

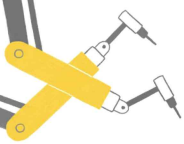
기술동향

2023

# 인공지능 반도체

KISTEP 성장동력사업센터 채명식 · 이호윤





# Contents

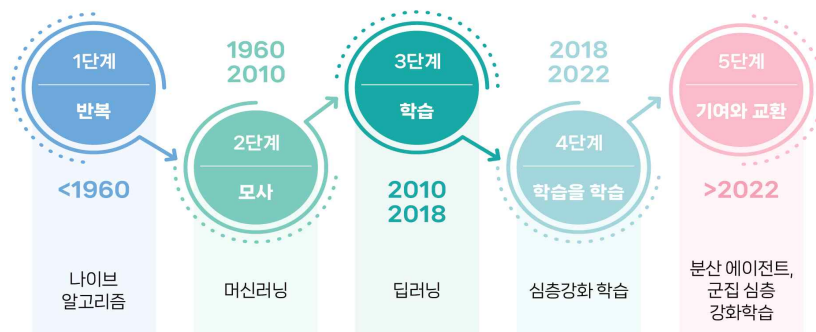
 제1장 배경 .....	1
 제2장 기술동향 .....	6
 제3장 산업동향 .....	11
 제4장 정책동향 .....	24
 제5장 R&D 투자 동향 .....	32
 제6장 결론 및 제언 .....	41

# 제1장 배경

최근 주요 분기\*에 따른 인공지능 기술(SW·알고리즘)의 비약적인 성능향상과 패러다임의 변화로 인공지능 반도체 등 HW 기술 역시 급격한 발전을 거듭

\*'06년 딥러닝 방식의 부상과 '16년 알파고의 성공 등

- 과거 인공지능 SW·알고리즘은 기계학습 기반 인식·추론 중심이었으나, 단순한 모델로도 더 뛰어난 성능을 갖는 딥러닝의 등장으로 맞춤형 HW(반도체) 개발 필요성·수요 급부상
  - '10년 전후로 인공지능 학습·구현 패러다임이 기계학습에서 딥러닝으로 전환되었으며, '16년 알파고 이후 다양한 산업 분야(모빌리티·모바일·제조·금융·컨텐츠 등) 활용 확산
    - ※ Sandeep Mittal(2022)의 분석에 따르면 '20년 이후로는 엣지디바이스를 활용한 분산 컴퓨팅 기법이 적용된 대규모 인공지능이 화두에 오르내리고 있으며 '22년 이후로 이는 IoT 응용에도 혁신적인 변화를 가져올 것으로 전망
  - 딥러닝은 과거 기계학습 대비 단순한 연산 메커니즘을 바탕으로 지속적인 알고리즘 최적화가 가능하며, 이를 구현하기 위한 목적·맞춤형 인공지능 반도체 수요가 지속해서 확대 중
    - ※ 이에 따른 대규모 학습 데이터에 대한 인공지능 연산력 수요 급증으로 HW가 갖는 발열·전력 소모 등 이슈가 존재하며, 이를 극복하기 위한 고성능·저전력 구현 필요성이 증대




[그림 1] 인공지능 기술의 발전 흐름

\* 출처: Sandeep Mittal, 2022 재구성

- 인공지능 반도체는 시스템반도체(비메모리 반도체)의 일종으로 인공지능 기술의 비약적인 발전과 이에 따른 수요 급증으로 주문형 소자를 중심으로 높은 시장성이 전망됨
  - ※ 인공지능 반도체 산업 역시 시스템반도체 산업이 갖는 설계·제조와 분리, 목적에 따른 주문형 생산 등 특성을 가짐
- 인공지능 반도체는 활용 목적에 따라 크게 ▲데이터센터(서버)와 ▲엣지컴퓨팅용으로 구분할 수 있으며, 앞서 언급한 '시스템반도체 산업 특징'은 주로 엣지컴퓨팅 분야에 한정

※ 현재 데이터센터용 인공지능 반도체는 기존 CPU와 GPU 제품의 조합이 일반적이며, 점차 엣지컴퓨팅 분야와 같이 특수 목적형 인공지능 반도체로 대체될 것으로 전망

- 인공지능 기술의 활용·확산에 따라 신규 인공지능 반도체가 개발·출시 중으로, 유수 매체에 따르면 시스템반도체를 비롯한 반도체 산업을 견인할 수 있는 잠재력을 가진 것으로 평가

 기존 범용 프로세서(CPU)는 딥러닝의 높은 정확성은 대량의 데이터 처리·연산에 한계가 존재, 이를 극복하기 위해 인공지능 분야에 특화된 다양한 ‘연산유닛 (Processing Unit)’이 등장

- (GPU, Graphic Processing Unit) 당초 고화질 그래픽 연산을 처리하기 위해 개발되었으나, 딥러닝의 반복적인 연산을 대부분 고도화된 병렬 연산유닛에 그대로 대응시킬 수 있다는 것에 착안, 인공지능 연산 가속의 핵심 반도체로 역할을 수행

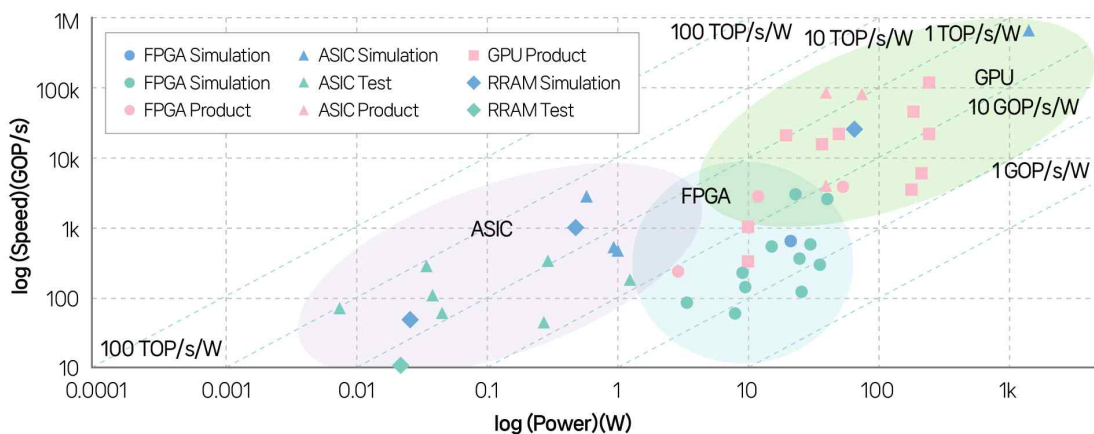
※ (GPGPU) ‘GPU의 범용연산 기능’으로 기존 CPU 프로세서의 역할을 GPU의 병렬연산을 통해 처리하기 위한 HW 및 API 기술. GPU를 통한 딥러닝 기술 구현을 위해 필수적인 기술로 대표적으로 NVIDIA社 전용 CUDA와 범용적으로 개발할 수 있는 OpenCL 등 존재

- (FPGA, Field-Programmable Gate Array) 목적에 따라 재프로그래밍이 가능한 반도체 구조\*로 짧은 개발시간과 높은 유연성에 따라 인공지능의 잦은 알고리즘 변화에 효과적으로 대응 가능

\* 전통적인 FPGA는 공급업체를 통해 로직 설계가 이루어지지 않은 상태로 납품되며, 프로그래밍 SW를 함께 제공함으로써 수요기업은 맞춤형 소자를 설계

※ 그간 타 구조에 비해 고성능·저전력 구현이 가능한 장점에 비해 높은 가격으로 범용적인 활용은 어려워 항공, 우주, 방산 등 특수 영역이나 전자 제품 생산 전 테스트용으로 사용

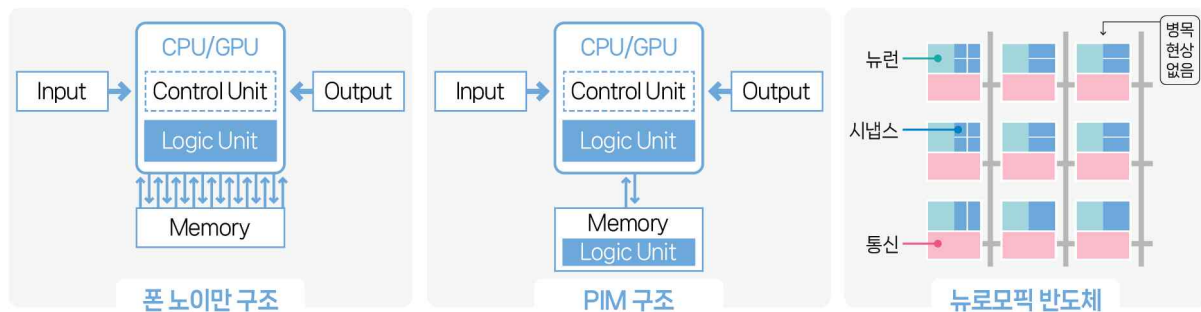
- (ASIC, Application Specific Integrated Circuit) 명확한 애플리케이션과 목적을 가진 시스템을 저전력으로 구동하기 위해 활용하는 주문형 SoC로, 낮은 범용성에도 불구하고 주요 빅테크·스타트업 기업들이 자사 제품·서비스에 특화된 인공지능 반도체를 개발·생산하기에 유리



[그림 2] GPU, FPGA, ASIC 기반 인공지능 반도체의 성능 비교

\* 출처: IEEE Signal Processing Magazine, 2019 재구성

- 또한, 기존 폰노이만 구조(GPU, FPGA, ASIC 등)에서 벗어나 인공지능 연산에 최적화된 ▲메모리 중심 컴퓨팅(Memory-Centric Computing) ▲뉴로모픽 반도체 등 구현을 위한 활발한 연구 진행 중
  - ※ 전통적인 폰노이만 컴퓨팅 구조는 연산(CPU)과 데이터 저장(메모리)을 담당하는 반도체를 직렬로 연결하여 각각의 역할을 독립적으로 수행하는 형태로, CPU와 메모리 성능 개선에도 불구하고 데이터 전송 성능의 한계로 인공지능 구현 등에 근본적인 한계가 존재
- (메모리 중심 컴퓨팅) 고성능 인공지능 구현을 위한 대용량 데이터의 고속 처리를 위해 메모리 소자 내에서 직접 연산을 수행하는 PIM 반도체 등 차세대 컴퓨팅 방식이 주목
- (뉴로모픽 반도체) 인간의 신경망 구조(뉴런-시냅스)를 모사하여 개별 칩을 병렬로 연결하여 연산·저장을 동시에 수행하며 초저전력·고성능 인공지능 연산의 구현이 가능



[그림 3] 기존 폰노이만 구조와 PIM 구조 및 뉴로모픽 반도체 구조 비교

\* 출처: 삼성전자, 각종 매체 등을 바탕으로 저자 재구성

## 인공지능 반도체는 인공지능 SW·알고리즘의 연산(학습·추론) 기능의 효율적인 연산에 특화된 반도체 소자

- 인공지능 연산은 ‘다중 병렬연산’에 기반한 ‘학습’과 ‘추론’으로 구분되며, 학습용 데이터를 통해 도출된 학습 모델은 추론 연산에 적용, 그 정확도는 도출된 학습 모델에 좌우
  - ※ 학습 모델 개발을 위한 학습용 지능형 반도체의 주요 성능 사양이 연산처리 속도와 소비전력이라면, 추론용 지능형 반도체는 정확도와 연산 레이턴시가 주요 성능 지표
- (학습용 반도체) 인공지능 구현을 위한 ‘학습’을 위해 여러 복잡한 비선형 부동소수점 연산이 필요하므로 고속 실행을 위해 범용 SW 지원이 가능한 서버용 GPU 반도체를 활용 중
- (추론용 반도체) 빠른 ‘추론’을 위해 ASIC 기반으로 상용화된 NPU나 연구·개발 중인 PIM 반도체를 바탕으로 특수행렬의 덧셈·곱셈과 같은 상대적으로 단순한 연산을 고속 실행
  - ※ 저전력·소형화 등 기존 GPU기반 컴퓨팅 플랫폼 적용이 어려운 분야와, 빠른 실행이 요구되는 분야에 주로 활용되며, 대표적으로 ▲모바일 AP내의 IP, ▲엣지 디바이스용 NPU칩, ▲대용량 인공지능 서버용 어레이 등 다양한 형태의 반도체가 개발·상용화

- (학습·추론 분리) 최근 인공지능 시스템 구조는 GPU 기반 고성능 서버(슈퍼컴퓨터)에서 대규모 데이터 학습 후 최적화·변환을 통해 엣지단(NPU 등)에서의 추론 연산이 일반화

※ NPU 기업은 학습을 위해 수십억~수백억원의 학습 서버를 구축·운영 중이거나, NVIDIA社 GPU 제품 기반의 클라우드 기반의 서버를 임대하여 활용


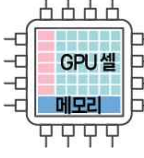

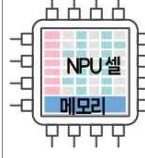
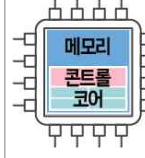
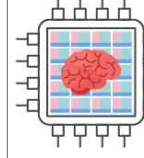
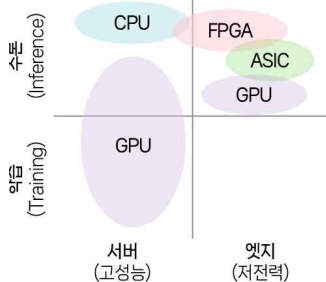
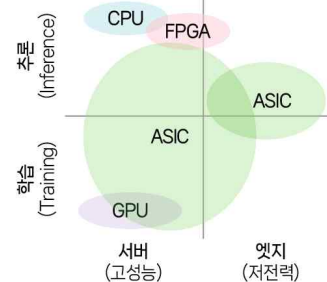
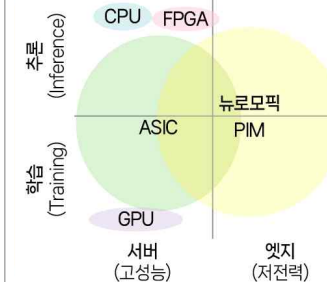
〈표 1〉 기존(1~2세대) 연산유닛 간 학습·추론 분야에 성능 비교

구분	학습		추론		범용성	정확도
	효율	속도	효율	속도		
CPU	기준 (1×)				매우 높음	~98-99.7%
GPU	~10-100×	~10-1,000×	~1-10×	~1-100×	높음	~98-99.7%
FPGA	-	-	~1-100×	~1-100×	보통	~95-99%
ASIC	~100-1,000×	~10-1,000×	~10-1,000×	~10-1,000×	낮음	~90-98%

\* 출처: CSET, 2020.

- 한편, 활발히 연구 중인 뉴로모픽 반도체는 아니라 폰노이만 컴퓨팅 구조의 한계를 극복할 수 있는 병렬형 컴퓨팅 구조를 위한 반도체 기술로 주목
  - CPU 중심의 폰-노이만 컴퓨팅 구조는 성능·미세공정·소모전력·소형화 등 기술의 급격한 발전에도 불구하고, 구조적 한계에 봉착하며 대규모 수요에 따른 산업혁신에는 역부족인 실정
  - 기존 메모리 병목·고성능·대용량 메모리 인터페이스 등 문제를 극복할 것으로 전망되나, 다층 인공신경망의 구조, 생체 학습 방법 등에서의 연구가 필요
- 이처럼 인공지능 반도체는 학습·추론을 저전력·고속으로 처리하는 데 특화되어 있으며, 연산유닛별로 아래와 같은 구분이 가능
  - (1세대 인공지능 반도체) CPU, GPU 등 기존의 반도체를 활용하여 인공지능 연산을 처리하여 비용이 저렴하고 범용성이 높지만, 상대적으로 연산성능이 낮고 소비전력은 높은 편
  - (2세대 인공지능 반도체) 특정 인공지능 연산에 특화·최적화한 FPGA, ASIC 등을 포함하며, CPU·GPU 대비 연산성능이 높고 소비전력이 낮지만, 높은 가격과 디자인된 알고리즘만 동작시킬 수 있다는 점에서 범용성은 다소 낮음
    - ※ CPU를 기준으로 GPU 및 ASIC의 학습 데이터 처리 능력은 1,000배, 추론의 속도는 ASIC이 100배, 정확도는 CPU > GPU > FPGA > ASIC순으로 알려짐
  - (3세대 인공지능 반도체) 기존 폰노이만 방식을 탈피한 PIM, 뉴로모픽 반도체 등이 대표적이며, 성능이 가장 우수할 것으로 보이나, 아직 기술성숙도가 낮고 특수한 구조상 범용성이 낮음
    - ※ 본 고에서는 PIM 반도체가 차세대 인공지능·컴퓨팅 시스템 구현을 위한 요소기술이라는 특수성을 고려하여 ‘3세대’ 인공지능 반도체로 분류



종류	1세대		2세대		3세대	
	CPU	GPU	FPGA	ASIC	PIM	뉴로모픽
특징						
	복잡 계산 순차처리	단순 계산 병렬처리	목적별 HW 재구성 가능	저전력·고효율 용도 맞춤형	메모리 내 연산 가능	뉴런·시냅스 모방 新구조
전망						
	수론 (Inference) 역산 (Training)		수론 (Inference) 역산 (Training)		수론 (Inference) 역산 (Training)	
	서버 (고성능)		서버 (고성능)		서버 (고성능)	
	엣지 (저전력)		엣지 (저전력)		엣지 (저전력)	

[그림 4] 인공지능 반도체의 구분과 특징·전망


\* 출처: 과학기술정보통신부(2021), 오윤제(2020) 등 재구성

본 고에서는 인공지능 반도체 분야 국내외 기술, 산업, 정책 및 정부 R&D 투자 동향을 살펴보고, 이를 통해 시사점을 도출하고자 함

- (기술동향) ①ASIC의 약진, ②인공지능 반도체 성능향상을 위한 주변 기술, ③오픈소스 설계SW인 RISC-V의 부상 등 산업 내 주요 기술 분야 이슈에 대해 논의
- (산업동향) ①GPU, ②FPGA, ③ASIC, ④차세대 인공지능 반도체(뉴로모픽·PIM반도체) 등 1~3세대 연산유닛 구분에 따라 국내외 주요 업체·기관 동향을 살펴봄
- (정책동향) ①인공지능·반도체를 둘러싼 미-중 간 기술패권 경쟁, ②주요국의 반도체 산업 및 인공지능 반도체 분야 지원 정책, ③국내 인공지능 반도체 지원 정책 등을 포함
- (투자동향) 그간 우리 정부가 지원한 인공지능 반도체 분야 ①R&D과제 및 ②주요 사업 이력을 바탕으로 투자 동향을 심층 분석

본 고는 KISTEP기술동향브리프 2019-01호 「인공지능(반도체)」 이후 인공지능 반도체 분야 최신기술·산업 동향·이슈를 포함하였으며, 작성 의도에 따라 분류 기준과 기술記述이 일부 상이할 수 있음

## 제2장 기술동향

 (응용·활용) 인공지능 반도체는 활용 목적에 따라 ▲데이터센터(클라우드·서버 등) ▲엣지 컴퓨팅(모바일·자율주행 등)용으로 분화·발전 중이며, 소분야에서 ASIC\*의 약진이 전망

\* 향후 인공지능 모델의 완성도가 갖춰지면 단말에서의 저전력 구동이 관건으로, GPU·FPGA 대비 성능·전력효율 향상에 유리한 ASIC이 상용화 인공지능 반도체에 궁극적인 형태가 될 것으로 전망

- 초기 엣지컴퓨팅을 위한 인공신경망 구현과 뉴로모픽 초기 모델을 시작으로 데이터센터용 인공지능 연산·작업 부하 증대로 서버·엣지컴퓨팅용 인공지능 반도체 연구가 지속

※ 특히, 알파고 이후('16), 서버·엣지(스마트폰, 자율주행차 등)에서 구글, 아마존, 애플, 삼성, 테슬라 등 AI 연산 기능을 위해 자사 제품에 인공지능 반도체를 적극 도입

- 클라우드 컴퓨팅을 위한 데이터센터용 인공지능 반도체는 '추론' 위주로 개발되었으나, 현재 NVIDIA 등 관련 칩 및 HW의 기술 진보로 '학습·추론'을 모두 서버에서 진행

- 데이터센터용 반도체는 CPU·GPU 조합을 기반으로 Intel과 NVIDIA가 시장을 주도 중이나, 최근 클라우드 시장과 인공지능 수요 확대로 ARM, RISC-V 기반 ASIC이 대체재로 부상

※ 일례로 아마존 웹 서비스(AWS), 마이크로소프트, IBM, 구글의 클라우드 회사에 사용되는 인공지능 가속기의 97%를 NVIDIA사가 점유

- '30년 서버 시장의 70% 이상을 ARM, RISC-V가 차지, PC 시장의 82%를 ARM\*이 점유할 것으로 예측되며, 특히 클라우드 시장의 표준이 될 것으로 전망\*\*

\* Apple사는 x86에서 ARM CPU로 전환 중이고, MS사는 ARM의 Windows 개발 중

\*\* 아마존은 클라우드 서버에 AWS Graviton 2(ARM CPU)를 채택, Intel보다 48% 가성비 우위 평가

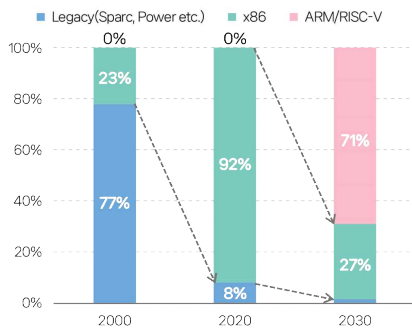
- CPU, 가속기, 메모리, 저장장치 등 데이터센터에서 성장 잠재력이 높은 반도체는 인공지능 가속기 분야로 '30년까지 연평균 21%로 성장이 전망

- 한편, 단말기에서 직접 인공지능 처리가 가능한 엣지컴퓨팅 분야에서도 기존 클라우드 컴퓨팅이 갖는 단점을 보완하며 ASIC의 주도권이 강화될 것으로 전망

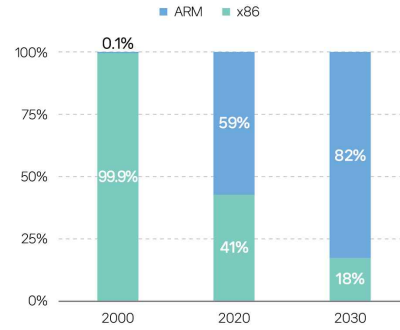
- 현재 대표적으로 적용되고 있는 분야는 모바일과 자율주행차용으로 TESLA, APPLE, Google, Qualcomm 등이 ASIC 기반의 독자적인 NPU, SoC 설계\*로 시장 주도 중

