시노스 2400', 울트라 모델에는 퀄컴의 '스냅드래곤 8 Gen3 for Galaxy'를 탑재

〈표 8〉 국내외 주요 팹리스 스타트업 현황

기업명	주요 내용	국가
Cerebras systems	- 웨이퍼 스케일 엔진(Wafer Scale Engine, WSE)으로 불리는 세계 최대 크기의 AI 가속기를 개발 - 가속기의 모든 요소가 하나의 칩에 내장되어 있어, 저지연, 저전력 등에 장점이 있음	미국
Sambanova systems	- FPGA와 같이 재프로그래밍이 가능한 NPU 개발 - HW·SW를 통합한 AI 시스템(DataScale SN30), AI 모델 구축 및 서비스 플랫폼(SambaNova Suite) 등 개발	미국
GraphCore	- IPU(Intelligence Processing Unit, 지능처리장치)주) 설계 - 신약 개발 및 의료공학, 자동화 시스템, 5G 통신 트래픽 관리용 등	영국
퓨리오사AI	- 데이터센터에 탑재되는 AI 추론용 NPU 등 - 국내 NPU 최초로 HBM3 탑재	한국
리벨리온	- 처리 속도가 빨라 실시간 트레이딩 등 금융 거래에 적합한 NPU '아이온' 개발 - AI 추론용 저전력 NPU '아톰' 개발	한국
모빌린트	- 엣지디바이스에 적합한 저전력 AI 반도체 '에리스', '레귤러스' 등 개발	한국
딥엑스	- 엣지디바이스에 적합한 저전력 AI 반도체를 개발 중으로, 지능형 영상 분석, 보안 시장 공략 중	한국

주 : 프로세서 내에 메모리가 내장된 아키텍처
자료 : 업계 및 언론 자료 종합

Ⅳ. 시사점

❑ 한국은 파운드리 및 메모리 반도체 시장 지위 양호한 편이나 비메모리 및 팹리스 역량은 부족한 상황으로, AI 반도체 기술수준은 최고기술국(美) 대비 90.7% 수준

○ '23년 파운드리 시장은 TSMC가 절반 이상을 점유하는 가운데, 삼성전자는 세계 2위 수준으로 UMC(대만), Globalfoundries(美), SMIC(中) 등이 경쟁 구도

〈표 9〉 주요 기업 파운드리 시장 점유율

(단위 : %, %p)

순위	업체명	국가	'22	'23	증감
1	TSMC	대만	58.1	60.1	+1.8
2	삼성 파운드리주)	한국	7.9	7.4	△0.6
3	Globalfoundries	미국	7.2	6.4	△0.4
4	UMC	대만	6.2	6.2	△0.3
5	SMIC	중국	5.6	5.5	+0.2

주 : 삼성 파운드리의 내부 매출 제외
자료 : Gartner('24)

○ 한국은 메모리 반도체에서 우수한 시장 지위를 확보하고 있으나, 비메모리 및 팹리스 분야에서는 선도국 대비 경쟁력 열위[29)]
- 한국의 글로벌 메모리 점유율은 60%인 반면, 비메모리 시장 내 점유율은 3%에 불과하여 전체 반도체에서는 11%를 차지
- 또한, 한국은 팹리스 100위 기업('23년 매출 기준) 내에 LX세미콘(16위), 서울반도체(30위) 2개社만 존재하는 등 선도국 대비 팹리스 역량이 열위

○ 한국은 우수한 메모리 역량에 기반한 지능형 메모리(PIM 등) 기술개발, 팹리스 스타트업의 성장 등으로 AI 반도체 분야에서 미국과의 기술격차를 좁혀가는 추세이며, 기술 수준은 글로벌 3위 수준[30)]
- '22년 기술수준(%) : (미국) 100 > (중국) 92.3 > (한국) 90.7 > (유럽) 89.9 > (일본) 85.8
- 최고기술국(美) 대비 기술수준 추이(%) : ('19) 87.9 → ('20) 90.2 → ('22) 90.7