\* 팹리스와 같이 반도체 설계를 전문적으로 수행하며, 생산은 주요 파운드리에 칩단공정을 활용





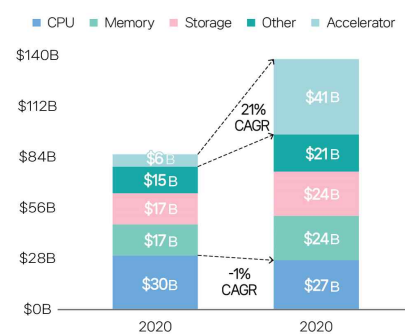
데이터센터용 인공지능 반도체 점유율 전망



PC용 연산유닛(CPU) 점유율 전망



Intel社(CPU)와 ARM社(ASIC) 기반 AWS 성능 비교



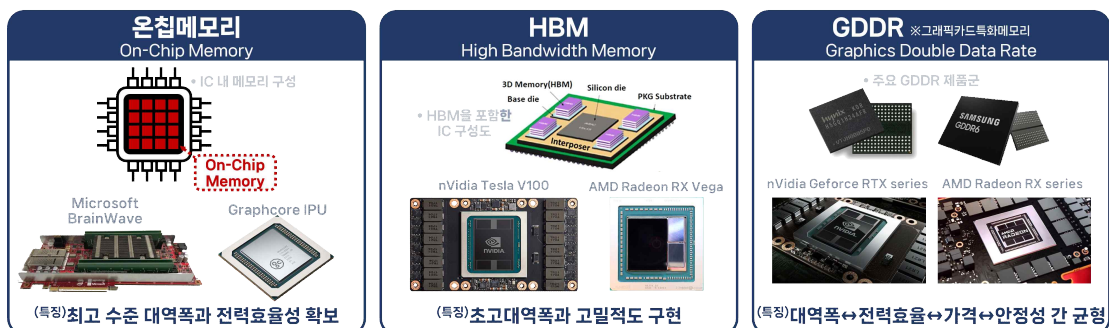
데이터센터용 인공지능 반도체 성장 전망

[그림 5] 인공지능 반도체 활용 전망

\* 출처: ARK Invest Big Idea, 2021 재구성

❖ (주변기술) 인공지능 반도체의 성능은 연산유닛 종류 외에도 ▲메모리 형태, ▲인터페이스, ▲패키징 구조, ▲단위 트랜지스터(Transistor, TR) 소자 등에 좌우

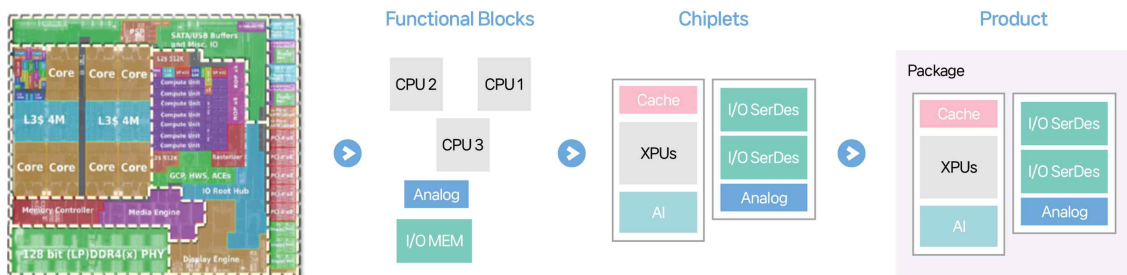
- (메모리) 인공지능 반도체의 메모리는 학습데이터 저장, 신경망의 병렬연산을 위한 핵심 요소
  - AI 연산의 빠른 처리와 저전력 연산을 위해 오프칩(Off-Chip) 형태의 DDR, GDDR 등과 연산 칩 내에 메모리를 집적한 온칩(On-Chip)으로 SRAM, MRAM 등으로 분류
  - SRAM의 성능이 가장 우수하나 가격·용량 면에서 불리하고, NAND 메모리는 처리 속도의 한계로 인공지능 반도체 온칩 메모리보다 오프칩 메모리가 주로 활용



[그림 6] 메모리 종류별 인공지능 반도체 활용 사례

\* 출처: Rambus, 2020 재구성

- (인터페이스) 연산유닛에 대한 대용량 학습데이터의 전송 속도와 대역폭이 주요한 성능 인자 중 하나로 이와 관련한 인터페이스 기술\*은 진화 중
  - \* 인터페이스는 외부 저장소(Storage)와 서버와의 연결 또는 컴퓨터 간 연결하는 Inter-Node 연결과 연산 컴퓨팅 내에 데이터 전송인 Intra-Node로 분류
  - ※ 연산능력 및 대역폭의 개별적인 2배 향상은 시스템 성능에 2배로 미치지 못하고, 두 개의 인자가 동시에 고려되어야 효과적인 시스템 성능향상이 가능
- Inter-Node 연결로는 Ethernet, Infiniband, BXI(Bull eXascale Interconnect) 등이 있고, 선도 기업인 Mellanox社의 'NDR(Next Data Rate)'이 가장 큰 대역폭(최대 400Gbit/s)으로 개발 중
- Intra-Node는 NVIDIA社의 NVlink가 최대 50GT/s로 기술을 선도하고 있으나 최근 CXL(Compute Express Link)\* 인터페이스의 활용도도 증가
  - \* CPU, GPU, 메모리 등을 효율적으로 활용하기 위한 PCIe를 대체하는 인터페이스로 메모리 용량의 확장성과 저지연성이 용이하여 삼성, SK하이닉스는 CXL을 지원하는 메모리 솔루션 출시('22)
- (패키징) 개별 연산유닛만큼 인공지능 반도체 성능 구현의 주요 인자로 최근 제조 단가와 성능 간 최적화를 위한 인공지능 반도체 구조(폼팩터)로 Chiplet(칩렛) 패키징이 대두
  - ※ 미세 공정과 다중코어 수의 증대에 따른 칩 면적은 대면적으로 확장되고 이는 제조 단가 상승의 주된 요인이므로, 이를 최적화를 위한 폼팩터가 Chiplet 구조
- Chiplet은 반도체 패키징의 한 종류로 다양한 기능을 집적한 단일칩(Monolithic) 구성이 아닌 다이(Die, 집적회로칩) 내 기능을 분리·단순화하고, 개별 다이를 인터커넥션으로 연결
  - ※ ▲고성능 다기능 칩, ▲수율 간 Trade-off, ▲확장성에 대한 요구 등으로 기존 '다기능 One Die'에서 '단(單) 기능 Multi Die 패키징'으로 진화 중
- AMD, INTEL, 삼성전자, TSMC, MS, Meta 등은 패키징 내 칩 간 고속 데이터 인터페이스, 칩 영역 외 엣지·데이터센터에서 활용가능한 개방형 인터페이스 표준인 UCle(Universal Chiplet Interconnect Express) 추진을 위한 컨소시엄 구성('22.03.)



[그림 7] 단일칩 구성과 Chiplet 구조 비교

\* 출처: Chip Scale Review, 2022 재구성

- (트랜지스터(TR) 소자) 연산유닛 성능의 근원으로 소모전력과 저지연성 등에 직접적인 영향을 주는 인자로 TSMC, 삼성전자 등 주요 파운드리 of 첨단 제조 능력에 따라 좌우
- 반도체 미세공정은 TR 소자의 게이트(Gate) 물리적 크기가 기준으로, 소자의 대역폭, 소비 전력, 주파수 응답 특성 등에 직접적인 영향

- 세계적으로 수nm 공정을 구현할 수 있는 파운드리인 TSMC(대만)과 삼성전자(대한민국) 2개社뿐으로 현재 3nm 공정 양산 및 수율 확보를 위한 경쟁 중

### (설계SW) 오픈소스인 RISC-V의 장점으로 스타트업 중심의 활용 증가와 기존 업체의 개발·투자 확대가 이루어지며, 기존 ARM 중심의 반도체 설계 분야 재편 중

- 저전력·고효율 소자 구현을 위한 미세화 공정에 따른 제조 단가 증가와 함께 설계자산(IP) 비용의 급증으로 신생 업체의 개발과 시장 진입에 걸림돌로 작용
  - 업계에 따르면 기존 28nm 공정 전주기 비용이 611억원 수준이었던 데 반해 5nm 공정 개발 비용은 6,461억원을 상회하며, 이중 절반 수준이 IP와 설계SW가 차지
    - ※ 일례로 10nm 이하 공정에 쓰이는 초고속 연결기술(인터페이스) ‘SerDes’ IP 가격은 80억원 안팎
  - 이에 따라, '10년 UC버클리에서 오픈소스로 공개한 설계SW(아키텍처)인 ‘RISC-V’가 최근 기존 IP 생태계를 대체할 대안으로써 급부상
- RISC-V가 갖는 경제성·저전력·보안성·비정치성 등 장점으로 적극적인 활용이 기대되나, 안정성·호환성 확보 등 개발위험 부담에 따라 ARM과 병·혼용 이후 점차 시장 안착 전망
  - RISC-V의 기본 ISA 수는 47개로 기존 CISC(Intel, AMD) 1,500개, RISC(ARM) 200개 대비 간소화하여 명령어 구조, 확장성, 활용 목적에 따른 유연한 적용과 저전력 소자 설계가 가능
  - NVIDIA社의 인수 무산, 퀄컴과의 소송전 등을 계기로 ARM社 설계자산 대체재에 관한 관심이 높아진 가운데 Intel社의 RISC-V 생태계 조성까지 속도를 내며 업계 이목 집중
    - ※ 그간 반도체 업계는 대체로 ARM社의 저전력 반도체 설계도와 명령어 세트에 의존하여 제품을 제작하였으나, NVIDIA의 인수 무산('22.02.) 등 과정에서 선택지 관점에서 업계 경각심 고조

〈표 2〉 주요 기업의 RISC-V 활용 사례

기업명	주요 사례
Intel	• 바로셀로나 슈퍼컴퓨터 센터와 협력으로 RISC-V 프로세서 개발에 4억 유로 투자 발표('22.06.)
Apple	• 임베디드 코어 일부를 AR에서 RISC-V 활용 추진('22.09.)
TESLA	• 데이터센터용 AI칩 D1내 데이터 흐름제어를 위해 RISC-V를 혼용
ThinkSilicon	• 세계 1위 반도체 장비사인 Applied Materials社의 자회사(그리스)로 RISC-V 기반의 GPU와 CPU를 집적한 AI 연산 가속기(28nm) 개발('21.03)
Western Digital(WD), Seagate 등	• SSD/HDD 컨트롤러를 기존 ARM 기반에서 RISC-V로 대체하며 관련 Chipset 개발 ※ (WD) SweRV코어('19.12.) (Seagate) HDD 성능을 3배 높이는 RISC-V코어 개발('21.05.)
Mobileye, Kneron 등	• RISC-V기반의 자율주행 자동차용 인공지능 반도체를 개발 ※ (Mobileye) 100W미만의 소비전력으로 42TFLOPS, 176TOPS의 성능으로 자율주행 4단계 지원하는 ‘EyeQ Ultra’를 발표하고, '23년 양산 예정('22.01) ※ (Kneron) ARM 코어(Cortex M)와 RISC V 가속기의 하이브리드 집적, 평균소비전력 500mW로 0.5TOP(INT8) 성능 발표 ('21.11)
Esperanto Technology	• RISC-V 기반 데이터 센터 추론·연산 용도(또는 추천 알고리즘용)의 저전력 AI Chip(ET-SoC-1, TSMC 7nm공정) 및 데이터센터 활용 사례 발표('21.08) ※ 최대 소비전력 120W, 최대 연산능력 800TOPS로 미국 삼성SDS 해당 칩 테스트 중('22.04)

※ 저전력의 장점으로 주로 자율주행을 비롯한 엣지용으로 개발되고 있으나 서버용으로도 확장 중

- 이와 같은 RISC-V의 특수성에 따라 미국·유럽·중국 등 지역별로 원천기술·활용성 확보를 위한 맞춤형 기술개발 추진 중
  - 미국은 스타트업과 대형 반도체 기업 등 민간을 중심으로 RISC-V 기반 반도체·주변기술 상용화를 위한 투자 확대 중
  - 유럽은 ETP4HPC\*을 필두로 MEEP, eProcessor, Red-Sea 등 R&D 프로젝트를 통해 RISC-V 관련 기술개발 진행 중

\* ETP4HPC : European Technology Platform for High Performance Computing

〈표 3〉 유럽의 RISC-V 관련 R&D 프로그램 현황

프로젝트	주요 내용	참여기관
eProcessor (’21.04~’24.03)	<ul style="list-style-type: none"> <li>• RISC-V 기반 HPC용 CPU 코어 개발</li> <li>• 우월한 전성비를 위한 HW/SW 개발</li> <li>• ML/DL로 확장 적용</li> </ul>	<ul style="list-style-type: none"> <li>• 주관기관 : BSC(스페인)</li> <li>• 참여기관 : FORTH(그리스), Sapienza大(이탈리아), Cortus(프랑스), Thales SA(프랑스), EXTOLL(독일) 등 9개 기관</li> </ul>
MEEP (’20.01~’22.12)	<ul style="list-style-type: none"> <li>• 자체 개발 IP로 엑사스케일 슈퍼컴퓨터 개발용 HW/SW 테스트 플랫폼 제공</li> <li>• RISC-V 기반 가속기 개발(ACME)</li> </ul>	<ul style="list-style-type: none"> <li>• 주관기관 : BSC(스페인)</li> <li>• 참여기관 : 자그레브대학(크로아티아) 등 2개 기관</li> </ul>
TEXTAROSSA (’21.04~’24.3)	<ul style="list-style-type: none"> <li>• 거대규모 슈퍼 컴퓨팅을 위한 기반 기술 및 가속기 개발</li> <li>• CPU 작업 스케줄링 및 저지연 통신을 위한 RISC-V 기반 IP개발</li> </ul>	<ul style="list-style-type: none"> <li>• 주관기관 : ENEA(이탈리아)</li> <li>• 참여기관 : Atos(프랑스), BSC(스페인) 등 10개 기관</li> </ul>
RED-SEA (’21.04~’24.03)	<ul style="list-style-type: none"> <li>• ARM, RISC-V, 하이엔드 CPU, FPGA, ASIC 등 이종 엑사스케일 컴퓨터 시스템 간 네트워크 표준 개발</li> </ul>	<ul style="list-style-type: none"> <li>• 주관기관 : Atos(프랑스)</li> <li>• 참여기관 : FORTH(그리스), EXTOLL(독일) 등 10개 기관</li> </ul>

\* 출처: European High-Performance Computing projects - HANDBOOK, 2021

- 한편, 중국은 미국의 제재에 대한 대안으로 적극적인 연구개발 추진 중으로 중국과학원은 RISC-V 기반 CPU\* 개발을 위해 베이징 카이신연구소(Kaixin Institute)\*\* 개소(’21.12)

※ 최초 RISC-V 국제적인 행사는 ’17년 중국에서 개최한 바 있으며, 중국 OS 독립을 위해 Kylin Ubuntu를 지원하는 RISC-V CPU 개발을 적극적으로 추진(’22.05)

\* RISC-V 기반 CPU인 장산(Xianshan)은 SMIC 14nm, 동작 클럭 2GHz 추정, Linux에서 동작 성공

\*\* 카이신 연구소의 역할 중 일부는 중국 내 RISC-V 에코 시스템 구축

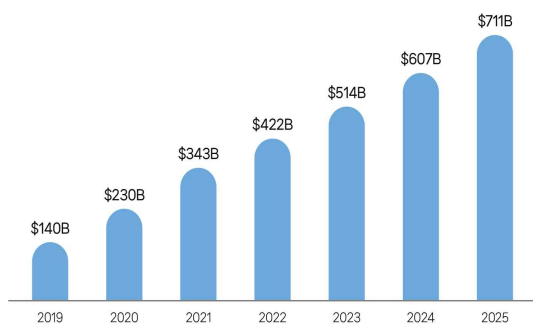
## 제3장 산업동향

향후 연평균 30%를 상회하는 폭발적인 시장 성장으로 반도체 산업을 견인, 응용 분야에서는 데이터센터 대비 엣지컴퓨팅 분야의 확대가 전망

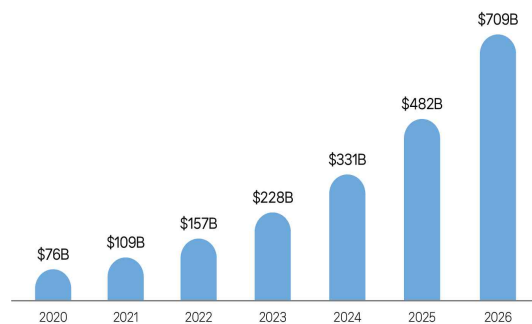
- 다수 유력 매체\*에 따르면 인공지능 반도체 시장 규모는 가파른 증가가 전망되며, 미래 주요 산업으로 성장할 것으로 전망(매체에 따라 현수준·예측치는 다소 상이하나, 대체로 큰 폭의 성장을 전망)

\* (Gartner) '20년 230억 달러 규모에서 오는 '25년 700억 달러까지 성장('21), '30년 1,179억 달러까지 성장하여 전체 시스템반도체 시장에서 약 31%를 차지할 것으로 전망('19)  
(Statista) '21년 109억 달러에서 '26년까지 약 709억 달러 규모를 달성(연평균 45.4% 성장)('22)  
(ReportLinker) '21년 96.4억 달러에서 '27년까지 약 664억 달러로 연평균 35.2%의 성장 전망('22)  
(OMDIA) 인공지능 가속기 시장규모는 '20년 153억 달러에서 '24년 428억 달러 규모로 기존 프로세서 대비 약 5배 높은 연평균 29.2% 성장률을 기록할 것으로 전망('20)

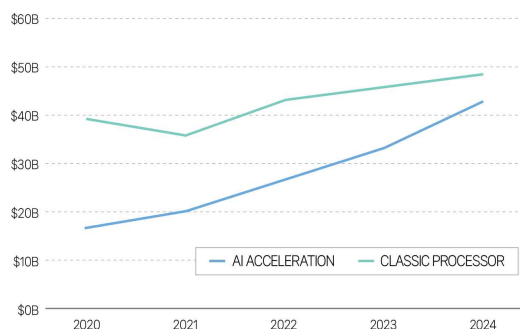
- 한편, 딥러닝 연산을 위한 인공지능 반도체 비중으로는 ASIC의 가파른 성장세, CPU는 답보, GPU의 폭발적인 성장세는 '22~'23년 이후 다소 둔화할 것으로 전망(Tractica, '19)



Gartner, '21.



Statista, '22.



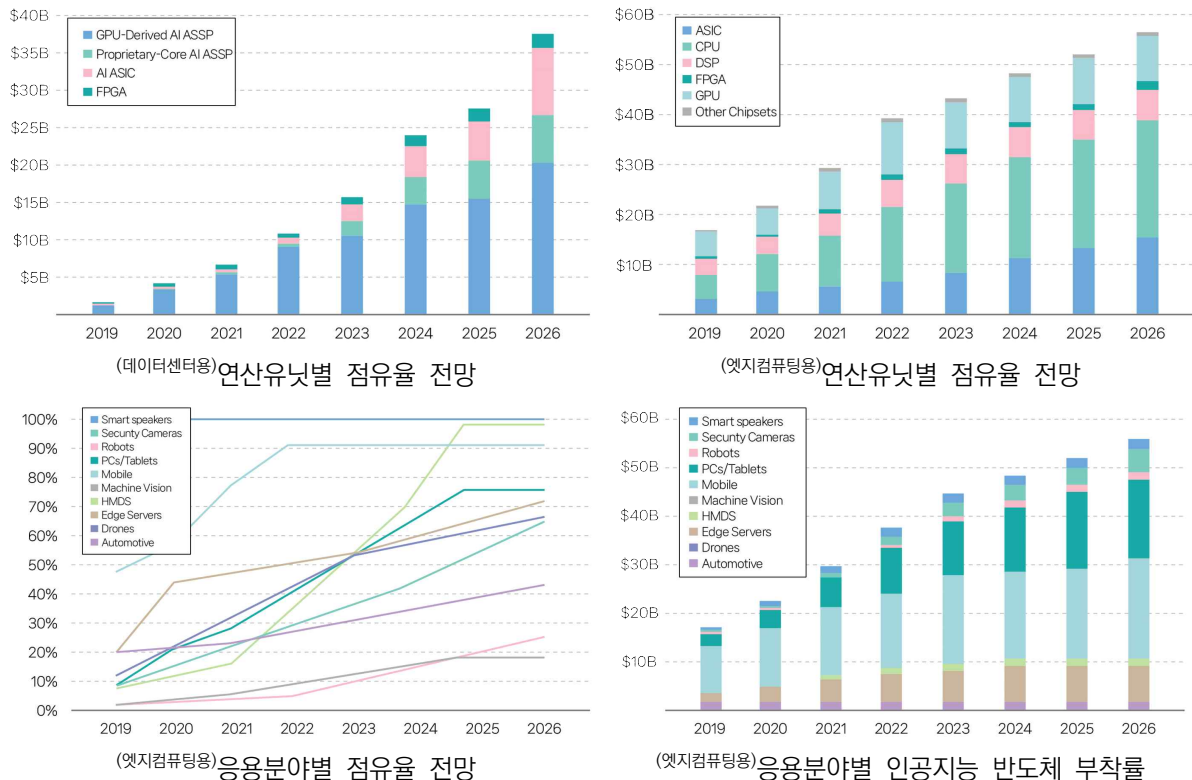
OMDIA, '21.



Tractica, '19.

[그림 8] 유력 매체의 인공지능 반도체 시장 전망 (각 통계 출처를 바탕으로 재구성)

- 인공지능 반도체의 부상은 시스템반도체 산업에서 새로운 기회를 창출하며, 인공지능 활용 제품·서비스의 성장에 따라 관련 시장이 성숙 중
  - 데이터센터(클라우드컴퓨팅) 분야는 시장 규모 확대\*와 함께 GPU의 강세\*\*가 이어질 것으로 전망되며, '23년 이후로 ASIC의 비중도 점차 확대할 것으로 전망
    - \* ('21) 6.9억불 → ('26) 37.6억불로 약 5배 이상 성장을 전망(Omdia, 2021)
    - \*\* '26년경에는 데이터 센터와 관련된 반도체 비용 중 절반 이상이 GPU 중심의 가속기들로 구성되며, 이중 NVIDIA社 제품이 약 80%로 절대적인 시장 점유율을 차지할 것으로 전망
  - 엣지컴퓨팅 분야 역시 지속적인 시장 규모 확대가 전망되며, 전력 소모 등 GPU의 한계\*로 '22~'23년 매출 정점에 도달한 후 ASIC·FPGA 등으로 대체될 것으로 보임
    - \* 2배 높은 성능 구현을 위해, 처리능력, 학습 시간 등 10배에 가까운 컴퓨팅 능력을 사용하고 있으며, GPU를 적용한 최신 AI/HPC 시스템은 이미 전력 소모가 매우 높은 상황
  - 엣지컴퓨팅 분야 응용·활용에 있어 모바일 분야가 가장 앞서 있으며, PC/Tablet 분야 역시 '19년 대비 '26년 약 8배 이상 가파른 성장세가 전망됨
    - ※ 한편, 응용분야 별로 부착률(attach rate) 면에서는 모바일을 비롯한 스마트 스피커, HMD가 90% 이상, PC/Tablet에서는 70%를 웃돌며 대부분의 단말에서 엣지컴퓨팅을 지원할 것으로 전망



[그림 9] 데이터센터·엣지컴퓨팅용 인공지능 반도체 시장 전망

\* 출처: Omdia, 2020 재구성



### 3.1 해외 산업동향

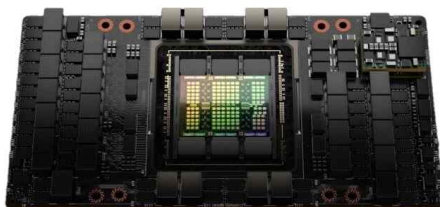
■ (GPU) 현재 데이터센터(서버) 기반 인공지능 상용화 서비스를 위한 연산 분야의 뛰어난 활용성으로 절대적인 시장 지배력을 갖는 인공지능 반도체(연산유닛)

- (NVIDIA) '10년 이후 딥러닝 기반 인공지능 기술은 사실상 NVIDIA GPU·SW에 의존하여 급격한 발전이 이루어졌으며, 현재 대부분의 데이터센터 등 상용화 서비스에 사용 중
  - ※ 인공지능 분야 NVIDIA GPU 제품의 경쟁력은 ▲CUDA 등 SW 기술(GPGPU)과 ▲CPU-GPU 컴퓨팅 구조, ▲최신 미세 공정과 최신 패키지(3D 적층) 기술이 접목된 제조 역량, ▲20년 동안 발전해온 다양한 AI SW 인프라와의 호환성으로, 현 시장에서 대체가 어려운 수준의 의존도를 가짐
- 전작 A100 대비 최대 30배 성능(Streaming Multiprocessor, Transformer Engine, NVlink, MIG, HBM3 등)의 데이터센터(학습·추론)용 반도체 GPU H100을 발표('22.03.)
  - ※ 학습용 AI반도체인 A100 GPU 모델은 6,912개의 Cuda Core, 40GB HBM2 메모리를 가지고 있으며, 이를 사용하여 구축된 NVIDIA DGX A100 서버는 640GB 메모리에서 5 PFLOPS, 10 POPS의 성능으로 슈퍼컴퓨터 수준의 대규모 서버로 확장이 가능

〈표 4〉 NVIDIA H100 주요 성능

주요 성능	특징
기본 사양	<ul style="list-style-type: none"> <li>• TSMC 4nm 공정으로 제작</li> <li>• 120TFLOPS (FP16)의 성능, 700W 소모전력</li> </ul>
Streaming Multiprocessor (SM)	<ul style="list-style-type: none"> <li>• 텐서코어(FP8, 16, 64, BF16, TF32, INT8 지원), Int 32, FP 32, 64 코어, 외부 메모리(HBM 등)와 Cuda 메모리간 데이터 전송을 가속하는 TMA(Tensor Memory Accelerator), 동적할당 AI 알고리즘 특화 연산 기능인 DPX(Dynamic Programming Instruction) 등이 탑재</li> </ul>
Transformer Engine	<ul style="list-style-type: none"> <li>• 자연어 학습 알고리즘(단어와 언어 어순 정보의 인코딩 및 디코딩 등 수행) 연산 가속기</li> </ul>
4세대 NVlink	<ul style="list-style-type: none"> <li>• GPU to GPU 인터페이스로 900GB 대역폭 제공(PCIe Gen5의 7배)</li> </ul>
2세대 MIG (Multi Instance GPU)	<ul style="list-style-type: none"> <li>• 1개의 GPU로 7개의 GPU Instance의 독립 연산을 지원</li> </ul>

- 한편, 자율주행을 비롯한 엣지컴퓨팅용 Jetson 시리즈의 제품을 출시('19~)하였으며, 중국 3대 전기차 기업인 Li Auto社 신차의 자율주행 기술에 Jetson AGX Orin을 적용('23.02.)
  - ※ '22년 출시한 Jetson AGX Xavier 모듈은 소모전력 10W에서 32 TOPS의 AI 가속 실행이 가능하며, NVIDIA CUDA의 JetPack SDK을 지원



(데이터센터용) H100 Tensor Core GPU



(엣지컴퓨팅용) Jetson AGX Xavier™

[그림 10] NVIDIA의 인공지능용 GPU 제품군

- (AMD) 고유의 GPU 구조인 CDNA2를 탑재한 듀얼 칩렛(Chiplet) 기반 AI 가속기 INSTINCT MI250를 발표('21.11.)
  - 내부 메모리 구조 HBM2e를 채택, 3.2TB/s 구현 및 최대 383 TFLOPS/500~600W의 성능으로 CPU-AI 가속기 또는 GPU-GPU 간 Infinity fabric Link로 데이터 전송 최대 800GB/s 구현
  - 한편, 경쟁사인 NVIDIA社 CUDA와 동일한 GPGPU 기능을 제공하는 'ROCm'을 구축, 시장 경쟁력 유지에 대응

〈표 5〉 CUDA와 ROCm의 주요 기능 비교

CUDA(NVIDIA)	ROCm(AMD)	주요 기능
cuBLAS	rocBLAS	• Basic Linear Algebra Subroutines
cuFFT	rocFFT	• Fast Fourier Transform Library
cuSPARSE	rocSPARSE	• Sparse BLAS+SPMV(Sparse matrix-vector multiplication)
cuSOLVER	rocSOLVER	• Lapack Library
Thrust	rocThrust	• C++ parallel algorithms library
cuDNN	MIOpen	• Deep Learning solver library

\* 출처: Nextplatform.com

- (INTEL) 독자적인 공정 기술을 적용한 GPU와 AI 가속기를 내재한 데이터센터용 CPU를 개발 중이며, Habana Labs社 인수('19)로 학습 및 추론용 AISC 기술을 확보
  - 시장의 GPU 수요 대응을 위해 128개의 코어(Xe core), HBM2e/캐쉬 메모리 등 이중 패키징 기술(EMIB\*, Foveros\*\*)을 적용한 GPU인 Ponte Vecchio(45TFLOPS(FP32기준)) 발표
    - \* EMIB : 다이와 다이 간 연결을 위한 인터포저(Interposer)를 대체하는 인텔고유의 이중 (Heterogeneous) 다이(Die)간 임베디드 형태의 국소공간 배선 기술. 인터포저 대비 TSV 및 추가 금속 배선 층 없는 공정으로 칩렛(Chiplet) 구현을 위한 인터커넥트 기술
    - \*\* Foveros : EMIB가 다이 간 확장의 2차원 연결인 반면 3D 적층 연결을 위한 액티브(Active) 인터포저를 적용한 패키징 기술
  - DDR5, HBM, 옵테인 메모리, 학습 및 추론을 위한 AI 가속기\* 등 탑재한 데이터 센터용 CPU Sapphire Rapids를 발표('21.08)하였으며, '23년 양산·출시 전망
    - \* NVIDIA의 행렬연산 코어 텐서코어와 동일한 AMX(Advanced Matrix Extension), 데이터 전송을 가속화 하는 DSA(Data Streaming Accelerator) 탑재 등, 데이터 센터 시장에서 시장 점유율 하락에 대한 대응
  - 한편, 기존 CPU·GPU 외 ASIC으로는 Habana Labs社 인수 후 2세대 학습 및 추론용 가속기 GAUDI2\*와 Goya의 후속 제품인 GRECO를 발표('22.05)
    - \* 1세대 대비 공정노드는 16nm → 7nm, 연산유닛코어 8개 → 24개로 증가 등 NVIDIA A100대비 2배 성능 보고하였으며, 5nm 공정노드와 메모리 용량을 개선한 3세대 제품 출시 예정('23)

## (FPGA) 시장을 주도하는 선도업체가 기존 반도체 업체에 인수되며, FPGA를 통합한 새로운 구조의 반도체 개발 중

- 업계 1위인 Xilinx社は FPGA의 시장 주도 기업으로 Edge에서 클라우드까지 적용할 수 있는 FGPA 제품군인 Versal 출시 중으로 최근 AMD社에 인수(490억달러, '22.2.)
  - ※ AMD社は Xilinx社 인수 이후 딥러닝에 특화된 데이터센터용 FPGA인 VCK 5000을 발표('22)
- ARM의 Cortex A(고성능 연산) 및 R(실시간·보안성 처리 특화)의 듀얼코어 CPU와 별도의 AI 엔진\*을 FPGA를 집적
  - \* 행렬연산(2차원 텐서) 아닌 벡터 연산유닛(1차원 텐서) 구조로 AI 연산외 DSP 등 다양한 혼합 연산에 대응
  - ※ 엣지용은 10W미만의 소비전력으로 14TOPS, CPU 연산을 보조하는 AI 가속기는 479Tops/75W의 성능으로 NVIDIA의 제품(Jetson AGX Xavier)과 비교 최대 4.7배 성능
- 한편, Intel社は FPGA 2위 기업인 Altera社 인수('15) 후 Agilex™ FPGA 제품군을 출시('19)하고, 인공지능 특화 기술을 확보하기 위해 독립형 FPGA와 통합형 FPGA(FPGA-CPU) 제공
- 다만, 다양한 인공지능 연산에 대응이 가능한 장점에도 불구하고, 성능의 한계로 대용량 학습 모델개발보다 ASIC으로 전환 전 초기 모델개발이나 추론용으로 시장 점유 전망

## (ASIC) 목표 애플리케이션과 대상 플랫폼이 명확해지면서 인공지능 알고리즘 연산에 최적화된 ASIC 반도체(NPU, TPU 등) 개발에 박차

- ASIC은 특정한 용도에 맞추어 개발된 반도체 회로이며, FPGA나 GPU 대비 동작 속도를 높이거나, 전력효율을 높이고, 칩의 면적을 줄이기가 용이
- 인공지능 모델의 완성도가 갖춰진 경우엔 실제 기기에 적용하기 위해 전력 소모 감소에 초점이 맞춰지기 때문에 ASIC은 인공지능 반도체의 궁극적인 형태가 될 것으로 예상
- 대표적으로 Tesla와 Google 등 미국 기업을 중심으로 수년간 ASIC 형태의 인공지능 반도체를 개발하고 있으며, 영미권뿐만 아니라 EU와 중국에서도 이를 개발하기 위한 지속적인 투자가 이뤄지고 있음
- (TESLA) TESLA는 자율주행 차량의 인공지능 연산을 처리하는 FSD(Full Self Driving) 칩과 NVIDIA와 협력을 통해 다수의 D1칩을 패키징한 데이터 센터용 칩셋을 개발
- 자율주행 연산\*을 위해 '18년 12개의 ARM Cortex-A72@2.2Ghz와 600 GFLOPS GPU, 73 TOPS(int8) NPU를 포함한 260mm<sup>2</sup> 의 FSD 칩을 개발
  - \* FSD 컴퓨터를 통해 8개의 카메라, 레이더, GPS, 초음파센서, 휠 눈금, 스티어링 각도, 맵데이터가 동시에 입력되며, 소비전력은 72W 수준

※ 자율주행을 차량용 보드는 FSD 2개의 칩, 소비전력 100W로 144TOPS(int8) 구현, 1개의 FSD칩 소비전력은 36W 수준

- 한편, 자율주행 완성도를 높이기 위해 최상의 인공지능 학습 성능을 보유한 슈퍼컴퓨터 Dojo 시스템\*에서 대용량 정보 처리를 담당할 전용 반도체 칩(NPU) ‘D1’ 발표(‘21.08.)

\* D1칩 3,000개를 조합해 초당  $10^{18}$ 번 연산이 가능한 1.1엑사플롭스(exaFLOP)급 성능을 제공하며, 기존 엔비디아 제품 대비 20% 작은 크기에 1.3배 전력효율, 4배의 성능을 발표

※ D1은 TSMC 7nm 공정을 기반으로 제작하며, 연산 성능은 362 TFLOPS 수준

- 이처럼 자체 개발한 NPU인 D1, 슈퍼컴퓨터 시스템(Dojo)과 테슬라봇(Tesla Bot)\* 등 다양한 차세대 인공지능 기술을 공개(‘21.08.)

\* 코드명 ‘옵티머스’로 명명된 로봇에는 테슬라의 자율주행차 핵심 기술인 ‘오토파일럿’과 인공지능을 고도화하는 슈퍼컴퓨터 ‘Dojo’ 등 최첨단 기술을 적용할 전망



[그림 11] TESLA社 NPU 및 관련 기술 (왼쪽부터 FSD칩, FSD컴퓨터, 테슬라봇)

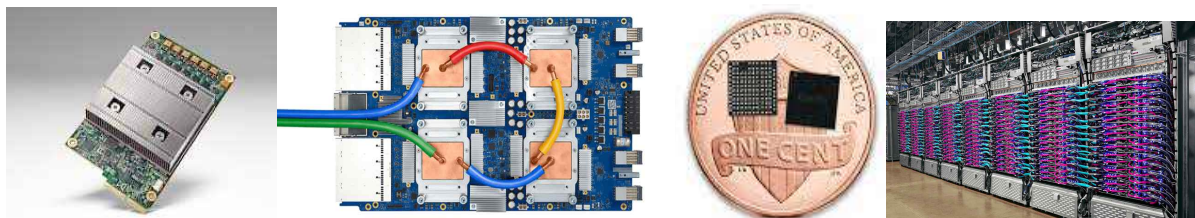
- (Google) '16년 알파고 연산에 사용된 TPU v1\*를 시작으로 '21년 7nm 공정을 기반으로 개발된 TPUv4는 138 TOPS의 성능 공개

\* TPU(Tensor Processing Unit)는 자사 오픈소스 딥러닝 라이브러리인 TensorFlow에 특화된 ASIC으로써 2세대 제품부터는 클라우드 기반 서비스를 제공('18~)

- '20년 4,096개의 TPUv3 반도체와 수백개의 CPU를 연결하여 총 430 PFLOPS의 성능 (100 PFLOPS/TPU-Pod) 을 보임('20), TPUv4는 TPUv3의 2.7배의 성능

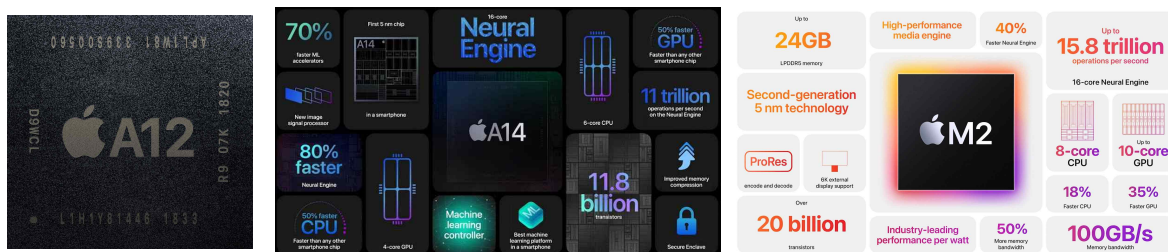
- '22년 Google은 TPUv4를 기반으로 대화형 서비스, 다국어 번역과 같은 자연어 처리를 위한 거대인공신경망 처리 컴퓨터(엑사스케일, 3000억개 이상의 파라미터수)를 구축

※ 한편, '19년 소형 장치에서 추론이 가능한 Edge TPU 칩셋을 발표하고, 칩셋이 장착된 Coral Dev. Board와 USB Accelerator를 보드를 출시하였고, 음성 및 영상 인지에 활용이 가능



[그림 12] Google社 TPU (왼쪽부터 TPUv2, TPUv3, Edge-TPU, TPU기반 AI 데이터 센터)

- (APPLE) '18년부터 자체 설계한 AP인 A12 Bionic부터 영상처리, 음성/얼굴 인식, 게임 등의 다양한 응용을 위해 NPU IP 탑재하여 자사 모바일·컴퓨터 제품군에 적용
  - \* Apple은 Intel, NVIDIA 등 기존 CPU·GPU 칩메이커로부터 독립하여 자체적인 모바일·컴퓨팅 SW 생태계(iOS, macOS 등)에 최적화된 자체 SoC로의 전환(transition) 계획을 발표('20)
- 애플의 NPU는 다층 인공신경망의 추론이 단말에서도 동작하도록 CPU에 비해 MAC 연산 작업의 효율이 50배 높은 가속기이며, 2020년 이후 랩탑용 M1/2 프로세서에도 탑재
- '22년 발표한 M2칩은 TSMC 2세대 5nm 공정에서 제조되었으며, 16개의 NPU코어를 탑재하고 15.8 TOPS의 성능을 구현



[그림 13] Apple社 NPU IP (왼쪽부터 A12, A14, M2)

- (Qualcomm) 기존 시장을 선도하고 있는 모바일용 AP(Snapdragon)의 인공지능 처리 성능을 대폭 향상, 최근 AR용 SoC(Snapdragon AR2 1세대)를 발표('22.11.)하는 등 영역 확장 중
  - 자사 최신 칩셋인 'Snapdragon8 2gen'를 통해 모바일AP 최초로 'INT4(32비트)\*'를 지원하며, 전 세대 대비 4.35배의 인공지능 성능과 추론 연산에서 와트당 60% 성능향상을 발표('22.11.)
  - ※ 퀄컴은 '19년 모바일 AP의 인공지능 가속을 위해 'Snapdragon 855' 내 DSP인 '퀄컴 헥사곤 프로세서(QDSP6)'에 인공지능 프로세서 런타임(AIP runtime) 소프트웨어 스택을 도입하여 DSP 최대 사용 시 모든 범용코어 활용 대비 3배 이상인 7TOPS 수준 연산력을 제공
- (ARM) 엣지에서 클라우드 데이터센터까지 활용 가능한 코어 IP와 SW 솔루션\*을 구축
  - \* ARM 설계에 최적화된 SW 라이브러리 ARM-NN(추론용), ARM Computer Library(영상처리 및 학습용), CMSIS - NN(Cortex M Processor의 학습용 라이브러리)을 구축·제공
- Cortex CPU, Ethos NPU, Mali GPU에 AI 가속기를 탑재 시장에 대응

〈표 6〉 ARM社 주요 IP 종류와 특징

종류	주요 특징
Cortex	<ul style="list-style-type: none"> <li>• A 시리즈는 모바일AP 또는 데이터 센터에서 CPU와 함께 실시간 AI 연산 등에 활용               <ul style="list-style-type: none"> <li>※ 8개 코어를 하나의 클러스터로 묶어서 사용하는 ARM 고유의 DynamIQ 기술로 AI 연산에 대응</li> </ul> </li> <li>• M 시리즈는 IoT와 Edge에서 이상 신호, 객체 검출 등 저용량 AI 연산용</li> </ul>
Ethos NPU	• Edge, 클라우드, 자율주행차 등에서 추론용 연산 반도체 (10~640TOPS)
Mali GPU	• 클라우드 등에서 대용량 학습용 AI 반도체



- (IBM) 인공지능 기술의 금융·의료 분야 특화 제품을 위한 인공지능 추론 전용 가속 칩 ‘Telum’과 이를 활용해 양자 내성 암호 시스템을 탑재한 IBM z16\*\* 메인프레임(‘22.04.)과 z16 기반 AI 연산에 특화된 AIU(Artificial Intelligence Unit) 연이어 발표(‘22.10.)

\* 거래 처리(금융·매매·보험 등) 중에 AI 추론 기술을 적용할 수 있도록 온칩 가속 기술을 포함한 프로세서로 삼성전자 파운드리 7nm EUV 공정으로 제조

\*\* IBM 텔럼프로세서 기반 방식으로 AI 추론을 결합해 안전한 대용량 거래 처리를 지원. 1밀리초(ms)의 지연 시간으로 하루에 3,000억 개 추론 처리 가능

- 미국을 필두로 유럽 등 다수의 스타트업에서 ASIC 기반 상용 인공지능 반도체 개발에 박차

〈표 7〉 미국·유럽 스타트업의 ASIC 기반 인공지능 반도체 개발 현황

업체명(국가)	개발 내용
<b>Kneron</b> (미국)	<ul style="list-style-type: none"> <li>• '19년에 자사에서 개발한 NPU를 포함한 엡지향 칩셋인 KL520 발표 후</li> <li>• 2021년에는 트랜스포머 네트워크와 4-bit INT를 추가로 지원하는 KL530을 발표하였으며, 8-bit INT 기준으로 이전 세대보다 2배의 성능(0.5 TOPS)을 보이고, 4-bit INT 기준으로 1 TOPS의 성능을 500mW에서 보임</li> </ul>
<b>Cerebras Systems</b> (미국)	<ul style="list-style-type: none"> <li>• '19년 8월에 세계 최대 규모의 인공지능 프로세서인 WSE(Wafer Scale Engine) 발표에 이어 '21년 3분기에는 1세대보다 두 배 이상의 성능을 자랑하는 2세대 인공지능 프로세서인 WSE-2*를 발표</li> <li>* 85만개의 인공지능 연산유닛, 20 PB/s에 달하는 메모리 대역폭, 220 Pbit/s에 달하는 입출력 대역폭, 그리고 40 GB에 달하는 on-chip SRAM을 집적</li> </ul>
<b>Syntiant*</b> (미국)	<ul style="list-style-type: none"> <li>• '21년에 발표한 IoT나 스마트폰을 대상으로 한 인공지능 반도체인 NDP120는 대표적인 딥러닝 모델인 CNN과 RNN을 구동</li> <li>※ 8-bit INT 연산 기준으로 최대 6.4 GOPS의 성능을 보여주며 전력소모는 500 uW보다 적게 유지 가능</li> <li>* '17년 설립된 인공지능 반도체 회사로, 3년 연속 EETimes 100개 이머징 스타트업으로 선정된 스타트업</li> </ul>
<b>Mythic</b> (미국)	<ul style="list-style-type: none"> <li>• '20년 11월에 제조, 영상 감시, 스마트 홈, AR/VR, 드론 등 다양한 영역에 엡지 인공지능을 배치하는 걸 목표로 M1108* 아날로그 매트릭스 프로세서(AMP)를 발표</li> <li>* 저전력·고성능 구현을 위한 아날로그 프로세서로, 아날로그 메모리 내 연산(compute-in-memory)이 가능한 Mythic 아날로그 컴퓨트 엔진(ACE)</li> <li>• 또한, 대량의 온칩 버퍼를 통해 외부 메모리(DRAM) 접근을 최소화 하여 범용 연산 제품 대비 높은 전성비를 자랑하는 제품군들을 보유</li> </ul>
<b>Sambanova</b> (미국)	<ul style="list-style-type: none"> <li>• TSMC 7nm 공정 기반 300mb 이상 On-Chip 메모리, 300 TFLOPS(BF16) 이상의 연산속도를 갖는 데이터센터용 인공지능 반도체 생산</li> <li>• AI학습과 추론 연산을 지원하고, AI 연산 시 칩 내 데이터 흐름을 개선, 학습 모델에 따른 재구성이 가능하며, 복합 작업 지원</li> <li>• 8개의 SN10을 칩을 적용한 렉시시스템으로 수 천개의 GPU가 필요한 거대 모델(고해상도 영상 판독, 자연어 학습 등) 학습 데이터를 파편화하지 않고 연산가능한 시스템 구현</li> <li>※ 소프트뱅크, 삼성전자, 블랙락, 구글, 인텔, SKT 등으로부터 누적투자금액 20억달러(2조4천억원)를 돌파하며 동종 업계 최고 수준의 투자유치 달성</li> </ul>
<b>Blaize</b> (미국)	<ul style="list-style-type: none"> <li>• 미국의 인공지능 컴퓨팅 플랫폼 회사인 Blaize社(舊 Thinci)는 '21년 엡지컴퓨팅 프로세서인 Pathfinder P1600을 발표</li> <li>※ 16개의 인공지능 연산프로세서를 이용해 16 TOPS의 성능을 구현</li> </ul>
<b>Groq</b> (미국)	<ul style="list-style-type: none"> <li>• 높은 플랫폼 유연성을 지닌 인공지능 반도체인 TSP를 발표('20)</li> <li>• 기존 경쟁사들 대비 최대 10배 이상 빠른 820 TOPS 성능을 확보한 것으로 발표</li> </ul>
<b>GraphCore</b> (영국)	<ul style="list-style-type: none"> <li>• 여타 인공지능 반도체와 달리 MIMD(multiple instruction multiple data) 명령어에 기반을 둘뿐만 아니라 3차원 패키징 기술을 도입함으로써 더 높은 성능 및 전력효율 향상을 도모하는 인공지능</li> </ul>