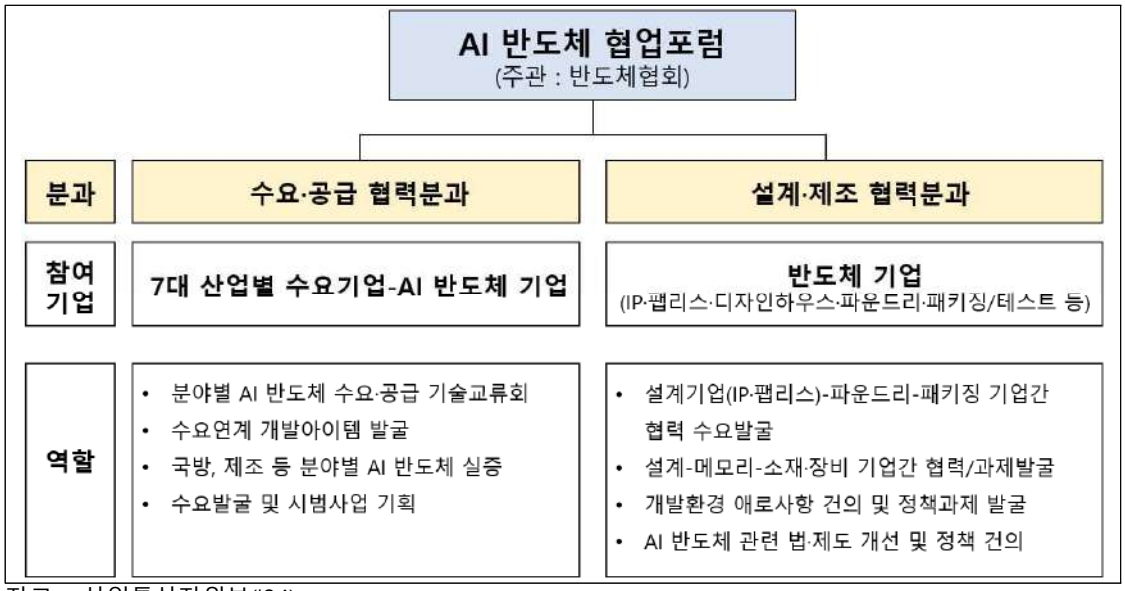
29) Gartner('24)
30) IITP, ICT 기술수준조사 및 기술경쟁력분석 보고서('21~'24)

❑ 한국 정부와 기업들은 국산 AI 반도체 개발의 중요성을 인식하고, 생태계 구축과 경쟁력 확보를 위해 노력 중

○ 정부는 '23.6월 'K-클라우드 프로젝트' 1단계[31]에 착수하였으며, 국내 CSP 3社(네이버클라우드, KT클라우드, NHN클라우드)와 팹리스 스타트업, AI서비스社와 함께 국산 NPU 실증 데이터센터 구축을 추진

○ 또한, 정부는 「AI 반도체 협업포럼('24.4)」을 통해 ① 분야별 AI 반도체 실증 및 수요매칭 기획, ② 설계기업(IP·팹리스)·파운드리·패키징 기업간 협력방안 기획 등을 지원할 계획

〈그림 14〉 AI 반도체 협업포럼 추진 체계도



자료 : 산업통상자원부('24)

31) 「국산 인공지능 반도체를 활용한 K-클라우드 추진방안('22.12)」의 일환으로 (1단계, '23~'25) NPU → (2단계, '26~'28) 저전력 PIM → (3단계, '29~'30) 극저전력 PIM 단계로 구성

〈표 10〉 '24년 AI 반도체 주요 지원사업

구분	내용	주관
R&D	(NPU) 차세대지능형반도체 기술개발사업('20~'29) - 서버·모바일·엣지 분야에 활용 가능한 고성능·저전력 NPU 개발 지원	산업부·과기정통부
	(PIM) PIM AI 반도체 핵심기술개발사업('22~'28) - 프로세서와 메모리를 융합한 PIM 반도체 개발 지원	산업부·과기정통부
인프라	시제품제작 지원 - 첨단칩 시제품 제작 지원을 위한 10nm 이하 초미세 공정 국비 지원 신설('24) 및 지원규모 확대(('23) 24억원 → ('24) 50억원))	산업부
	시험·검증 지원 - '시스템반도체 검증지원센터'를 신설해 AI 반도체 시험·검증장비 구축 및 서비스 제공	산업부
	K-클라우드('23~'30) - 국산 AI 반도체 기반 클라우드 데이터센터 구축 실증 및 이에 특화된 HW·SW 기술생태계 조성	과기정통부
수요연계	COMPASS - 수요-공급기업간 온-디바이스 AI 반도체 제품개발 매칭 시 수시 선정평가를 통해 시제품제작 등 총개발비 50%를 과제비로 지원	산업부
인력·금융	AI 반도체 대학원 - 서울대·한양대·KAIST에 AI 반도체 특화 심화교육 과정, 기업 인턴십, 글로벌 우수대학과 공동연구 등 운영	과기정통부
	반도체 생태계 펀드('24.4~) - 설계, 소·부·장 등 반도체 기업 지원을 위해 조성한 '반도체생태계 펀드(3,000억원, '23.7 조성) 집행	산업부