업체명(국가)	개발 내용
	<p>연산 프로세서인 BOW IPU*를 발표('19)</p> <p>* 360 TOPS 수준의 연산성능 구현 및 전세대 대비 성능(40%)·전력효율(16%) 향상</p> <ul style="list-style-type: none"> <li>2세대 AI 플랫폼 IPU M2000*은 대규모 인공지능 인프라를 위한 시스템으로 ~PFLOPS의 처리능력 달성('21.11.), '22년에는 4개의 BOW IPU가 들어간 서버인 BOW-2000을 발표</li> <li>* 594억 개의 트랜지스터를 단일 다이에 집적해 TSMC의 7nm 공정으로 제작</li> <li>자사 제품을 위한 ML 프레임워크 호환 각종 라이브러리, 디바이스 드라이버, 리소스 관리 및 가상화 SW를 지원하며 Microsoft Azure, DELL-DSS 및 국내 이동통신 및 플랫폼 기업에서도 도입</li> </ul> <p>※ Microsoft, Dell 등 기존 빅테크의 투자를 받은 것으로 알려짐</p>
Kalray (프랑스)	<ul style="list-style-type: none"> <li>'08년 프랑스 CEA lab으로부터 분사한 Karlay社は MPPA(massively parallel processor array) 아키텍처 프로세서를 설계</li> <li>'13년에 발표된 MPPA 1세대 프로세서(Andey)를 시작으로 '15년, '20년에 차례로 2세대(Boston), 3세대(Coolidge) 프로세서를 공개</li> <li>Coolidge 프로세서는 8-bit INT 기준으로 최대 25 TOPS의 성능을 보이고, 16-bit 소수점 연산 기준 최대 3 TFLOPS의 성능을 보유했다</li> </ul>

- 중국 역시 앞서있는 인공지능·반도체 기술 역량을 바탕으로 다수의 기업에서 자사 제품·서비스 활용을 위한 자체적인 인공지능 반도체 개발 진행 중

※ 한편, 미국의 반도체 분야 대중국 수출규제에 대한 대응 수단으로도 활용

〈표 8〉 주요 중국 기업의 ASIC 기반 인공지능 반도체 개발 현황

업체명	개발 내용
Baidu	<ul style="list-style-type: none"> <li>'11년부터 자체적인 인공지능 반도체 개발을 위한 'Kunlun AI Chip' 프로젝트 추진</li> <li>'18년 14nm 공정 기반 'Kunlun 1세대'를 발표하였으며, '21년에는 'Kunlun Chip Technology'로 분사하여 'Kunlun 2세대' 발표</li> <li>* 7nm 공정 기반으로 INT8 기준 512 TOPS 성능과 120W 전력 소모</li> </ul>
Huawei	<ul style="list-style-type: none"> <li>통신장비 및 스마트폰 제조 업체로 SoC 전담 기업인 HiSilicon社에서 인공지능 기능을 추가한 모바일 SoC를 지속적으로 제작 중이었으나, 미국의 반도체 분야 대중국 제재 강화에 따라 현 시점('23.03.)에서는 신규 제품 설계는 다소 답보 중</li> <li>(Ascend 310) '18년 인공지능 HW 아키텍처인 'Da Vinci'에 뒤이어 발표한 저전력 인공지능 반도체로 12nm 공정 기반 INT8 기준 16 TOPS 성능과 8W 소모전력을 가짐</li> <li>(Ascend 910) '19년 7nm 공정 기반 INT8 기준 640 TOPS 성능과 310W 전력소모를 갖는 데이터센터용 인공지능 반도체</li> </ul>
Alibaba	<ul style="list-style-type: none"> <li>'19년 자체 클라우드용 인공지능 추론을 위한 'Hanguang 800'을 발표</li> <li>* 12nm 공정 기반으로 INT8 기준 825 TOPS 연산성능과 280W 전력 소모</li> </ul>
Biren Technology	<ul style="list-style-type: none"> <li>'22년 발표한 BR100은 여타 GPGPU와 동일하게 SIMT(Single Instruction Multiple Thread) 아키텍처 기반이나, 내부에 텐서 연산유닛을 도입해 인공지능 연산을 가속</li> <li>연산 효율 극대화를 위해 다른 코어의 텐서 연산유닛 간에 직접적인 통신이 가능하며, INT8 기준 2048 TOPS의 성능을 550W 전력소모 내에서 제공(TSMC 7nm 공정 기반)</li> <li>NVIDIA의 'A100' GPU 제품 2.8배 성능을 발표하며 인공지능 반도체 내재화 가능성을 내비쳤으나, 미국의 대중국 규제에 따라 국외(TSMC) 제조 불가로 제품화에 난항</li> </ul>
Bitmain	<ul style="list-style-type: none"> <li>세계 최대 암호화폐 채굴기 생산업체로 고속·저전력 맞춤형 칩 설계 및 연구개발 추진</li> <li>저전력·고성능 16nm 프로세스 집적회로의 대량생산 경험이 있고, 다양한 ASIC과 통합 시스템</li> </ul>

업체명	개발 내용
	생산·구축하여 100개국 이상의 중소기업·개인 사용자에게 판매 중 • 기존 GPU 방식의 암호화폐 채굴기의 성능을 획기적으로 개선하여, 세계 최초로 ASIC 채굴기를 개발·판매
Enflame Technology	• '18년 3월 설립된 스타트업으로, 자체적으로 개발한 인공지능 반도체를 DTU(Deep Thinking Unit)로 명명 • DTU 기반의 20TFLOPS 성능 AI 가속기 솔루션 클라우드블레이저(CloudBlazer) T10을 출시('19.12.)
Horizon Robotics	• 스마트 모빌리티, 감시 카메라, 그리고 다른 스마트 디바이스에 탑재되는 자체적인 인공지능 반도체 개발하고, 이를 통한 자율주행, 딜리버리 로봇 등 솔루션 개발

### 한편, 고성능 인공지능 반도체를 지원하기 위한 다양한 메모리 중심 컴퓨팅, 뉴로모픽 반도체 등 세계적으로 차세대 컴퓨팅 시스템과 연산유닛이 개발 중

※ 3세대 인공지능 반도체 분야는 아직 시장이 형성되지 않은 기술개발 초기 단계로, 세계적으로도 주요 산·학·연을 중심으로 상용화를 위한 기초연구를 수행 중

- (메모리 중심 컴퓨팅) 대규모 학습 데이터셋의 효과적인 처리를 위해 메모리단에서 빠르게 연산이 가능한 “PIM” 기술과 데이터를 효과적으로 저장할 수 있는 메모리 구조 연구 활발

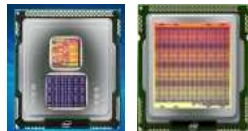


※ 삼성전자, SK하이닉스社 등 국내 주요 업체의 PIM 반도체 개발 사례는 ‘국내 산업동향’에서 논의

〈표 9〉 PIM 반도체 분야 주요 개발 사례

수행 주체	개발 내용
스탠포드 대학	<ul style="list-style-type: none"> <li>• 딥러닝 애플리케이션의 에너지 효율을 개선하기 위해 3차원 메모리 구조와 로직다이를 결합한 PIM 기술을 발표(ASPLOS, 2017)</li> <li>• 기존 ASIC 형태의 인공지능 반도체 대비 1.5배의 에너지를 절약하고 4.1배의 성능향상</li> <li>• 22년, RRAM 기반 CIM(Compute-in-memory) 칩 발표(Nature, 2022)</li> <li>• 메모리 자체 내에서 AI 처리를 수행하는 RRAM구조로 구글 음성 명령 인식 정확도는 84.7%, 베이지안 데이터의 이미지 재구성 오류 70% 감소</li> </ul>
메타 (舊페이스북)	<ul style="list-style-type: none"> <li>• '20년, 페이스북에서 개인용 추천시스템 내에서 임베딩 테이블을 참조하는 연산을 가속하기 위해 PIM 기술인 AxDIMM의 프로토타입을 시뮬레이션 레벨에서 시연(ISCA, 2020)</li> <li>• '21년, 삼성과의 협력을 통해 그 기술을 실제 하드웨어 반도체로 검증하는 논문을 연구성과로 발표함(IEEE Micro, 2021)</li> </ul>
UPMEM (프랑스)	<ul style="list-style-type: none"> <li>• 프랑스 스타트업 기업으로 '19년 DDR4 메모리 기반 AI 가속기용 PIM 칩을 출시</li> <li>• 'Hot Chips 31 4GB DRAM' 칩은 기존의 AI가속 연산 기준 DRAM에 비해 20배의 성능향상과 10배의 에너지 절약, 10배의 비용 절감의 성능으로 발표</li> <li>※ 메모리 모듈 내부에 인공지능망 실행을 위한 MAC 연산기능을 구현하여 PCIe와 같은 시스템버스에 연결되는 GPU 및 NPU와는 달리 상대적으로 속도가 빠른 메모리 버스에 연결이 가능한 장점</li> </ul>

- (뉴로모픽 반도체) 뇌과학 분야에 근간한 연구 및 개발이 활발히 진행 중이며, 이를 통해 뇌의 동작 및 학습의 구조에 관한 연구가 진행 중
  - 생체의 뉴런·시냅스의 모델을 에뮬레이션하는 CMOS기반 아날로그·디지털 반도체 집적화 연구 및 메모리 소자·디바이스가 적용된 신개념의 반도체 기술개발이 진행 중

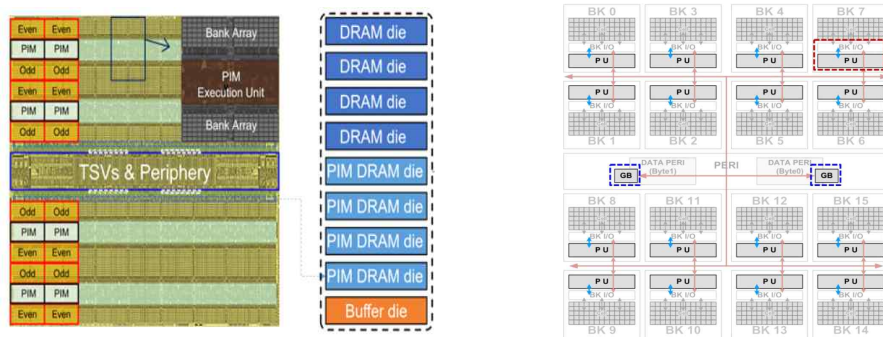
〈표 10〉 뉴로모픽 반도체 개발 사례

수행 주체 (국가)	개발 내용	
드레스덴 공과 대학 (독일)	 BrainScaleS (2012)	<ul style="list-style-type: none"> <li>• TFLOPS급 슈퍼컴퓨터 시스템을 혼합하여 98,604개의 시냅스, 384개의 뉴런을 내장한 뉴로모픽 칩</li> <li>• Plasticity model의 대용량 에뮬레이션 가능</li> </ul>
Qualcomm (미국)	 Zeroth (2013)	<ul style="list-style-type: none"> <li>• 생물체처럼 학습하는 스파이크 형식의 뉴런 활동을 구현한 'Zeroth' Core 발표</li> <li>• SoC-SW ZDS(Zeroth Development Studio)-Xilinx K7 FPGA 플랫폼 및 Zeroth NPU 지원('15)</li> </ul>
맨체스터 대학 (영국)	 SpiNNaker (2013)	<ul style="list-style-type: none"> <li>• 대규모 병렬처리 뉴로모픽 슈퍼컴퓨터로써 130nm 공정을 통해 18개의 ARM968 코어와 synapse weight 저장을 위한 128MB SDRAM이 NoC로 연결됨</li> <li>• 518,400개의 프로세서로 확장될 수 있는 구조를 이용하여 10개의 뉴런 시뮬레이션 가능</li> </ul>
IBM (미국)	 TrueNorth (2014)	<ul style="list-style-type: none"> <li>• 4,096개의 코어로 인간 뇌 100만 개의 뉴런(neuron)과 2억 5,600만 개 시냅스(synapse)를 재현</li> <li>• 초기 뇌모사 컴퓨팅 구조로(Brain-inspired computer architecture) 삼성 28nm CMOS 공정 사용</li> <li>• 70mW의 전력으로 실시간 연산을 통해 인간처럼 물체를 식별하고 패턴 인식 등의 처리 가능</li> </ul>
Intel (미국)	 Loihi (2018)	<ul style="list-style-type: none"> <li>• 비동기식 SNN(Spiking neural network) 알고리즘 구현</li> <li>• Intel 14nm 공정으로 제작되었으며, 128개의 뉴로모픽 코어와 3개의 lakemont x86 코어로 구성</li> <li>• 칩 간의 확장이 가능하게 설계되었으며 최대 4096개의 온-칩 코어와 최대 16,384 개의 칩을 지원</li> </ul>
	 Loihi2 (2021)	<ul style="list-style-type: none"> <li>• 전 세대 대비 최대 10배 빠른 동작 속도와 8배 많은 뉴런 제공, 4배 넓어진 링크 대역폭 제공</li> <li>• 다양한 스파이크 신호 프로그래밍 기능과 프로그래밍 가능한 뉴런 및 고속 이더넷 인터페이스 지원</li> </ul>
GML (프랑스)	 GrAI One (2019)	<ul style="list-style-type: none"> <li>• 초저지연·초저전력 엣지 프로세싱에 최적화된 최초의 인공지능 반도체</li> <li>• 총 20만개의 뉴런에 대해 로컬 뉴런·시냅스 메모리가 장착된 196개의 뉴런 코어로 조직</li> <li>• TSMC 28nm 공정으로 제작되었으며 인간 뇌를 모사한 NeuronFlow 기술을 기반으로 응용프로그램 대기시간 감소</li> </ul>
	 GrAI VIP (2022)	<ul style="list-style-type: none"> <li>• 풀스택 AI SoC 플랫폼으로 실시간 이벤트 기반 컴퓨팅 솔루션</li> <li>• 16비트 부동 소수점 기능을 탑재한 업계 최초의 근접 센서 AI 솔루션</li> </ul>

## 3.2 국내 산업동향

국내 대기업은 PIM 등 차세대 인공지능 반도체, 스타트업은 모빌리티·IoT·금융 등 응용 분야 특화 인공지능 반도체 개발·상용화 추진 중

- 삼성전자, SK하이닉스 등 메모리 반도체 분야 초격차 기술을 바탕으로 메모리에 연산유닛을 집적한 PIM구조의 차세대 메모리 개발 중
  - 삼성전자는 DRAM 메모리와 AI연산기를 하나로 결합한 HBM-PIM 발표('20)
    - ※ 데이터센터와의 연결이 없는 모바일과 환경에서 독자적인 인공지능 수행 성능(On-Device AI)과 관련해 시뮬레이션을 통해 음성인식, 번역, 챗봇 등에서 2배 이상의 성능향상 발표
  - SK 하이닉스는 퍼듀 대학 연구진들과 함께 HBM2E 구조의 GPU 같은 이중 컴퓨팅 장치용 메모리를 위한 GDDR6-AiM 칩\* 발표('22)
    - \* Non-PIM 기반 시스템과 GPU보다 각각 약 10배, 54배의 성능향상을 보이며, 기존 동작 전압인 1.35V보다 낮은 1.25V에서 구동되어 전력 효율 향상이 높음



[그림 14] (좌)삼성전자 HBM-PIM 구조, (우)SK하이닉스 GDDR6-AiM 구조

\* 출처: 삼성전자, SK하이닉스

- 국내 중소·중견 팹리스는 ASIC 기반 응용분화 특화 인공지능 반도체를 개발·상용화 추진 중

<표 11> 국내 주요 팹리스의 ASIC 기반 인공지능 반도체 개발 현황

업체명	개발 내용
사피온	<ul style="list-style-type: none"> <li>• SKT로부터 분사·독립하였으며, 6.7kFPS/60W의 추론용(서버용) 인공지능 반도체 Sapeon X220(28nm, '20)을 개발하고 MLPerf에서 상용화 등급(Available) 인정</li> <li>• SKT의 5G 인공지능 서비스 및 자율주행차 등 상용화 추진 중</li> </ul>
퓨리오사 AI	<ul style="list-style-type: none"> <li>• AI 반도체 벤치마크 MLPerf에서 엔비디아의 T4 대비 4배 성능의 영상인식에 특화된 NPU 'Warboy' 개발('21) ※ 64TOPS, 이미지 분류 0.74ms/2,634fps</li> <li>• 추론성능에서 NVNDIA A100과 경쟁이 가능한 기술로 발표하였으며, '23년 GPT-3와 같은 자연어 처리를 위한 거대 인공지능경망용 반도체 출시 계획</li> <li>• 자율주행차·클라우드·의료분야 영상진단 등 최첨단 기술에 활용과 카카오, 네이버 등의 컴퓨터비전-메타버스-하이퍼스케일 분야의 사업화 추진 중</li> </ul>

업체명	개발 내용
리벨리온	<ul style="list-style-type: none"> <li>• 인텔 Goya보다 성능 우위 30%의 금융에 특화된 NPU인 아이온 칩 발표('21.12)</li> <li>• 실시간 트레이딩과 같이 빠른 처리속도가 중요한 금융분야 AI 응용에서 nVidia A100보다 연산 속도는 10배 빠르고, 전력 소모는 10W로 경쟁제품의 10% 수준으로 발표</li> <li>• 트랜스포머 모델 지원하는 데이터센터용 아톰(ATOM)* 출시('23.02.)</li> </ul> <p>* KT초거대 인공지능 '믿음' 지원</p>
딥엑스	<ul style="list-style-type: none"> <li>• 엣지 디바이스, 자율주행차, 데이터센터 등 각 어플리케이션에 특화된 NPU인 '제네시스(GENESIS)*' 개발('22.)</li> </ul> <p>* 테슬라 NPU대비 5배이상의 연산 성능으로 10TOPS/W의 우수한 전력 효율(1W 소모)을 발표</p>
모빌린트	<ul style="list-style-type: none"> <li>• 자율주행차, 사물인터넷(IoT) 기기, 블랙박스, 폐쇄회로(CCTV), 도어록 등 저전력 인공지능 반도체 개발에 주력</li> </ul>
디퍼아이	<ul style="list-style-type: none"> <li>• 팹리스 기업으로서 NPU를 내장한 CCTV 및 로봇용의 AI 반도체 SoC 양산 예정</li> <li>• 다수의 AI 반도체 개발 특허를 보유하고 있으며 엣지용 AI 반도체를 위한 NPU 기술을 독자 확보한 것으로 알려져 있음</li> </ul>
뉴블라	<ul style="list-style-type: none"> <li>• '한화임팩트'가 주도하여 설립한 업체로 NPU반도체 설계 및 IP 개발에 집중</li> </ul>
LG전자	<ul style="list-style-type: none"> <li>• 기존 TV의 고화질처리용 AI반도체를 가전제품에 적용할 수 있는 범용 칩 자체 개발</li> <li>• CTO 부문 내의 SIC센터가 담당한 것으로 영상 음성처리 및 제어 기능이 가능한 범용 SoC 및 MCU로 볼 수 있음</li> </ul>
ETRI	<ul style="list-style-type: none"> <li>• 국내 최초 40TFLOPS급 AI 반도체인 'AB(Artificial Brain)9' 개발('19.12)하였으며, 5PF급인공지능 서버를 개발 완료('21.07)</li> </ul>




[그림 15] 국내 인공지능 반도체 (왼쪽부터 (주)사피온 X220, (주)퓨리오사AI Warboy, (주)리벨리온 아이온)




## 제4장 정책동향

본 장에서는 주요국의 반도체 산업 전반에 대한 정책을 함께 살펴보고  
인공지능 반도체 분야의 주요 시사점을 도출하고자 함

### 4.1 해외 정책동향

 디지털전환 등 대외환경 변화와 반도체 산업에 대한 산업·안보 관점에서의 인식변화로  
주요 선도국은 자국 공급망 확보와 기술우위를 위한 대대적인 지원 정책 추진 중

- ‘디지털전환’의 가속화에 따라 전 산업에 대한 인공지능·반도체의 활용·확산으로 기존 PC  
부품에서 모바일, 엣지컴퓨팅, 모빌리티, 네트워크 등 수요 산업 증가
  - 반도체 분야는 그간 수출·입 품목 측면의 무역·통상 측면에서 산업·안보적 전략자산 관점  
으로 급격한 인식변화가 이루어지며 국가적 차원의 핵심기술로 자리매김
  - 한편, 코로나19 팬데믹과 여러 자연재해\*로 비롯된 수요예측 불확실성 및 공급 불안  
등으로 그간 공급망에 대한 취약성 노출
    - \* (대만) 가뭄('21년 봄)으로 주요 파운드리 가동 중단, (일본) 지진('21.02.) 및 화재('21.03.)로 르네사스社 공장  
복구에 3개월 소요, (미국) 텍사스 한파정전('21.02.)에 따라 주요 차량용 반도체 제조업체 공장 가동 중단
  - 이에 따라 주요국은 이른바 ‘자국중심주의’를 앞세워 자국 내 공급망 확보와 첨단기술 선점을  
위한 대대적인 지원 정책을 추진 중

 (미국) 세계 반도체 시장의 과반을 차지하며 산업을 주도 중인 미국은 꾸준한  
정부투자와 상대국에 대한 지속적인 견제를 바탕으로 기술패권을 유지

- 바이든 대통령은 「반도체과학법(Chips and Science Act)」에 서명하며 자국 반도체 산업역량  
강화를 위해 520억 달러 규모의 투자 지원('22.08.)
  - ※ 한편, 반도체법 이행에 관한 행정명령(Executive Order)에 서명('22.08.25.)하여 반도체 법의 세부기준  
마련을 위한 조정위원회(CHIPS Implementation Steering Council)를 설치하고, 법률 이행 우선순위를 설정
  - 첨단반도체 생산 촉진을 위한 연구개발 지원과 자국 반도체 생산역량 제고를 위한 390억  
달러 규모의 인센티브\* 등을 포함



- \* 동 법안에는 반도체 관련 투자기업에 25% 공제를 지원하는 내용도 포함되어 있으며, 이를 통해 국내 공급 확보 및 고속련 제조 일자리 창출 등 민간 투자 촉진을 유도
- 한편, 인센티브를 받은 기업이 우려 대상 국가(중국, 러시아, 이란, 북한 등)에 향후 10년간 반도체 시설에 대한 투자를 금지하는 ‘가드레일 조항’을 적용하여 기술개발을 견제
  - ※ 단, 레거시 반도체(28nm 이상)의 생산·특정거래의 경우 정부의 승인으로 가드레일 조항에서 제외 가능
- '22년 8월 이후 첨단 인공지능·반도체 기술에 대한 중국, 러시아 등의 접근을 차단하는 수출통제 정책을 강화
  - 중국·러시아에 대해 ‘국가안보위협 기술개발 가능성’을 이유로 고성능 그래픽카드의 수출 허가제를 실시(8.26.)
    - ※ '22년 8월 이전까지 미국은 반도체 및 생산재료, 설계SW·제조장비에 대해 전방위적인 규제를 지속하였으나, 인공지능 반도체 완제품에 대한 규제는 모호한 측면이 존재
  - 또한, 중국의 인공지능·반도체 기술·산업 발전 억제를 위해 지난 '22년 두 차례(9월, 10월)에 걸쳐 글로벌 공급망 길목(Check Point) 통제 조치를 강화
    - ※ (9월) 자국 반도체 설계기업(AMD, 엔비디아 등)의 인공지능·슈퍼컴퓨터 활용 반도체 대중국 판매 제한 (10월) 반도체 수출 통제리스트를 강화하고 미검증 기업 목록을 확대하는 등 수출 통제
  - 이와 같은 수출통제에 따라 중국 기업은 막대한 타격\*을 입게 되었으며, 주요 반도체 업체는 미·중 기술경쟁 장기화에 대비
    - \* 일례로 화웨이社は 미국 제품에 의존하던 많은 생산시설의 운영을 중단했고 SMIC社は ASML 社の 반도체 제조기기 수입 금지로 첨단반도체 생산역량 일부 상실
  - ※ 중국은 이러한 수출통제 조치를 WTO 규범 위반으로 제소('22.12.)하였으며, 대만·한국 주요 반도체 기업은 미국 수출통제를 염두하여 공급망 재편, 미국 내 공장 건설 추진 중

〈표 12〉 미국의 반도체 분야 수출·접근통제 주요 내용

접근 제한 품목	주요 내용
첨단 인공지능 반도체	<ul style="list-style-type: none"> <li>• 대규모 인공지능 모델을 위한 데이터센터 및 슈퍼컴퓨터 구현 등을 위한 일정 성능 이상의 첨단반도체*는 상무부의 수출 허가가 필요               <ul style="list-style-type: none"> <li>* 강력한 연산성능(300TFLOPS 이상)과 빠른 데이터 입출력 속도(600GB/s 이상)</li> <li>※ 미국의 NVIDIA와 AMD는 이러한 첨단 반도체 설계·생산이 가능한 대표적인 업체로, 이들 제품을 사용하지 않는 경우 인공지능 구현을 위한 범용SW 생태계에서 제외될 수 있음</li> </ul> </li> </ul>
설계SW	<ul style="list-style-type: none"> <li>• 해외직접생산품규칙*을 통해 중국 팹리스 기업이 자국 SW를 활용하여 설계한 반도체는 중국 외에서 생산할 수 없도록 제한               <ul style="list-style-type: none"> <li>* Foreign Direct Product Rule : 제3국 생산 제품이라도, 미국의 기술이나 SW를 활용하여 생산한 경우 미국(상무부) 허가 없이 수출금지</li> <li>※ 반도체 설계 SW(EDA) 선도 기업은 대부분 미국에 본사를 두고 있어 이와 같은 조치를 통해 중국 등 해외 반도체 설계 산업 성장 억제 가능</li> </ul> </li> </ul>
반도체 제조장비	<ul style="list-style-type: none"> <li>• 반도체 종류(로직, DRAM, NAND 등)에 따라 수출제한 조치를 차별화*하여 적용하고, 기존 제조 장비 뿐 아니라 검사·측정장비 등 응용 도구의 공급 제한               <ul style="list-style-type: none"> <li>* (로직) FinFET 구조나 16/14nm 이하, (DRAM) 18nm 이하, (NAND) 128-layer 이상</li> </ul> </li> </ul>
자국 제조 부품	<ul style="list-style-type: none"> <li>• 중국의 자체적인 생산 장비 개발을 막기위해 관련 부품의 수출을 금지               <ul style="list-style-type: none"> <li>※ 반도체 제조 기기·부품 분야는 미국·일본·네덜란드 등 소수의 주요국이 지속적인 경쟁 우위를 가짐</li> </ul> </li> </ul>

## (중국) 국가전략 수립, 펀드 운용, 세제지원 강화 등을 통해 반도체 자급화를 추진 중이나, 미국 등 주요국의 거센 압박 속 난관에 봉착

- 중국은 '14차 5개년 계획 및 2035 중장기 목표('21.03.)'의 7대 전략육성 분야\* 중 하나로 반도체 분야를 선정하는 등 대대적인 지원을 추진 중이나 자립화율은 답보 상태

\* 인공지능, 양자정보기술, 반도체, 뇌과학, 유전자 바이오 기술, 임상의학, 심해·극지·우주 분야

- 동 계획은 '중국제조 2025'로 널리 알려진 '13차 5개년 계획('16)'의 후속 국가전략으로 미국의 제재 분야\*에 대한 전략적 지원이 이루어질 것으로 전망

\* 반도체 설계툴, 주요 소재·장비, 제조기술, 화합물 반도체 등 개발을 구체적으로 언급

- 대규모 펀드와 각종 세제지원 등 전방위적인 자국 반도체 산업 육성정책을 추진 중이나, 여전히 자립화율은 15% 수준으로 답보 상태

※ '14년부터 '국가집적회로산업투자펀드(약칭 대기금)' 조성을 통해 자국 제조분야와 공백영역에 대한 지원으로 생태계를 육성, 반도체를 비롯한 첨단기술 보유 기업에 대한 세제혜택 강화 등을 추진

- 한편, 앞서 언급한 미국의 대중국 수출금지 등 압박과 설계·장비 분야에 대한 자체적인 기술혁신 난항으로 자립화에 지속적인 어려움에 봉착

- 미국의 그래픽카드 무역 통제가 자국 인공지능 기술 연구에 크게 영향을 미치지 않을 것이라는 관측\*도 존재하나, 반도체 산업에 대한 전방위적 규제로 산업 기반 타격 불가피

\* 그래픽카드 전량에 대한 규제가 아닌 이상 대체품을 통한 연구개발이 가능하며, 자국 기업(Biren Technology社 등)이 유사 성능의 제품을 발표함으로써 중장기적으로 내재화 가능성도 존재

- 유력 매체 보도('23.01.)에 따르면 세수, 코로나19 확산 등 재정 부담에 따라 그간 투자 주도 접근을 중단하고, 대안으로 반도체 소재 가격 인하 등을 모색하는 방안을 검토

※ 그간 450억 달러(약 57조원) 규모의 '대기금' 펀드를 통해 SMIC, YMTC 등 자국 기업을 지원해 왔으나, 기술개발 돌파구 마련보다는 뇌물 등 부패와 미국의 제재를 불러왔다는 시각이 존재

## (대만) 반도체 제조 분야 기술우위를 기반으로 정부 지원을 통해 인공지능 반도체를 비롯한 핵심 장비·소재 분야 전략자원·기술의 내재화를 추진

- 첨단반도체 분야 기술·산업선도를 위해 ▲제조기반 고도화 ▲기술 및 핵심소재·장비 경쟁력 강화 ▲고급인재 확보 등을 지속 지원하고, 생태계 활성화를 위한 법·정책 등 기반 마련

- 대내적으로는 고급인재의 육성과 확보, 연구개발 등을 중점 추진하고, 대외적으로 글로벌 소재·장비 기업에 대한 대만 유치 확대하여 장비·소재·SW의 내재화 추진('21.09.)

- 또한, 일명 '대만형 CHIPS'법으로 일컫는 「산업혁신 조례 수정안」 통과('22.11.)로 첨단 공정개발 촉진을 통한 경쟁 우위와 반도체 업체 활동기반 강화('23년 연내 시행)

※ 기술혁신·세계 공급망에서 중요한 위치를 차지하는 기업이 연구개발 및 선진 생산공정 설비에 투자하는 경우 각각 투자비의 25%, 5% 세액을 공제하는 등 인센티브 제공

- 한편, 대만 정부는 '18년부터 「인공지능 반도체 제조공정 및 칩 시스템 R&D 프로젝트 (Semiconductor moonshot Project)」를 추진하고 4년간 약 1,526억원 규모의 예산 투입
  - 인공지능 반도체 제조공정, 칩 시스템 연구개발에 집중하여 6대 유망기술\*을 개발하고, 산업 경쟁력 제고를 위해 설계 인재 육성에 집중
    - \* ▲센서 관련 소자·전자회로·시스템 ▲차세대 메모리 설계 ▲인지 컴퓨팅 및 AI 칩 ▲사물인터넷 시스템·보안 ▲자율주행차·AR/VR 관련 소자·전자회로·시스템 ▲반도체 제조공정·재료·소자 관련 신기술
  - 동 정책의 일환으로 '대만반도체연구소(TSRI)'를 설립('19.01.)하여 인공지능 반도체 개발 역량을 강화하기 위해 설계·공정·검증·인력양성 등 통합 서비스를 제공
  - 또한, 'AI on chip Taiwan Alliance(AITA)'를 통해 ITRI(산업기술연구소), UMC社 등 유관 기관\*이 4개의 SIG\*\*에 참여하여 AI반도체 관련 산업의 수직통합을 통한 생태계 강화 추진('19.01.)
    - \* 출범 당시 56개 참여사로 출발하여 152개 국내외 주요 반도체 업체로 확대('23.01.기준)
    - \*\* AI반도체의 이기종 통합을 위한 ▲공동인터페이스 ▲고성능 아키텍처 구현 ▲설계 가속 SW ▲AI 시스템 통합·적용 기술

#### (유럽) EU회원국은 「반도체 법안(Chips Acts)\*」 통과, 인공지능 반도체 설계기술 확보를 위한 EPI(European Process Initiatives) 수립 등 지역 내 경쟁력 강화 추진 중

\* '30년까지 세계시장 점유율 20%를 목표로 민간·공공협력을 통해 430억 유로 투자에 합의

- 유럽의 반도체 법안은 공급망 위기에 대응하기 위한 ▲권고(Recommendation)\*와 첨단반도체 기술개발·생산확대를 주요 골자로 하는 중장기계획인 ▲규정(Regulation)\*\*으로 구성
  - \* EU와 회원국 간 반도체 시장 상황 모니터링을 통한 조기경보 및 정보공유 체계 구축
  - \*\* ▲연구·혁신 ▲설계·생산 ▲공급안정 ▲위기관리 ▲국제협력 등 주요 부문별 정책 방향 제안
- 동 법안의 차질 없는 이행을 위해 110억 유로 규모의 기금으로 '반도체 이니셔티브(Chips for Europe Initiative)'를 발족하고, 유럽 내 ▲집적기술·설계 능력 ▲접근성 높은 파일럿 라인 ▲설계·생산·장비 기업 간 협력체계 등 확보를 통해 산업생태계 강화를 추진
  - ※ 특히, 생산 측면에서 뉴로모픽 소자, 인공지능, 양자기술 등 혁신 기술 연구개발을 위한 파일럿 라인의 구축을 지원
- 또한, 「공동관심 주요 프로젝트(Important Project of Common European Interest, IPCEI)」를 통해 ▲AI 프로세서 ▲엣지컴퓨팅 ▲모빌리티 ▲5G·6G 등 기술개발 프로젝트 지원
  - ※ 한편, 2nm 반도체, 인공지능, 신소재, 3D 이종집적(heterogeneous) 패키징 기술, 첨단 설계 솔루션 등 기반기술 강화 병행 추진

- 한편, 인공지능 반도체 분야 설계 기술력 확보를 위한 이니셔티브 운영('20), 엣지컴퓨팅 응용을 위한 HW 개발 로드맵 발표 등('21) 기술주도권 확보를 위한 지속적인 노력 중
- 빅데이터 처리 초고성능 컴퓨팅 신생 첨단 분야 등 적용·활용을 위해 저전력·고성능 CPU 설계 R&D EPI(European Process Initiatives)\*를 수립·운영('20~)
  - \* EU 10개국 28개 기관이 컨소시엄을 구성하여 가속기, 차량용 반도체 등 분야의 기술 개발 및 '24년까지 로드맵 수립을 추진
  - ※ GlobalFoundries社 12nm 공정을 통해 제작된 RISC-V 기반 가속기 칩(1GHz, 200Gbit/s(인터페이스))에 대한 EPAC 테스트 완료('21.06.)
- EPoSS(European Technology Platform on Smart System Integration)\*는 엣지컴퓨팅용 인공지능 반도체·알고리즘 분야 기술개발 로드맵 발표('21)
  - \* 마이크로·나노-스마트 시스템의 융합을 위한 요구사항 정의, 정책·전략을 개발하는 산업 기반 협의체
  - ※ '31년 엣지에서 학습방법을 학습하는 메타러닝 알고리즘과 이의 연산을 보조하는 뉴로모픽 반도체 개발 포함

## (일본) 지난 20여 년간 쇠퇴한 반도체 산업의 재부흥과 디지털전환 등 대외 환경 변화에 대응하기 위한 「반도체전략」을 발표('21.06.)

- 일본의 반도체 제조공장 수는 세계 1위로 알려져 있으나, 대부분 노후화되어 30~40nm 수준의 공정이 대다수를 차지하는 등 첨단반도체 제조능력은 부재
- '80년대 말 정점을 찍은 일본의 반도체 세계시장 점유율(50%)은 '90년 대부터 하락하기 시작하여 '20년 기준 9% 수준으로 쇠퇴
- 다만, 지난 '19년 수출규제에서와 같이 일본은 여전히 반도체 분야 핵심 소재·부품·장비에서 세계적인 경쟁력\*을 보유
  - \* 세계 반도체 시장에서 일본의 반도체 공급·생산 점유율은 각각 9%, 19%이나, 제조 장비·소재는 32%, 56%를 차지(STEPI, 2021)
- 일본은 「반도체전략」을 통해 자국 반도체 산업경쟁력을 재건하고자 ▲첨단반도체 양산체제 구축 ▲설계·개발 강화 ▲그린 이노베이션 ▲제조기반 구축 등 추진
- 강점 분야인 소재·장비 분야의 격차를 유지하고, 글로벌 파운드리 유치로 통해 자국내 첨단반도체 양산시설 구축 추진
  - ※ 일본 정부는 구마모토현에 TSMC·소니·덴소의 합작법인 JSAM이 운영하는 TSMC 신공장 구축을 허가('22.06.)하고 건설 비용의 절반 수준인 4760억엔(약 4조6천억원)을 지원을 약속하였으며, 최근 다수 매체에서 제2공장 건설 검토 관련 보도('23.01.)
- 특히, 발전된 5G·AI·IoT 등 디지털기술을 활용한 자율주행, 공장 자동화, 스마트시티 등 어플리케이션 시스템 등에 필요한 로직 반도체 설계 및 개발 추진

## 4.2 국내 정책동향

우리 정부는 '19년 이래로 매년 국가 주도의 반도체 분야 경쟁력 강화를 위한 지원 정책을 발표 중이며, '20년부터는 '인공지능 반도체' 중심의 지원방안을 지속 제시

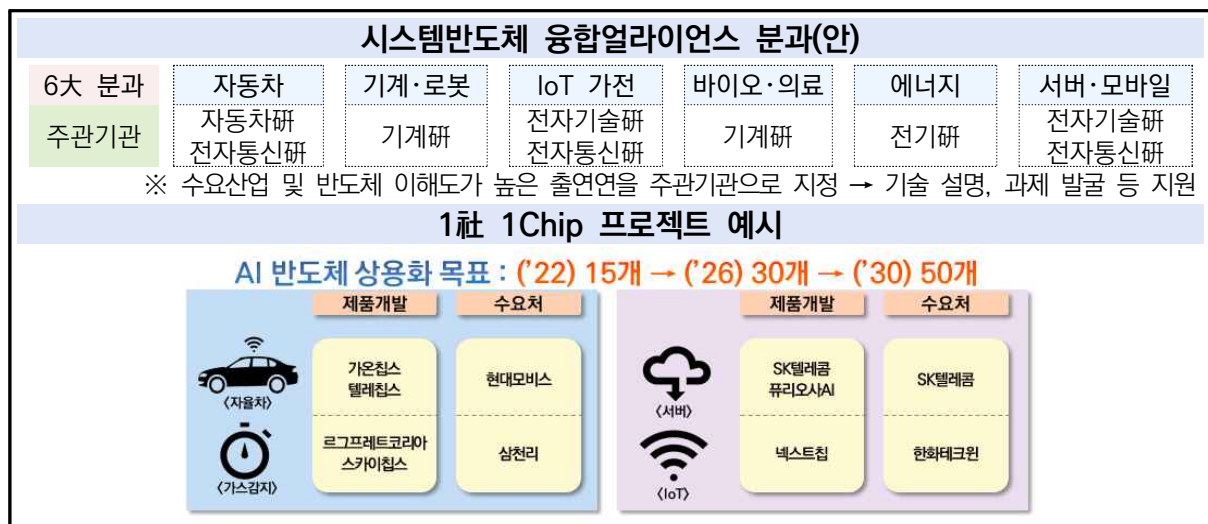
- ('20년) 「인공지능 반도체 산업 발전전략(시스템반도체 비전과 전략 2.0)」을 통해 최초로 인공지능 반도체 분야 중심의 지원전략을 제시
- 개발 초기 단계인 인공지능 반도체 시장을 선점하고, 선도국가로 도약하기 위한 기술개발·인력확보와 산업생태계 조성을 위한 전방위적인 지원 추진
- 대표적으로 수요기업-팹리스간 연계(융합얼라이언스2.0)를 바탕으로 인공지능 반도체 공동 개발을 통해 '1社 1Chip 프로젝트'로 '30년까지 인공지능 반도체 50종 개발\*' 등을 제시

\* 既구축된 융합얼라이언스2.0을 6개 분과로 확대하고, 분과별 주관기관을 선정하여 산업 맞춤형 수요발굴

〈표 13〉 인공지능 반도체 산업 발전전략('20.10.) 주요 내용

목표	<ul style="list-style-type: none"> <li>• 인공지능 반도체 분야 시장점유율 : ('26)10% → ('30)20%</li> <li>• 인공지능 반도체 혁신기업 : ('26)10개 → ('30)20개</li> <li>• 인공지능 반도체 고급인력 : ('24)1,000명 → ('30)3,000명</li> </ul>
추진과제	<ul style="list-style-type: none"> <li>• (퍼스트무버형 혁신 기술·인재 확보) 인공지능반도체 플래그십 프로젝트 추진, 신개념 PIM 반도체 기술개발, 국가 인공지능·데이터 기반 AI반도체 시범 도입, 민·관 공동투자를 통한 고급인재 양성</li> <li>• (혁신성장형 산업 생태계) 1社1Chip 프로젝트를 통해 '30년까지 수요맞춤형 AI칩 50개 출시, 기업간 연대 협력을 통해 AI반도체 설계 역량 강화, AI반도체 혁신기업 스케일업 촉진을 위한 펀드 지원, "인공지능 반도체 혁신설계센터" 신규 구축</li> </ul>

\* 출처: 관계부처 합동, 「인공지능 반도체 산업 발전전략」, 2020. 재구성



[그림 16] 인공지능반도체 시장 조기 성숙을 위한 수요-공급 연계 확대 방안

\* 출처: 관계부처 합동, 「인공지능 반도체 산업 발전전략」, 2020.