자료 : 산업부, 과기정통부('24)

○ 국내 대기업들은 AI 패권 경쟁에 대응하기 위해 팹리스 스타트업에 투자하며 자사 기기 및 서비스에 특화된 AI 반도체를 개발 중

- SK ICT 연합(SK텔레콤, SK하이닉스, SK스퀘어)과 한화 그룹은 지분 100%를 보유한 팹리스 자회사 '사피온', '뉴블라', '비전넥스트' 등을 설립하였으며, LG전자는 자사 제품에 AI 기술을 적용하기 위해 '에임퓨처'에 투자

〈표 11〉 국내 대기업-팹리스 스타트업 투자 현황

팹리스	대기업	대기업 지분(%)	투자 내용
사피온	SK텔레콤 SK하이닉스 SK스퀘어	62.5 25.0 12.5	데이터센터용 AI 반도체
뉴블라	한화그룹	100.0	군사·항공·우주산업 AI 반도체
비전넥스트	한화그룹	100.0	영상 보안 등 Vision AI 반도체
리벨리온	KT	13.0	데이터센터용 AI 반도체
에임퓨처	LG전자	11.7	가전 등 온디바이스 AI 반도체
보스반도체	현대차그룹	미정주)	전기차 및 자율주행차량용 AI 반도체

주 : 조건부지분인수계약으로 후속투자 시점에 재평가하여 지분 결정
자료 : 언론 기사 종합

❑ AI 반도체 기술력 확보를 위해서는 하드웨어 개발 외에도 이에 대응하는 클라우드, AI 애플리케이션 등 수요처와의 유기적 연계를 통한 생태계 구축이 중요

○ 미래 첨단 기술의 경쟁력 강화를 위해서는 국산 AI 반도체의 유의미한 실증 사례 축적과 전문인력 양성 등 장기적 관점의 민·관 협력을 지속해야 할 것임

참고문헌

[국문자료]

과학기술정보통신부(2023), "인공지능 반도체 Team Korea, 'K-클라우드 프로젝트' 1단계 본격 착수" 보도자료('23.6.26.)

박영준(2020), "AI 반도체 시장 동향 및 우리나라 경쟁력 분석", ETRI

산업통상자원부(2024), "인공지능(AI) 반도체 시장 선점을 위해 수요·공급기업이 함께 하는 「AI 반도체 협업포럼」 출범" 보도자료('24.4.2.)

신창환, 김영우(2022), "AI 반도체 생태계 분석", NIA

이미혜(2024), "AI 반도체 시장 현황 및 전망", 한국수출입은행

이선재(2022), "AI와 AI반도체 생태계 특징 및 시사점-팹리스 스타트업을 중심으로", ETRI

전황수, 최새솔, 민대홍(2024), "글로벌 파운드리 Big3의 첨단 패키징 기술개발 동향", ETRI

정보통신기획평가원(2024), "2022 ICT 기술수준조사 및 기술경쟁력분석 보고서"

홍석우(2024), "지능형(AI) 반도체 산업 현황과 관련 기술 동향", IITP

[영문자료]

Gartner(2024), "Forecast: AI Semiconductors, Wordwide, 2022-2028"

_____(2024), "Semiconductors and Electronics Forecast Database, Worldwide"

_____(2024), "Forecast: Semiconductor Foundry Revenue, Supply and Demand, Worldwide"

_____(2024), "Market Share: Semiconductors by End Market, Worldwide, 2023"

Omdia(2023), "Processors for Graphics and AI Market Tracker – 2H23 Analysis"