- ('21년) 「K-반도체 전략」은 글로벌 패권경쟁 대응과 반도체 공급망 안정화를 위해 '30년까지 510조 원 이상 자금 투입 계획을 제시

- 세계 최고 반도체 공급망 구축을 목표로 ①“K-반도체 벨트” 조성 ②반도체 제조 인프라 지원 확대 ③인력·시장·기술 등 성장기반 강화 ④반도체 위기 대응력 제고
- 또한, 인공지능 반도체 분야 선도기술 확보를 위해 ①선도형 원천기술, ②상용화 응용기술, ③수요연계 실증 등 추진

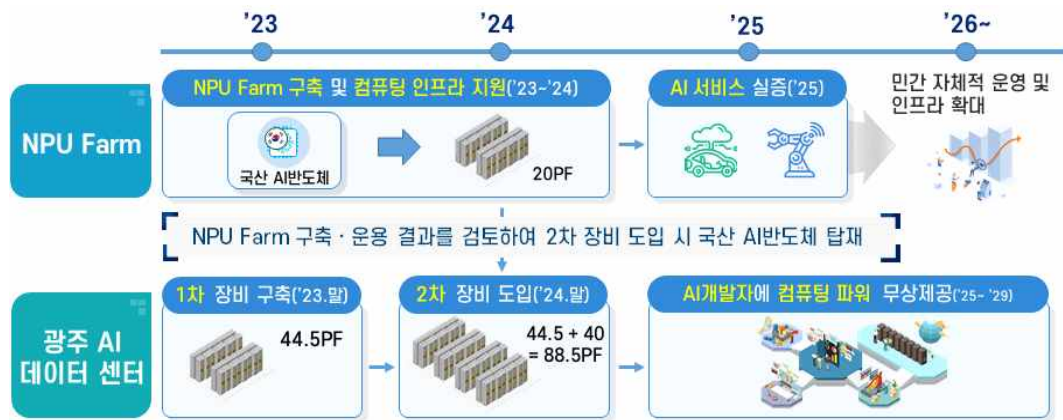
〈표 14〉 인공지능 반도체 선도기술 확보 방안

추진 전략	주요 정책
원천기술	<ul style="list-style-type: none"> <li>• 고성능·저전력 NPU* 등 독자적 기술력을 확보(~'24)하고, 핵심 기술간 연계·융합으로 차세대 AI반도체 개발(~'29)</li> <li>* 서버·모바일·엣지용 NPU 플랫폼 기술 개발 → 민간 개방을 통한 성과 공유·확산 촉진('22~)</li> <li>** 세계 최고 수준인 “1 PFLOPS”급 연산능력과 1mW급 전력 구현</li> <li>- 미래 컴퓨팅 패러다임을 바꿀 신개념 PIM 반도체 기술, Lv.4 이상 자율주행용 AI 반도체 HW·SW 플랫폼 개발 신규 추진</li> </ul>
응용기술	<ul style="list-style-type: none"> <li>• 1조원 규모의 대규모 R&amp;D* 활용 AI 반도체 응용기술 개발</li> <li>- 민간 수요 연계 산·학·연 협력 응용기술 개발* 지원 및 국내 기업이 취약한 SW 역량 강화 등 기술·사업화 장벽** 해소('21~)</li> <li>* 학·연이 보유한 R&amp;D 원천기술(특허 등)을 팹리스에 기술이전, 인력지원</li> <li>** 미세공정 전환(7nm 급), 신규 설계자산(IP) 개발·활용, 시스템SW 최적화 등 맞춤형 지원</li> </ul>
실증지원	<ul style="list-style-type: none"> <li>• 국내 개발된 AI 반도체 기술·제품을 민·관 데이터센터 및 디지털 뉴딜 프로젝트에 적용하여 상용화 실적(레퍼런스) 확보('21~)</li> </ul>

\* 출처: 관계부처 합동, 「K-반도체 전략」 (2021)

- 한편, '21년 「인공지능 반도체 선도국가 도약 지원사업」을 발표하고 3대 분야 13개 사업을 중심으로 ▲핵심기술 개발, ▲혁신기업 육성, ▲산업기반 조성을 추진하고, '30년까지 인공지능 반도체 세계시장 20%를 점유, '제2의 DRAM'으로 육성하는 것을 목표로 설정
- ('22) 「인공지능 반도체 산업 성장 지원대책」은 ▲초격차 기술력 확보, ▲초기 시장수요 창출, ▲산·학·연 협력 생태계 조성, ▲전문인력 양성 등 전략적 산업 육성을 위한 주요 추진과제를 제시
  - ※ '(1차) 인공지능 반도체 최고위 전략대화('22.6.27.)' 제1호 안건
- NPU, PIM, 뉴로모픽 新소자 등 초격차 기술 선점을 위한 원천기술 개발을 강화하고, 시스템SW, 초거대 AI시스템 등 SW 기술력 확보와 관련 국제협력\* 등 강화
  - \* 한미 정상회담('21.5.) 후속 조치로 미국과 첨단기술 교류·협력 강화를 위한 공동연구 추진
- NPU Farm 구축, 공공인프라 도입\*을 바탕으로 민간 확산을 추진하고, 적용 제품·서비스 실증 지원\*\* 및 공공분야 도입·확산으로 국내 인공지능 반도체 기반 조성
  - \* 광주 AI 집적단지에 조성 예정인 대규모 데이터센터 컴퓨팅 장비에 국산 AI반도체를 도입·활용('24~)
  - \*\* AI반도체 실증사업('23, 55억원), AI바우처 사업에 AI반도체 전용트랙('23, 200억원) 신설





[그림 17] 인공지능 반도체 인프라 적용·확산 (출처: 과학기술정보통신부, 2022)

- 반도체 대기업과 산·학·연 간 연대·협력을 강화하여 개방형 혁신생태계\*를 조성하고 유망 기업을 육성\*\*하여 기술혁신과 산업 성장의 선순환 유도
  - \* (첨단공정 협력) PIM 예타사업 기반기술 과제총괄·협력 창구인 PIM HUB에 삼성·SK하이닉스가 기술자문위원으로 참석하고, 우수 R&D 결과물에 대해서는 공정 적용을 검토
  - \*\* ▲중소 팹리스 기업에 대한 기술이전 및 상용화 R&D 지원 ▲고가의 설계를 공동활용 지원 ▲창업초기 전문 엑셀러레이터와 연계를 통해 투자·용자 및 해외진출 지원
- ①공공 인프라 연계 현장중심 교육, ②연구 중심 특성화 대학원, ③실무형 학부과정 등 AI 반도체 특화 교육 신설·강화를 통한 5년간 7,000명 규모의 최고급 전문인력 양성
- ('22년) 「국산 인공지능반도체를 활용한 K-클라우드 추진방안」을 통해 ①국산 AI반도체 고도화, ②AI반도체용 SW 개발, ③데이터센터·AI서비스 실증, ④산·학·연 협력 강화 방안 제시
  - ※ '(2차) 인공지능 반도체 최고위 전략대화('22.12.12.)' 논의 사항에 대한 후속 조치
- (국산 AI반도체 고도화) ('23~'25) NPU → ('26~'28) 저전력 PIM(DRAM 기반) → ('29~'30) 극저전력 PIM 구현(비휘발성메모리(NVM) 기반)
- (AI반도체용 SW 개발) 신규 대형사업을 통해 초고속·극저전력 알고리즘 구현 SW 및 상용자원공유 적용을 위한 주변 기술개발 추진
- (데이터센터·서비스 실증) 인공지능 반도체 고도화 단계별로 데이터센터 적용, 인터넷 자원공유 기반 인공지능서비스\* 제공을 위한 실증사업\*\* 추진
  - \* 서버형 인공지능반도체(지능형 가전가구, 금융기술 등) 및 엣지형 인공지능반도체(자율차, XR 등)
  - \*\* (1단계) NPU Farm(팜) → (2단계) PIM Farm → (3단계) NVM-PIM Farm
  - ※ '국산 NPU 데이터센터 구축사업(신규 예정)' 및 '인공지능·인터넷기반자원공유 서비스 개발 사업'을 연계하여 '25년까지 3년간 약 1천억원 규모로 투자 예정
- (산·학·연 협력 강화) '국산 인공지능반도체 기반의 케이-클라우드 얼라이언스'를 구성하여 민·관 협업 창구 마련과 주요 과제 발굴 추진하고, 산·학 협력을 통해 산업 수요 맞춤형 설계역량을 갖춘 인력양성을 위해 '인공지능반도체 대학원' 신설('23년~, 3개교)


## 제5장 R&D 투자 동향

본 장에서는 국내 정부 R&D 투자동향을 파악하기 위해 ①NTIS 키워드 검색\*을 통한 R&D 과제\*\*와 ②인공지능 반도체 분야 주요 R&D 사업\*\*\* 분석을 수행함

\* (과제검색키워드) 인공지능 반도체, \*\* (분석 기준) 저자 판단에 따라 '인공지능 반도체' 분야 상관성에 따라 세부 분류 및 제외, \*\*\* (정부R&D사업) 유관 분야 세부사업 '19년~'23년 사업비(국회안) 기준(내역·내내역사업 제외)

### 5.1 정부R&D 과제 수행 현황 ※ 과제 결산 시점을 고려하여 최근 4년('18~'21)을 기준으로 분석

※ NTIS 과제 검색 결과를 저자가 임의 분류한 결과로 주관기관·부처의 의도와는 상이할 수 있음

 지난 4년('18년~'21년)간 인공지능 반도체 분야 정부R&D과제 총투자액은 3,380억 원 규모로 연평균 81.3% 투자 확대 중

- (주관부처) 4개 부처가 전체 투자의 약 99.2%를 차지하고 있으며, 이 중 과학기술정보통신부(이하 과기정통부)의 그간('18~'21) 투자 규모는 2,602억 원으로 전체의 77.0%를 차지(연평균 76.0% 투자 확대)

※ 과기정통부는 산·학·연 모든 수행 주체를 대상으로 기초·개발·응용연구 등 개발 단계 전반을 지원 중

- 뒤이어 산업통상자원부(이하 산업부)가 532억 원(15.7% 비중)을 차지하였으며, 주요 부처 중 가장 높은 연평균 투자 증가율(112.9%)을 보임

※ 산업부는 중소·중견기업을 중심으로 개발·응용연구를 지원 중

- 대학을 중심으로 개인·기초연구지원을 추진하는 교육부(3.7%)와 중소기업에 대상으로 개발 연구를 중점 지원하는 중소벤처기업부(2.8%)가 뒤를 이어 투자를 추진 중

※ 특히, 교육부는 개발연구보다는 기초·응용 연구의 비중을 강화하는 등 연평균 97.2%의 투자 확대를 통해 '21년 이후 개발연구 중심의 중소벤처기업부 투자액을 상회

〈표 15〉 인공지능 반도체 주관부처별 투자 동향('18~'21)

(단위 : 백만원)

부처명	'18년		'19년		'20년		'21년		총 투자액		연평균 증가율
	예산	비중	예산	비중	예산	비중	예산	비중	예산	비중	
과학기술정보통신부	23,016	84.8%	37,184	75.4%	74,637	74.7%	125,425	77.6%	260,263	77.0%	76.0%
산업통상자원부	2,435	9.0%	8,950	18.1%	18,286	18.3%	23,509	14.5%	53,180	15.7%	112.9%
중소벤처기업부	816	3.0%	1,666	3.4%	3,282	3.3%	3,765	2.3%	9,529	2.8%	66.5%
교육부	863	3.2%	1,172	2.4%	3,762	3.8%	6,618	4.1%	12,416	3.7%	97.2%
기타	-	-	373	0.8%	-	-	2,257	1.4%	2,630	0.8%	-
총 계	27,130	-	49,346	-	99,967	-	161,574	-	338,017	100.0%	81.3%

- (수행 주체) 지난 4년간 총 투자액을 바탕으로 산·학·연 과제 수행 주체 비중은 37.4%, 36.0%, 23.4%로 나타났으며, '20년부터 중견·대기업에 대한 투자 규모가 확대
  - 4년간 중소·중견기업에 대한 투자가 큰 폭으로 확대(연평균 121.0%, 144.7%)되고, '20년부터 대기업의 신규 참여가 이루어지는 등 산업계를 대상으로 한 활발한 정부 지원 추진 중
    - ※ 중소기업의 일부 기초연구를 제외하고 중소·중견·대기업은 대체로 개발·응용단계가 80% 이상을 차지
  - 대학은 기초연구를 중심(54.0%), 국공립·출연(연)은 응용연구를 중심(41.1%)으로 수행 중

〈표 16〉 인공지능 반도체 분야 수행주체별 정부R&amp;D과제 투자 동향('18~'21)

(단위 : 백만원)

수행주체	'18년		'19년		'20년		'21년		총 투자액		연평균 증가율
	예산	비중	예산	비중	예산	비중	예산	비중	예산	비중	
중소기업	4,432	16.3%	10,076	20.4%	33,696	33.7%	47,824	29.6%	96,027	28.4%	121.0%
중견기업	945	3.5%	1,653	3.3%	4,730	4.7%	13,843	8.6%	21,171	6.3%	144.7%
대기업	-	-	-	-	3,600	3.6%	5,600	3.5%	9,200	2.7%	-
대학	10,087	37.2%	15,242	30.9%	31,881	31.9%	64,508	39.9%	121,718	36.0%	85.6%
국공립·출연(연)	11,647	42.9%	18,214	36.9%	21,294	21.3%	28,040	17.4%	79,195	23.4%	34.0%
기타	20	0.1%	4,160	8.4%	4,765	4.8%	1,760	1.1%	10,705	3.2%	344.8%
총 계	27,130	-	49,346	-	99,967	-	161,574	-	338,017	100.0%	81.3%

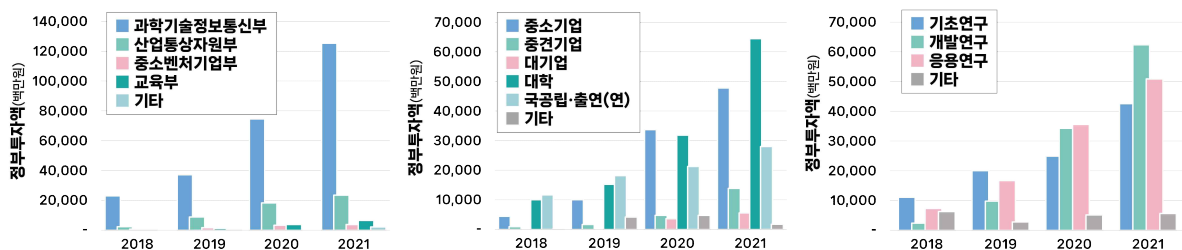
- (연구개발단계) '18년 기초연구 중심(40.8% 비중)에서 그간 개발·응용연구에 대한 지속적인 투자 확대로 누적 투자액 기준 64.9%(개발32.2%, 응용32.7%)를 차지

※ '19~'20년 이래로 상용화 중심의 기술개발·기업지원 등 인공지능 반도체 분야 정책 방향성과 부합

〈표 17〉 인공지능 반도체 분야 수행주체별 정부R&amp;D과제 투자 동향('18~'21)

(단위 : 백만원)

수행주체	'18년		'19년		'20년		'21년		총 투자액		연평균 증가율
	예산	비중	예산	비중	예산	비중	예산	비중	예산	비중	
기초연구	11,067	40.8%	20,063	40.7%	24,957	25.0%	42,627	26.4%	98,714	29.2%	56.8%
개발연구	2,430	9.0%	9,794	19.8%	34,290	34.3%	62,372	38.6%	108,887	32.2%	195.0%
응용연구	7,327	27.0%	16,691	33.8%	35,588	35.6%	50,903	31.5%	110,509	32.7%	90.8%
기타	6,306	23.2%	2,798	5.7%	5,132	5.1%	5,672	3.5%	19,907	5.9%	△3.5%
총 계	27,130	-	49,346	-	99,967	-	161,574	-	338,017	100.0%	81.3%



[그림 18] 왼쪽부터 ▲주관부처, ▲수행주체, ▲연구개발 단계별 투자 동향('18~'21)

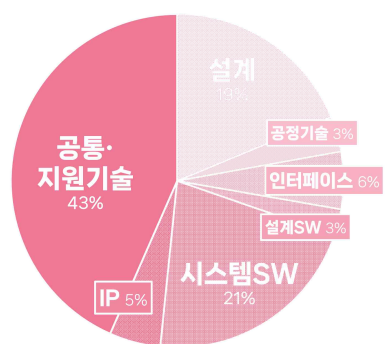
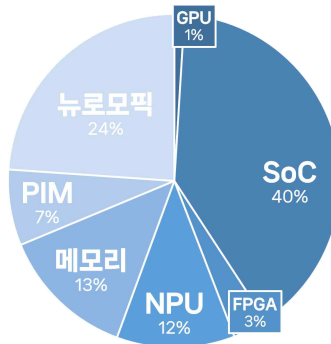
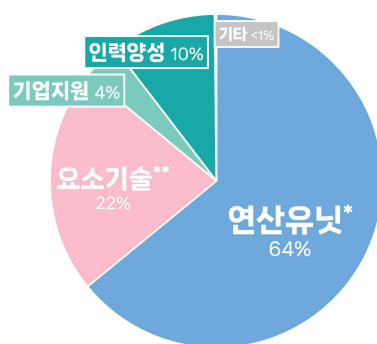
## 정부R&D과제는 대체로 연산유닛·요소기술 분야에 맞춰 투자 중이며, 일부 과제를 통해 기업지원·인력양성 등 생태계 강화를 위한 지원 중

- 연산유닛 구분에 따라 2세대 인공지능 반도체 기반의 상용화 분야와 뉴로모픽·PIM 반도체 등 차세대 기술 선점을 위한 R&D 투자 규모 지속 확대 중(연평균 94.4% 확대)
  - 단기적인 산업응용·활용을 위해 SoC\*, NPU\*\*, FPGA 등 2세대 인공지능 반도체 기반 개발·응용연구가 중점적으로 이루어지는 중(총투자액의 35.1% 비중을 차지)
    - \*이미지·침단 센서 등 IoT·제조·가전·바이오 등 산업 분야에 인공지능 접목을 위한 다양한 반도체 소자를 포함
    - \*\*ASIC 구조를 기반으로 데이터센터(서버)·모빌리티(자율주행) 및 인공지능을 접목한 차세대 산업 분야 활용을 위한 특수 목적형 기술개발 과제가 대다수이며, 모든 분야 중 가장 높은 연평균 증가율(419.4%)이 나타남
    - ※ 비록 소규모이나 GPU를 기반으로 한 산업응용(제조 분야 적용 등) 고도화 연구와 GPGPU·고성능·고신뢰성 인공지능 시스템 구현 연구가 지속적으로 이루어지는 중
  - 3세대 인공지능 반도체 분야로는 뉴로모픽 반도체가 '18년부터 비교적 높은 비중(21.9%('18) → 12.9%('21))으로 지속적인 지원(연평균 51.9% 투자 확대)이 이루어졌으며, PIM 반도체는 유관 분야 대형사업\* 추진 시점에 따라 '20년부터 본격적인 투자가 이루어진 것으로 나타남
    - \* 차세대지능형반도체기술개발사업('20~'29), 신개념PIM반도체선도기술개발('21~'24) 등
  - 한편, 국내 기업이 선도중인 메모리 반도체 분야 역시 ▲DRAM 등 기존 구조의 고도화, ▲차세대 메모리(MRAM·ReRAM 및 뉴로모픽向 구조 등) 요소기술 확보를 위한 투자 중
- 요소기술을 통해 ▲설계·공정기술, ▲시스템SW, ▲설계SW, ▲IP 등 인공지능 반도체 구현을 위한 주변기술을 지원 중이며, 그간 총투자액의 12.4%를 차지
  - ※ 본 고의 통계에서 '공정' 분야는 시스템반도체 분야 공통 기술개발이 다수 포함된 관계로 인공지능 반도체 분야 특화 지원 분야 집계에 다소 한계가 존재
  - 그간 설계 분야 지원 규모는 지속해서 확대되었으나, 연평균 증가율이 평균(81.3%)을 하회하는 수준(45.6%)으로 연간 투자액의 비중은 해마다 감소하는 추세\*로 나타남
    - \* 이는 설계·공정 분야가 주요 연산유닛(소자)별로 특성화·흡수되어 지원 중인 이유로도 해석이 가능
  - 연산유닛을 비롯하여 HW 시스템 작동을 위한 시스템SW 분야 R&D 지원(연평균 104.6% 확대)과 인공지능 반도체 설계SW와 IP 내재화를 위한 투자도 지속적으로 추진 중
    - ※ 한편, 본 고에서 언급한 RISC-V 관련 R&D는 '19년부터 일부 과제를 통해 추진중인 것으로 나타남
  - 요소기술 중 가장 높은 비중을 차지하는 공통·지원기술에는 차세대 인공지능 반도체 구현을 위한 주요 소재·부품·장비 개발을 포함
- 인공지능 반도체 분야 인력양성은 대체로 대학\*을 중심으로 이루어지고 있으며, 기업지원은 기 확보된 기술의 사업화·상용화\*\*를 중점 지원 중
  - \* '인공지능 반도체 융합전문인력육성사업', '인공지능시스템 융합 연구센터' 등
  - \*\* '차세대 반도체 기술개발사업 상용화 지원', '혁신형 창업과제 혁신성장 패키지' 등

〈표 18〉 인공지능 반도체 세부 분야별 정부R&amp;D과제 투자 동향('18~'21)

(단위 : 백만원)

세부 분야	'18년		'19년		'20년		'21년		총 투자액		연평균 증가율
	예산	비중	예산	비중	예산	비중	예산	비중	예산	비중	
<b>기술개발</b>	<b>23,350</b>	<b>86.1%</b>	<b>43,706</b>	<b>88.6%</b>	<b>84,416</b>	<b>84.4%</b>	<b>139,057</b>	<b>86.1%</b>	<b>290,529</b>	<b>86.0%</b>	<b>81.3%</b>
연산유닛	14,462	53.3%	29,928	60.6%	66,010	66.0%	106,251	65.8%	216,652	64.1%	94.4%
GPU	219	0.8%	335	0.7%	359	0.4%	1,210	0.7%	2,123	0.6%	76.7%
SoC	4,188	15.4%	10,583	21.4%	29,647	29.7%	41,770	25.9%	86,188	25.5%	115.2%
FPGA	1,446	5.3%	2,191	4.4%	1,740	1.7%	1,917	1.2%	7,294	2.2%	9.8%
NPU	99	0.4%	194	0.4%	10,707	10.7%	13,875	8.6%	24,875	7.4%	419.4%
메모리	2,572	9.5%	4,193	8.5%	8,777	8.8%	12,908	8.0%	28,450	8.4%	71.2%
PIM	-	-	85	0.2%	2,090	2.1%	13,759	8.5%	15,934	4.7%	-
뉴로모픽	5,938	21.9%	12,347	25.0%	12,690	12.7%	20,812	12.9%	51,788	15.3%	51.9%
요소기술	3,767	13.9%	6,852	13.9%	8,810	8.8%	22,383	13.9%	41,812	21.9%	81.1%
설계	1,958	7.2%	2,520	5.1%	3,340	3.3%	6,039	3.7%	13,857	4.1%	45.6%
공정기술	105	0.4%	280	0.6%	1,261	1.3%	878	0.5%	2,524	0.7%	103.0%
인터페이스	-	-	500	1.0%	1,800	1.8%	1,800	1.1%	4,100	1.2%	-
설계SW	-	-	-	-	-	-	1,895	1.2%	1,895	0.6%	-
시스템SW	1,201	4.4%	3,015	6.1%	1,261	1.3%	10,286	6.4%	15,763	4.7%	104.6%
IP	503	1.9%	537	1.1%	1,148	1.1%	1,485	0.9%	3,673	1.1%	43.5%
공통·지원기술	5,121	18.9%	6,925	14.0%	9,597	9.6%	10,422	6.5%	32,064	9.5%	26.7%
<b>기업지원</b>	<b>1,170</b>	<b>4.3%</b>	<b>1,950</b>	<b>4.0%</b>	<b>5,290</b>	<b>5.3%</b>	<b>3,775</b>	<b>2.3%</b>	<b>12,185</b>	<b>3.6%</b>	<b>47.8%</b>
<b>인력양성</b>	<b>2,500</b>	<b>9.2%</b>	<b>3,520</b>	<b>7.1%</b>	<b>10,171</b>	<b>10.2%</b>	<b>18,552</b>	<b>11.5%</b>	<b>34,743</b>	<b>10.3%</b>	<b>95.1%</b>
<b>기타(정책·전략)</b>	<b>110</b>	<b>0.4%</b>	<b>170</b>	<b>0.3%</b>	<b>90</b>	<b>0.1%</b>	<b>190</b>	<b>0.1%</b>	<b>560</b>	<b>0.2%</b>	<b>20.0%</b>
<b>총 계</b>	<b>27,130</b>	<b>-</b>	<b>49,346</b>	<b>-</b>	<b>99,967</b>	<b>-</b>	<b>161,574</b>	<b>-</b>	<b>338,017</b>	<b>100.0%</b>	<b>81.3%</b>



세부 분야별 투자 비중

\*연산유닛 유형별 투자 비중

\*\*요소기술별 세부 투자 비중

[그림 19] 연산유닛·요소기술 등 세부 분야별 투자 비중('18~'21 총투자액 기준)



그간 정부 R&D 과제 총투자액 중 과반(49.4%) 규모로 ▲데이터센터(서버), ▲엣지컴퓨팅, ▲산업응용 중심의 응용·활용을 위한 R&D 지원이 이루어짐

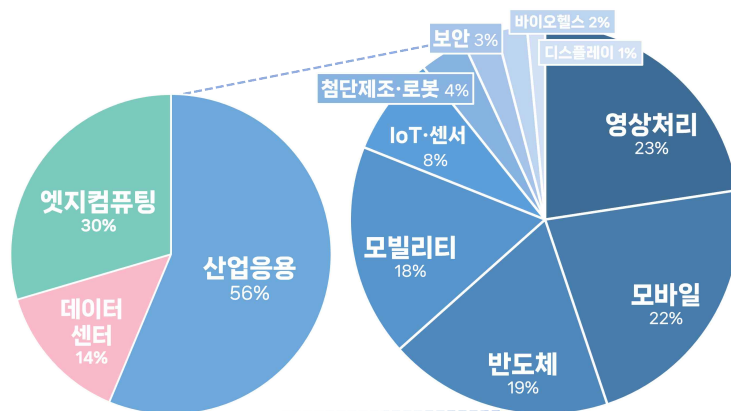
- 응용·활용 분야 중 데이터센터가 14% 수준을 차지하는 반면, 엣지컴퓨팅과 산업응용 분야가 각각 30%, 56%를 차지하며 기술·산업 동향과 부합하는 추세가 나타남
  - ※ 엣지컴퓨팅과 산업응용 분야를 명확히 구분하기 어려운 점이 존재하나, 본 고에서는 엣지 분야 공통 기술 분야는 '엣지컴퓨팅'으로, 응용분야(데이터센터 제외)가 명확한 과제는 '산업응용'으로 구분
  - 앞서 언급한 바와 같이 데이터센터(서버) 분야는 상용 CPU·GPU 제품군이 시장을 선도하는 만큼, 정부R&D 투자는 NPU·PIM·뉴로모픽 등 차세대 인공지능 반도체를 통한 ▲서버용 프로세서, ▲신개념 컴퓨팅 아키텍처(뉴로모픽, PIM 등) 등을 중점적으로 지원 중
  - 엣지컴퓨팅 분야 R&D과제는 대체로 산업 응용을 위한 목적 보다는 인공지능 반도체 구현을 위한 ▲회로·구조 설계 및 ▲시스템SW 등 기반기술 확보를 위한 지원 중
- 산업응용 분야는 세부적으로 9가지의 활용 목적(산업)으로 구분할 수 있으며, 영상처리 및 IoT·센서 등 분야는 타 산업과의 연계성을 포함
  - 그간 산업응용 분야 투자액 중 가장 높은 비중(23%)을 차지하는 '영상처리' 분야는 ▲사물·행동 인식, ▲영상 기록을 위한 지능형 카메라, ▲시각정보 처리를 위한 기반·응용 기술을 포함하고 있으며, 타 분야에 대한 폭넓은 확장성이 존재
    - ※ 영상처리 관련 응용·활용 기술 중 목적성이 뚜렷한 과제는 해당 산업으로 분류(예: 자율주행)
  - 자율주행기술을 필두로 '모빌리티' 산업 패러다임 변화로 인공지능 반도체 수요가 증가함에 따라 정부R&D투자 규모 역시 타 분야 대비 가장 높은 연평균 투자 증가세(167.5%)가 나타남
    - ※ 자율주행 레벨4급 기능안전성 구현, C-ITS용 영상기반 다중 객체 검출 NPU 개발, 인공지능 기반 차량용 통신 기술 향상을 위한 NPU 등 시스템 개발 등 추진 중
  - 그간 산업응용 분야에서 모바일\*, 반도체\*\* 분야의 투자 비중은 각각 22%, 19% 수준으로 상위권을 유지하였으나, 점차 투자 규모가 둔화 중인 것으로 나타나 정부 지원보다는 민간 중심의 자체 개발이 이루어지고 있는 것으로 보임
    - \* 모바일 프로세서(NPU)뿐 아니라 인터페이스, 시스템SW, 모바일용 뉴로모픽 시스템 기술개발
    - \*\* 반도체 공정의 예지·예측 및 인공지능 기술을 도입한 장비 개발을 위한 연산유닛·SW 개발
  - 인공지능 반도체를 도입한 IoT·센서\*와 첨단제조·로봇\*\* 산업 분야 과제 지원의 비중은 작은 편(응용·활용 분야 내 8%, 4% 차지)이나, '20년을 기점으로 큰 폭으로 증가하는 추세로 나타남
    - \* 단기적으로는 스마트가전에 적용할 수 있는 지능형 소자 기술개발이 이루어지고 있으며, 중장기적으로 제조·로봇분야에 적용할 수 있는 지능형 센서 분야도 포함
    - \*\* 완전 자동화, 협업시스템 구축, 디지털트윈과의 접목 등 기존 제조 환경의 지능화를 위한 기술개발 추진 중



〈표 19〉 인공지능 반도체 응용·활용분야별 정부R&amp;D과제 투자 동향('18~'21)


(단위 : 백만원)

구 분	'18년		'19년		'20년		'21년		총 투자액		연평균 증가율
	예산	비중	예산	비중	예산	비중	예산	비중	예산	비중	
<b>응용·활용</b>	<b>12,961</b>	<b>47.8%</b>	<b>23,201</b>	<b>47.1%</b>	<b>52,553</b>	<b>52.6%</b>	<b>78,040</b>	<b>48.3%</b>	<b>166,755</b>	<b>49.4%</b>	<b>81.9%</b>
데이터센터	1,209	4.5%	2,213	4.5%	8,530	8.5%	11,642	7.2%	23,594	7.0%	112.8%
엣지컴퓨팅	3,275	12.1%	2,845	5.8%	16,153	16.2%	27,026	16.7%	49,298	14.6%	102.1%
산업응용	8,477	31.2%	18,143	36.8%	27,870	27.9%	39,372	24.4%	93,863	27.8%	66.8%
영상처리	1,665	6.1%	4,091	8.3%	6,130	6.1%	9,310	5.8%	21,194	6.3%	77.5%
모바일	2,849	10.5%	3,016	6.1%	7,579	7.6%	7,447	4.6%	20,890	6.2%	37.8%
반도체	2,103	7.8%	3,730	7.6%	4,855	4.9%	6,746	4.2%	17,434	5.2%	47.5%
모빌리티	300	1.1%	5,782	11.7%	4,738	4.7%	5,740	3.6%	16,559	4.9%	167.5%
IoT·센서	-	-	-	-	1,775	1.8%	5,943	3.7%	7,718	2.3%	-
첨단제조·로봇	138	0.5%	224	0.5%	1,238	1.2%	2,014	1.2%	3,614	1.1%	144.4%
보안	441	1.6%	777	1.6%	744	0.7%	717	0.4%	2,679	0.8%	17.6%
바이오헬스	982	3.6%	117	0.2%	263	0.3%	1,006	0.6%	2,367	0.7%	0.8%
디스플레이	-	-	408	0.8%	550	0.6%	450	0.3%	1,408	0.4%	-
<b>기타</b>	<b>14,170</b>	<b>52.2%</b>	<b>26,145</b>	<b>53.0%</b>	<b>47,414</b>	<b>47.4%</b>	<b>83,534</b>	<b>51.7%</b>	<b>171,263</b>	<b>50.7%</b>	<b>80.6%</b>
공통기술	11,560	42.6%	23,305	47.2%	39,570	39.6%	68,916	42.7%	143,351	42.4%	81.3%
기타(인력양성 등)	2,500	9.2%	2,670	5.4%	7,754	7.8%	14,428	8.9%	27,352	8.1%	79.4%
정책연구	110	0.4%	170	0.3%	90	0.1%	190	0.1%	560	0.2%	20.0%
<b>총 계</b>	<b>27,130</b>	<b>-</b>	<b>49,346</b>	<b>-</b>	<b>99,967</b>	<b>-</b>	<b>161,574</b>	<b>-</b>	<b>338,017</b>	<b>100.0%</b>	<b>81.3%</b>



응용분야별 비중(기타 제외)      산업응용 분야 세부 비중  
 [그림 20] 응용·활용분야별 세부 투자 비중('18~'21 총 투자액 기준)

## 5.2 정부R&D 사업 투자 동향 ※ 사업비 확정 시점을 고려하여 최근 5년('19~'23)을 기준으로 분석

 최근 5년간('19~'23) 인공지능 반도체 분야 정부R&D사업 총투자액은 6,976억 원 수준으로 연평균 74.2%로 투자 확대 중

※ 인공지능 반도체는 시스템반도체 분야 정부R&D투자·지원의 일부로써 인력양성, 인프라 사업을 공유하나 본 고에서는 인공지능 반도체 개발을 주요 목적으로 하는 R&D 사업을 중심으로 분석

- 인공지능 반도체 분야는 크게 기술개발과 기업지원을 중심으로 지원 중이며, '20년 이후 신규 예타 사업 추진에 따라 지원 규모가 큰 폭으로 증가('19~'21년간 매년 약 3배 수준으로 투자 확대)
  - '19년도 이후 인공지능 반도체 관련 정부 R&D 투자액은 매년 증가(연평균 74.2%)하고 있으며, '23년 2,408억 원(국회안 기준)으로 최대 규모
    - ※ '19년부터 매년 발표한 정부의 반도체 산업 육성 방향이 '인공지능 반도체'를 중심으로 한 시스템 반도체 산업생태계 육성인 점을 고려하였을 때, 정책적 방향성과도 일치도가 높음
  - 정부는 R&D사업을 통해 단기적으로는 상용화 지원과 산업생태계 조성, 중장기적인 투자를 통한 뉴로모픽, PIM 반도체 등 차세대 기술 선점하는 Two-Track 전략 추진 중

〈표 20〉 인공지능 반도체 분야 주요 R&D사업 투자 동향('19~'23년)

(단위 : 백만원)

지원분야	연차별 정부R&D사업 투자규모					총 투자액		연평균 증가율
	'19년	'20년	'21년	'22년	'23년	예산	비중	
기술개발	15,000	57,448	137,306	209,151	218,665	637,570	91.4%	70.9%
기업지원	-	1,847	12,595	23,447	22,160	60,049	8.6%	86.1%
총 계	15,000	59,295	149,901	232,598	240,825	697,619	100.0%	74.2%

- (기술개발) 다부처 대형 예타사업을 중심으로 뉴로모픽·PIM 반도체 등 차세대 인공지능 반도체 구현을 위한 설계·소자 제조 분야별 기술개발 지원
  - 과기정통부·산업부가 다부처 예타사업으로 추진 중인 '차세대지능형반도체기술개발사업'을 통해 차세대 지능형 반도체 소자 원천기술의 선제적인 확보, 설계 및 미세화 제조·장비 융합 연구 등 반도체 연구 전주기 지원
    - ※ (소자) 신소자 원천기술(초저전력·고성능 신소자 개발), 조기 상용화 가능 신소자 개발, IP확보를 위한 웨이퍼 레벨의 집적·검증 기술 및 소자 기초기술 개발 지원(신소재·신공정)
    - (설계) 인공지능 연산에 최적화된 인공지능 프로세서(고성능·저전력), 초고속 인터페이스, 시스템SW 등 핵심 설계기술 개발 지원
    - (제조) 주력산업과 연계한 상용화 중심 시스템반도체 개발 지원과 차세대 반도체 제조에 필요한 공정·장비 기술개발 지원

- 또한, 과기정통부·산업부 다부처 예타사업인 ‘PIM인공지능반도체핵심기술개발사업’은 세계 최선도국인 우리나라 메모리 반도체 기술과 프로세서 기능을 융합한 PIM 반도체 기술개발을 통해 차세대 인공지능 반도체 초격차 기술 확보 추진
    - ※ (소자) 신소재, 집적공정 기술개발 등 PIM소자 원천기술 확보
      - (설계) 프로세서·로직과 메모리를 융합한 PIM반도체 개발 및 성능검증 용 칩 제작, 시스템 SW 최적화, 아키텍처·인터페이스 등 기반 기술 지원
      - (제조) PIM 반도체 제조를 위한 국내 소재·부품·장비 기술지원과 국내 팹리스 기업 경쟁력 강화, 수요기업 연계를 통한 상용화 지원
  - ’23년 신규 사업 추진을 통해 NPU-PIM 기반 차세대 데이터센터 컴퓨팅 플랫폼 개발\*과 엣지컴퓨팅에 특화된 인공지능 반도체 및 SW 특화 기술 확보\*\* 등 지원 강화
    - \*거대인공신경망인공지능반도체SW기술개발사업, \*\*인공지능반도체SW통합플랫폼기술개발사업
  - 이 외에도 통신·헬스케어 분야 엣지디바이스 구현, 자율주행용 인공지능 반도체 개발을 위한 인공지능 반도체 분야 산업응용·활용을 목적으로 주요 세부사업 추진 중
    - ※ 스마트엣지디바이스기술개발사업, 자율주행용인공지능반도체핵심기술개발 등
- (기업지원) 기존 대규모 설비 투자 중심 위주였던 메모리 반도체와 달리 인공지능 반도체는 기술집약적 산업으로 전문 설계역량과 IP를 보유한 팹리스 기업을 중심으로 지원
    - 유망 팹리스 기업 선정하여 인공지능 반도체 개발 전 주기 과정(혁신기술개발-설계-SW 구현·시제품제작-테스트·상용화)을 맞춤형으로 지원
      - ※ 인공지능반도체응용기술개발사업(’20~’24), 인공지능반도체혁신기업집중육성(’21~’24) 등
  - (인력양성) 인공지능 반도체 분야 전문인력 양성을 위한 맞춤형(전용) 사업은 아직 존재하지 않지만, 유관 예타 사업 내에서 인력양성을 지원하거나 시스템반도체 인력양성 사업 내에 이를 포함하는 형태로 지원 중
    - ‘PIM인공지능반도체핵심기술개발사업’은 PIM통합 인력양성 센터 구축을 통해 실무 연계 교육으로 PIM 인공지능 반도체 전문설계 인력양성을 지원
    - ‘민간공동투자반도체고급인력양성사업’은 인공지능 반도체 산업계 수요를 반영하여 첨단 기반기술 R&D를 통해 핵심기술과 고급전문연구인력 양성 확보를 동시에 지원
    - ‘국가반도체연구실(NSL)핵심기술개발사업’은 로직, 메모리 및 공통기반 기술 반도체 분야 핵심 원천기술 확보와 인력양성 기초단위인 반도체 연구실(Lab)을 지원
    - 한편, ‘시스템반도체융합전문인력육성사업’을 통해 인공지능, IoT·가전, 바이오헬스 등 유망 신산업 분야의 차세대 시스템반도체 제품개발 및 시장 선도를 위한 융합전문인력육성 지원


〈표 21〉 인공지능 반도체 주요 R&amp;D 사업

(단위 : 백만원)


분야		부처명	사업명	기간	사업내용	예산	
						'22년	'23년
기술 개발	대형 사업 (예타)	다부처 (과기정통부·산업부)	차세대지능형반도체 기술개발사업 (소자·설계·제조)	'20~'29	저전력 신소자, 인공지능 반도체 설계, 시스템반도체 설계 및 장비·공정 기술개발	145,469	134,837
		다부처 (과기정통부·산업부)	PIM인공지능반도체 핵심기술개발사업 (소자·설계·제조)	'22~'28	PIM 반도체 설계기술 및 PIM 특화 신소자 및 장비·공정 핵심 선도기술 개발 확보	50,894	57,935
	기반 기술	과기 정통부	신개념PIM반도체 선도기술개발사업	'21~'24	PIM 반도체 설계 자산(IP) 및 시제품, 반도체 구동을 위한 SW 등 확보	10,985	8,846
			거대인공신경망인공지능 반도체SW기술개발사업	'23~'27	AI반도체(NPU, PIM)을 기반으로 한 거대인공신경망 처리 컴퓨팅 구현 기술 확보	-	4,000
			인공지능반도체SW 통합플랫폼기술개발사업	'23~'27	엣지 AI반도체 상용화 촉진을 위한 특화된 SW 통합플랫폼 개발을 통해 엣지컴퓨팅 SW 통합플랫폼 및 공개 가능 모듈형 인공지능 반도체 확보	-	5,100
	응용· 활용	과기 정통부	스마트엣지디바이스 기술개발사업	'22~'26	DNA(데이터·5G·AI 등) 생태계 조성 및 국산 인공지능 반도체 활성화를 위한 5G 융합 맞춤형 스마트 엣지 디바이스 기술개발	4,270	5,700
			자율주행용인공지능 반도체핵심기술개발	'22~'25	레벨 4 이상을 지원하는 자율주행용 인공지능반도체 기술 확보 및 모빌리티 적용을 위한 통합 모듈 개발	7,800	10,400
	기업 지원	과기 정통부	인공지능반도체 응용기술개발사업	'20~'24	반도체 설계 팹리스 기업 중심으로 산·학·연 협력을 통한 인공지능반도체 조기 상용화 및 응용기술개발 지원	12,900	12,000
			인공지능반도체 혁신기업집중육성	'21~'24	유망기업(팹리스, IP기업 등)의 연구 개발 및 기술 사업화에 대한 맞춤형 집중 지원을 통한 '인공지능 반도체 혁신기업' 육성	10,547	10,160
	인력 양성	산업부	민관공통투자 반도체 고급인력양성사업(예타)	'23~'32	기업수요형 R&D 수행으로 검증된 산업형 R&D 고급전문인력 양성을 통한 반도체 산업 글로벌 경쟁력 확보 및 지속 가능발전 도모	-	10,046
과기 정통부		국가반도체연구실 (NSL)핵심기술개발	'23~'28	연구개발 및 인력양성의 기초단위인 반도체 연구실(Lab)의 역량강화를 위한 연구개발 지원	-	6,475	

## 제6장 결론 및 제언

### 6.1 요약 및 결론

 인공지능 반도체는 최근 미세화 공정의 한계, 최근 메모리 반도체 다운 사이클 등 위축된 반도체 산업 전반에 새로운 기회를 창출하며 큰 폭의 성장세가 전망

- 지금까지 데이터센터(서버) 중심으로 초기 시장이 형성되었으나, 기술개발 동향을 고려하였을 때 모바일·모빌리티를 포함한 에너지·바이오 등 엣지컴퓨팅 분야에 대한 활용 확대가 전망
  - 제조·통신·헬스케어 등 산업 전반에 대한 인공지능 역할 증가는 종단 단말(엣지) 기기에 대한 인공지능 반도체 적용과 직결되며 유관 분야 수요가 폭증할 것으로 전망
    - ※ 인공지능 연산력 고도화 → 인공지능 반도체 수요 → 파운드리 시장 촉진 → 저전력·고밀도 집적을 위한 미세 공정의 수요 증대 → 유관 분야 지원기술(소·부·장) 시장 성장
- 반도체 산업 생태계에서 인공지능 반도체 분야는 ‘팹리스 중심’ 산업구조를 가지며, 기존 CPU(x86), GPU, 메모리 반도체와 달리 목적에 따른 맞춤형 다품종 주문생산이 가능
  - 데이터센터용 CPU+GPU는 절대적인 영향력에 따라 사실상 대체가 불가능한 품목\*이나, 모바일·자율주행 등 엣지컴퓨팅 분야는 기존 빅테크기업\*\*에서부터 중소·중견 팹리스 난립·경쟁 심화
    - \* 다만, 데이터센터용 인공지능 반도체 역시 기존 x86 CPU와 NVIDIA GPU를 대체하기 위한 맞춤형 반도체 기술·제품 개발이 지속 중
    - \*\* Apple, Google, Tesla 등 자사 인공지능 기반 제품·서비스의 효과적인 제공을 위해 자체적인 모바일AP, NPU 등을 개발

 인공지능 반도체 시장의 급격한 확대와 함께 ▲ASIC의 약진 ▲RISC-V의 부상 ▲주변 기술(패키징 공정 등) 고도화 ▲뉴로모픽·PIM 개발 등이 주요 이슈로 부상

- 인공지능 반도체는 데이터센터용 GPU를 중심으로 적용 범위를 확장하는 중이며, 빅테크·선도기업들은 자사 제품·서비스에 특화된 ASIC 기반 인공지능 반도체(NPU 등) 개발에 박차
  - 데이터센터용 인공지능 반도체는 여전히 Intel社 CPU(x86)와 NVIDIA社 GPU가 높은 범용성과 학습·추론 능력으로 세계시장에서 독점적인 지위를 차지
  - 다만, 현재 엣지컴퓨팅 분야에서 활발한 상용화가 이루어지고 있는 ASIC 기반 인공지능 반도체는 강력한 수요 증가와 기술 발전에 따라 데이터센터 분야까지도 주도권 확대가 전망

※ 현재 ASIC은 모바일 AP, NPU, 자율주행용으로 널리 활용 중이며, 이미 아마존은 자사 서버(AWS)에 ARM CPU를 도입하여 기존 x86(Intel) 대비 우수한 가성비를 확보한 사례가 존재


- 설계 분야에서는 오픈소스 설계자산인 RISC-V의 점유 확대가 전망되며, 고효율·고성능·고용량 인공지능 연산 구현을 위해 주변기술(패키징·인터페이스 등) 동반 고도화가 요구
  - RISC-V는 그간 선도기업(Intel, AMD, ARM 등)의 독점적인 코어 시장 지배구조를 흔들며, 근 미래에 산업 내 주력 코어 중 하나로 정착할 것으로 전망
  - ※ 오픈소스·비정칙성·저전력 구현의 경쟁력을 바탕으로 그간 IoT, 엣지컴퓨팅 일부 분야에 대한 적용에서 최근 자율차·데이터센터·모바일(랩탑) 등으로 적용 산업 확장 중
- 기존 컴퓨팅 구조의 한계로 연산유닛과 메모리 간의 데이터 병목과 저전력화의 해결 방안으로 뉴로모픽 반도체, PIM 반도체 등이 활발히 연구 중이나, 단기간 내 상용화는 요원
  - 뉴로모픽 반도체는 SNN에 기반한 학습·추론 알고리즘 개발 중으로 PRAM, MRAM, RRAM 등 신경 전달 함수 모사를 위한 메모리 소자 연구 역시 진행 중
  - 한편, 해외 선도기업은 서버용 PIM 솔루션을 사업화 중이며, 국내 대기업 2개社は 각각 PIM 반도체 제품을 발표

### 주요국은 공급망 내재화, 기술패권경쟁 등 대외환경 변화와 인공지능 반도체의 산업·사회적 파급력을 고려해 생태계 강화와 기술 확보를 위한 대대적인 지원을 추진 중

- 사회 전 분야의 디지털전환 가속화로 인공지능·반도체 활용·확산으로 무역·통상 품목에서 전략자산 관점으로 인식이 급변하며 주요국 간 기술·통상 경쟁 심화
  - 이에 따라 지난 수년간 반도체 산업을 중심으로 미-중 패권경쟁이 지속 중이며, 최근 미국은 제재 수단 확대(일부 그래픽카드, 설계SW, 제조장비 등)를 통해 중국의 인공지능 기술 확보를 저지
- 미국을 필두로 대만, 유럽 등 주요국은 이른바 「반도체법」을 마련함으로써 국가적 차원에서 반도체 분야 ▲기술개발 ▲생산역량 확충 ▲인력양성 등 대대적인 지원 중
  - 주요국은 기존 지원 정책·전략을 넘어 <sup>(미국)</sup>「반도체과학법(‘22.8.)」을 필두로 <sup>(대만)</sup>「산업혁신 조례 수정안(‘22.11.)」, <sup>(EU)</sup>「반도체법(‘22.12.)」 등 법 제정을 통한 상위 지원 근거를 마련
  - 특히, 시스템반도체 분야를 바탕으로 인공지능 반도체 분야의 ▲설계(인터페이스·아키텍처 등) ▲제조역량 ▲첨단 패키징 ▲엣지컴퓨팅 분야 응용·활용성 제고 등에 주안점
- 우리 정부 역시 '20년 이래로 인공지능 반도체 분야를 중심으로 국가적 차원의 기술개발, 산업경쟁력 확보를 위한 지원 정책을 발표 중
  - 특히, 산업생태계 조성을 통한 민간 경쟁력 확보를 위해 ▲융합얼라이언스 확대·운영 ▲상용화·수요 연계 실증 추진 ▲공공·민간 적용 확산(NPU Farm 구축 등) 등 지속적인 지원전략 제시



## 6.2 제언

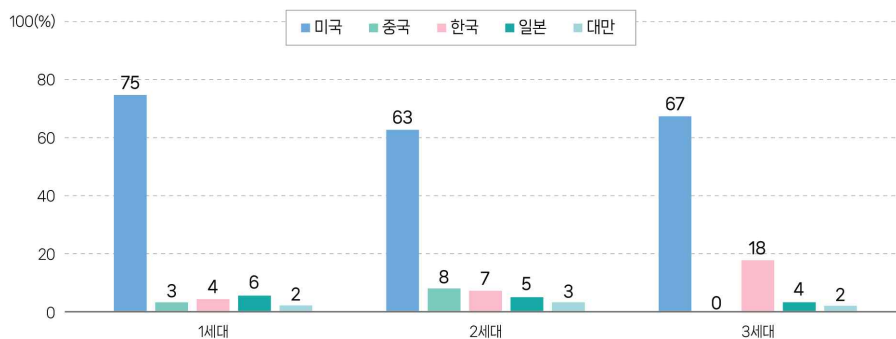
 인공지능 반도체는 우리나라가 취약한 시스템반도체 분야와 연관성이 높아 생태계 조성 및 차세대 기술 선점을 위한 전략성 강화와 중장기적인 지원이 필요

- NPU 등 ASIC 구조를 중심으로 급성장 중인 인공지능 반도체 산업에 대응하고 시장을 선도하기 위해 메모리 대기업 중심의 국내 반도체 산업 구조의 체질 개편이 절실
  - 국내 반도체 분야는 국내 수출액의 약 20%를 차지('21)하는 주력 품목이나, 인공지능 반도체를 비롯한 시스템반도체 분야 전반의 글로벌 경쟁력은 여전히 열세
    - ※ 메모리 반도체는 국내 대형 IDM 2개사가 세계시장의 60~70%를 차지하고 있으나, 시스템 반도체 시장의 점유율은 3% 수준으로 중국·대만에 비해서도 월등히 낮은 수준
  - 향후 인공지능 반도체 산업에 대한 중요성 등 인식에 따라 '19년 이래로 유관 분야의 정부 지원 정책이 매년 발표 중이나, 아직 실효성 있는 결과 확보에는 미흡한 실정
    - ※ 팹리스 분야의 세계 시장 점유율은 1%대에 불과하며, 파운드리 분야는 세계 2위이나 점유율 면에서 1위인 TSMC(53.4%) 대비 열세(16.4%)로 지속적인 하락세가 전망됨(출처: IC Insights(2021), TrendForce(2022))
- 글로벌 반도체 공급망 이슈에 관한 대처와 반도체 산업의 재도약을 위한 국가적 육성과 지원을 위해 통합적 거버넌스 구축과 추진체계 개편 방안 모색이 필요
  - ※ 최근 정부의 「인공지능(AI) 반도체 산업 성장 지원대책」 등 다양한 지원책이 추진 중이나, 데이터센터용 인공지능 반도체 대표적인 스타트업 기업 Graphcore 6.8억달러, SambaNova 20억달러 투자 규모를 감안하면 지원 규모 재고 역시 필요
  - 최근 세계적인 반도체 공급망 내재화·재편 움직임에 따라 유럽, 미국, 중국 등 반도체 제조의 자립화 의지, 시장 지배력 확보를 위한 치열한 경쟁이 예상
    - ※ 반도체는 재료, 설계, 장비, 소자 기술 등이 융합된 산업으로, 장비 및 설계 등 국내 산업 환경은 외산에 점유된 상태
  - 정부 연구개발 구조상 반도체 분야 지원을 위해 과기정통부·산업부 등 주요 부처별 역할이 분담되어있어 대외환경 변화에 따른 기민한 대응에 한계 가능성 존재
  - 따라서 인공지능 반도체 시장 선점을 위한 국가 차원의 장기적인 기술개발 로드맵을 마련하고, 신기술 R&D 투자 및 반도체 기업 생태계 간의 가치사슬 공공 활용이 가능한 인프라 구축 등을 위한 통합된 거버넌스 확립 필요
    - ※ 「양자기술개발지원반(과)」가 유사 사례로, 과기정통부 내 1, 2차관의 양자 분야 사업을 총괄적으로 관리·기획하고, 효율적으로 운영을 위해 과기정통부장관 직속 조직으로 신설·운영('22.~)

- 인공지능 반도체는 여전히 발전 가능성이 높은 초기 시장 단계로, 국내 시스템반도체 산업의 경쟁력 강화를 위한 기회로 활용 가능
  - ※ 다만, CPU·GPU 등 데이터센터용 인공지능 반도체는 메모리와 같이 기존 IDM(Intel, NVIDIA)의 자체적인 설계·생산 품목으로 후발주자로서 시장 진입의 여지가 다소 낮음
- 인공지능 반도체 분야는 시스템반도체 산업의 일부로 설계·생산의 분업화가 일반적이며, 응용·활용 다변화로 다품종 수요가 확대 중인 만큼 중소·중견 팹리스와 스타트업의 시장 진입을 위한 실효성 있는 전략 마련 필요
  - ※ 전 산업의 인공지능 확산에 따른 데이터센터의 부담 가중으로, 이를 보완하기 위해 팹리스·스타트업에 진입이 유리한 엣지컴퓨팅 분야의 기술개발 필요성과 수요가 증가하는 추세
- 다양한 산업 분야에 고른 투자 전략보다는 엣지컴퓨팅 유망 분야 특화된 제품개발을 위해 수요기업·팹리스·데이터 전문기업·인공지능 전문가들이 참여한 이니셔티브·컨소시엄 등 협의체 구성을 통해 실효성 있는 지원전략 수립이 필요
  - ※ 국내 인공지능 반도체 기업들은 R&D 정책지원에 힘입어 팹리스-수요기업간 연계 활성화 중이나, 여전히 매출로 이어지지 못한 실정으로, 모바일·자율주행 분야 외에도 인공지능 활용이 전망되는 유망 분야(바이오헬스, UAM, 자율운행선박 등) 선점 노력 강화 필요

## 국내 반도체 기술 역량을 바탕으로 초기 시장이 형성 중인 인공지능 반도체 산업 선점을 위한 중장기적인 지원·노력과 함께 기술 종속 우려 해소를 위한 설계 분야 집중 투자 필요

- 차세대 인공지능 반도체로 불리는 뉴로모픽·PIM 반도체는 구조적 패러다임 변화가 필요한 분야로, 기존 메모리 분야에 경쟁력을 갖는 우리 산업 특성에서 긍정적인 요소로 작용
  - ※ 특히, PIM 반도체는 메모리 반도체 기술력에 기반하여 단기적으로는 기존 GPU·HBM에 활용이 예상되며 중장기적으로는 인공지능 연산유닛 전반에 활용도가 확장되며 지속적인 수요 증가가 전망
- 특허청, 경제추격연구소에서 발표한 연구 결과에 따르면 우리나라는 기존 1~2세대 대비 3세대 인공지능 반도체 관련 특허 신청 점유율이 세계 2위 수준으로 기술 선점 측면에서 우세



[그림 21] 인공지능 반도체 분야 국가별 특허 신청 점유 비중('00.~'21.)

\* 출처: 특허청 재구성

- 우리 정부는 대형사업\* 추진을 통해 고성능·저전력 NPU와 PIM 반도체의 독자적 기술력 확보를 위한 기술개발과 반도체 설계 관련 전문인력 양성을 위한 대대적인 지원을 추진 중
  - \* 차세대지능형반도체기술개발('20~'29), 총사업비 1조 96억원), PIM인공지능반도체핵심기술개발('22~'28, 총사업비 4,027억원)
  - ※ PIM 성능 구현에는 다양한 AI 연산 알고리즘에 대응할 수 있는 ISA의 컴파일러 구현과 함께 PIM 구조의 유용성 향상을 위한 추가적인 컴파일러 병행 개발이 필요
- 또한, PIM 반도체는 국내 대형 IDM의 우수한 메모리 기술력을 바탕으로 조기 상용화가 이루어질 수 있도록 민-관 기술 교류 및 사업·정책 기획 시 민간 수요를 반영 필요
- 다만, 뉴로모픽 분야는 여전히 초기 개발 단계로 전용 사업이 미비하며, 과제 단위에서 수행 중인 것으로 나타나 개발 능력을 갖춘 산·학·연을 연계하여 조기 상용화를 위한 지원 필요
- NPU를 비롯해 최근 약진을 거듭 중인 ASIC 기반 인공지능 반도체는 설계기술이 핵심\*으로 기존 ARM 등 주요 설계자산에 대한 종속성 해소를 위한 투자 강화 필요
  - \* CPU, GPU 등 1세대 인공지능 반도체는 일반적으로 대형 IDM에서 설계·제조가 이루어지나, 특수 목적형의 FPGA, ASIC 및 향후 뉴로모픽 반도체는 설계·생산 분화가 일반화
- 국내 반도체 산업은 세계적인 메모리·파운드리 경쟁력을 보유하고 있어 독자적인 설계 능력 확보에 따라 인공지능 반도체 분야 경쟁력이 좌우
- 최근 주목받는 RISC-V는 오픈소스의 특성상 M&A 등 경영권에 영향을 받지 않아 기존 시장을 점유하는 ARM의 강력한 대항마로 부상 중이며, 이미 국내외 팹리스·스타트업에서 적극적으로 도입 중
- 정부 차원에서도 RISC-V와 같은 대체제를 기반으로 반도체 설계 분야 ▲교육·훈련 ▲기반 조성 ▲기업지원 ▲Top-Down형 과제 추진 등 지원방안을 강화가 필요
  - ※ 일례로 인도 정부가 RISC-V의 프리미어 멤버(연간 25만불 회비 납부)인 점 등을 고려하여, 우리 정부 역시 RISC-V에 대한 R&D와 관련 인프라 조성에 대한 적극적인 검토가 필요
- 첨단 패키징·인터페이스 등 연산유닛을 보완할 수 있는 지원기술 고도화 필요성이 나날이 증가하는 추세로 정부 지원을 통한 민간 참여·교류 유도과 차세대 기술 확보가 절실
- 국내 우수 후공정 기업(OSAT)이 존재하는 만큼 정부 지원을 통해 팹리스와의 연계 및 차세대 기술개발을 위한 투자 확대가 필요
  - ※ 후공정 분야 기술개발을 위한 정부 R&D 세부사업은 전무한 실정으로 중소기업 중심의 개발연구가 과반 수준을 차지하는 것으로 나타남(KISTEP, 2020)
- 한편, 반도체 미세화에 따라 급증하는 설계·제조 비용에 대응하여 ▲민-관 협력, ▲팹리스-파운드리-수요기업 연계 및 다양한 형태의 ▲비R&D적 지원방안 등 모색 필요

- 공정 미세화에 따라 칩 제조 비용과 함께 초미세공정 반도체 설계에 EDA 등 SW 라이선스 비용 및 설계자산(IP) 및 20~40억 수준의 시제품 개발 비용(MPW)은 사업성이 확보되지 않은 이상 중소기업에 큰 부담
  - ※ 최신 통계에 따르면 28나노 공정 설계비용은 4,000만 달러 대비, 7나노 2.2억 달러, 5나노 4.2억 달러, 3나노 공정설계비용은 최대 5.9억 달러에 육박(IBS, 2022)
- 따라서 정부는 기업 간 협업을 강화하기 위한 정책·제도·자금 등 비R&D적 지원을 강화하고, 민간은 공급망 다변화와 개발·생산 역량을 강화하는 등 협력·역할 분담 추진
  - ※ 한편, 증가하는 개발비용 상쇄를 위해 다소 제한적인 국내 시장을 넘어 글로벌 시장을 목표로 제품개발 및 매출 확보를 위해 판로개척 등 기술개발 외적인 지원을 병행
- 또한, 정부는 초기 창업, 스케일업, 중소·중견·대기업 등 기업 규모별 맞춤형 협력 모델을 구축하여 생태계 다양성과 지속가능성을 확보 방안 마련
  - ※ 그간 반도체 소자 개발에 집중된 주요 펀드에서 유관 산업생태계의 동반성장을 위한 펀드로 개편하는 등 체질 개선을 위한 지렛대로 활용

## 인공지능 반도체를 비롯한 시스템반도체 산업의 핵심은 ‘설계’이며, 이를 위한 실무·고급인력의 확보가 필수적이나, 산업계는 지속해서 인력난을 호소

- 그간 정부의 반도체 분야 지원 확대에도 불구하고, 산업인력 수요 대비 맞춤형 전문인력 양성 기반은 여전히 미흡한 실정
  - ※ 반도체 산업 규모 확대에 따라, 산업인력은 현 17.7만명 수준에서 10년 후 약 30.4만 명까지 증가할 것으로 전망됨(한국반도체산업협회, 2022)
- 따라서 반도체 학과를 통한 전문인력 집중 육성, 기존 유관 학과에 대한 반도체 분야 트랙 신설·강화를 통해 산업 수요에 유연하게 대응
  - ※ 한편, 졸업자 및 타분야 산업인력을 대상으로 반도체 분야 재교육 기회를 제공하여 단기간 내 인력양성을 추진
- 전국 반도체 특성화 학과 25개 중 계약학과를 제외한 지방 대학에서는 미달이 발생하는 등 계약학과·수도권 편중이 실재하는 만큼 민-관 협력을 통해 ▲채용 연계형 계약학과 지원 확대, ▲고급·실무 프로그램 확충 등 유인책 마련이 시급
  - ※ 작성 시점 반도체 계약학과 8곳 중 6곳이 서울·수도권에 집중되어 있으며, 지방은 3개 학교를 제외한 13개 학과에서 70% 이상의 충원율을 보임(교육부, 2022)

## 참고문헌

- Bloomberg, “Chip Exports to China at Risk on New US Rules, Sparking Selloff”, 2022
- Bloomberg, “Battered by Covid, China Hits Pause on Giant Chip Spending Aimed at Rivaling US”, 2023
- CSET, “AI Chips: What They Are and Why They Matter”, 2020
- Semiconductor packaging trends: an OSAT perspective(Chip Scale Review, 2022
- AI타임즈, “GPU를 대신할 새로운 AI반도체는?”, 2021
- COMPASS, “COMPASS MAGAZINE - 인공지능반도체 산업 동향”, 2021
- IITP, “지능형 반도체 최신 기술 동향”, 2022
- KOTRA, “2021년 대만 반도체 산업 정보”, 2021
- KOTRA, “대만, 격변하는 글로벌공급망 속 반도체산업 육성 노력”, 2022
- KOTRA, “유럽반도체법 주요 내용 및 영향”, 2022
- S&T GPS, “2022년 반기 국내외 주요 기업의 AI반도체 개발 동향”, 2022
- S&T GPS, “2022년 주요국 과학기술정책 동향 및 시사점”, 2022
- S&T GPS, “미국, 인공지능 기술에 대한 중국의 접근을 제한하는 정책 추진”, 2022
- S&T GPS, “오픈소스 기반 ‘RISC-V’, 반도체 업계의 새로운 선택지로 주목”, 2022
- S&T GPS, “테슬라 AI 데이’…자율주행 이어 로봇까지 미래 전략기술 공개”, 2022
- S&T GPS, “퀄컴, ‘스냅드래곤 서밋 2022’ 개최…차기 첨단 반도체 공개”, 2022
- THE AI, “초거대 AI 시대 버팀목 된 한국형 ‘메모리 반도체’”, 2023
- 김&장 뉴스레터, “미국의 2022년 반도체법, 인플레이션 감축법과 국내 기업에 대한 영향”, 2022
- 대외경제정책연구원, 「미국의 대중국 AI 반도체 수출규제 영향과 시사점」, 2022
- 전자신문, “AI반도체, 중소 팹리스에 그림의 떡 설계비만 7,200억”, 2022
- 오윤제, “지능형 반도체 발전 방향 및 전략”, 2020
- 연합뉴스, “중저가 반도체 및 EDA 톨 투자 확대하는 중국, ‘반도체 굴기’ 가속”, 2022
- 연합뉴스, “중, 반도체굴기 포기 검토... 돈만 쓰고 성과 미미”, 2023
- <https://www.anandtech.com/show/16626>
- <https://www.rambus.com/blogs/memory-systems-for-ai-part-5/>

## Ⅰ 용어 해설 Ⅰ

### API | Application Programming Interface

응용 프로그램과 운영체제 통신을 용이하게 하고, 운영체제나 프로그래밍 언어가 제공하는 기능을 제어 가능한 인터페이스

### CNN | Convolutional Neural Networks

합성곱 신경망, 인간의 시신경 구조를 모방한 기술로써 이미지를 분석하기 위한 패턴을 찾는 데 유용한 알고리즘이며, 공간 정보를 유지한 상태로 학습이 가능한 모델

### CUDA | Compute Unified Device Architecture

GPU를 사용하는 소프트웨어 플랫폼, NVIDIA에서 GPU개발을 위한 툴이자 독자적으로 만든 병렬 컴퓨팅 플랫폼(API)로서 CUDA를 소개(2006)하였으며, 대량의 연산을 요구하는 분야에 병렬처리 연산 작업을 통한 성능향상에 유리

### DDR | Double Data Rate

메모리 종류로서 하나의 클럭당 기존의 2배 데이터를 처리하도록 만든 장치를 의미

### DNN | Deep Neural Network, 심층신경망

인공지능 반도체 성능 향상에 따라 기존 인공신경망(Artificial Neural Network, ANN)이 갖는 단점(학습 최적값 도출, 느린 학습시간 등)이 해결되며 은닉층(hidden layer)을 대폭 확장하여 학습의 결과를 향상시킨 기계 학습 알고리즘 중 하나

### FLOPS | Floating point Operations Per Second

초당 부동소수점 연산성능으로 주로 GPU 또는 학습용 NPU 성능 비교 시 사용 (예: T(Tera)FLOPS, 테라플롭스)

### GGPU | General-Purpose computing on Graphics Processing Units

컴퓨터 그래픽 처리를 위한 계산용으로 사용되던 GPU를 CPU처럼 일반 컴퓨팅 영역에 응용 가능하도록 프로그래밍한 범용 계산 장치로서 병렬처리에 적합한 뛰어난 연산속도를 가진다는 장점이 존재

### IP | Intellectual Property, 반도체 설계자산

독립적인 기능을 가지고 재사용이 가능한 회로 또는 칩 레이아웃 디자인으로써 반도체 설계 시 프로세서, 메모리, 디지털/아날로그 신호처리, 다양한 입출력(I/O) 등의 '기능블록'으로써 인공지능·시스템반도체 분야에 있어 ARM사의 모바일 프로세서 설계 자산인 Cortex가 대표적

### ISA | Instruction Set Architecture

Instruction Set Architecture) 명령어 집합체

### NPU | Neural Processing Unit

신경망 처리장치, 인공지능(AI)의 핵심인 딥러닝 알고리즘에 최적화된 프로세서, 빅데이터를 사람의 신경망처럼 빠르고 효율적으로 처리 가능하며 데이터의 병렬연산 처리에 최적화되어 있어 모바일 기기에 탑재되어 시가반의 애플리케이션이 저전력으로 빠르게 동작하도록 도와주며 효율을 높여주는 역할을 수행

### OpenCL | Open Computing Language

GPU를 사용하는 소프트웨어 플랫폼, 애플, NVIDIA, AMD, Intel, IBM에서 개발한 범용 병렬 컴퓨팅 프레임워크로 GPU뿐만 아니라 기타 프로세서(CPU 등)로 이루어진 이기종 컴퓨터 시스템(heterogeneous)을 위한 산업 표준 병렬 프로그래밍 모델

### OPS | Operations per Second

초당 정수 연산 성능으로 주로 추론용 NPU 성능 비교 시 사용하는 단위(예: TOPS, 테라(Tera)옵스)

### RISC-V | Reduced Instruction Set Computer

2010년 UC버클리에서 개발된 무료 개방형(오픈소스) 명령어 집합체(ISA)로 누구나 반도체 칩과 SW, IP를 설계·제조할 수 있는 것이 장점

### RNN | Recurrent Neural Network

순환 신경망, 딥러닝의 가장 기본적인 시퀀스 모델로서 시간에 의존적이거나 순차적인 데이터 학습에 활용

### 기계학습 프레임워크 | Machine Learning Framework

인공신경망의 개발, 학습, 실행, 검증의 기능을 수행하기 위한 공개 SW 프레임워크로서 주로 GPU를 기반으로 구축되며, 최근 NPU, PIM으로 변환 최적화하려는 인공지능 컴파일러 프레임워크와 연동되어 동작

### 레이턴시 | Latency, 지연시간

인공지능 연산을 위해 프로세서들을 대량으로 연결할 때 상호 접속된 연결부의 대역폭(Bandwidth)의 한계로 발생하는 지연시간으로 DNN을 결정하는 중요 요소 중 하나

### 반도체 패키징 | Packaging

반도체로서 기능을 수행할 수 있도록 전원공급, 신호 연결, 열 방출, 외부로부터의 보호할 수 있게 포장하는 기술

### 시스템반도체 | System-on-Chip, SoC

비메모리 반도체의 일종으로 연산·제어 등 정보처리 기능을 갖는 반도체, 기능에 따라 마이크로컴포넌트, 로직IC, 아날로그IC 등으로 구분

### 임베딩레이어 | Embedding Layer

자연어 처리 등에서 단어를 컴퓨터 연산기가 쉽게 인식할 수 있는 숫자 기반의 벡터로 바꿔주고 참조 또는 대조 작업등을 수행하는 연산 레이어

### 팹리스 | Fabless

반도체 제조시설 없이 설계만을 수행하는 전문 기업을 말하며 파운드리를 통해 위탁생산 후 제품을 판매함. 우수한 설계 기술 인력 확보를 필요로 하는 특징이 있음

### 파운드리 | Foundry

팹리스 업체가 설계한 반도체를 위탁 생산하는 기업을 말하며 전문생산업체로 초기에 대량 설비투자 비용이 필요하다는 특징이 있음



## | 저자 소개 |

채 명 식

한국과학기술기획평가원 성장동력사업센터 부연구위원

Tel: 043-750-2608 E-mail: mchae@kistep.re.kr

이 호 윤

한국과학기술기획평가원 성장동력사업센터 연구원

Tel: 043-750-2720 E-mail: hylee@kistep.re.kr

## | 편집위원 소개 |

전 승 수 연구위원

진 영 현 연구위원

유 형 정 부연구위원

정 두 엽 부연구위원

나 영 식 선임전문관리원

한국과학기술기획평가원 사업조정본부

Tel: 043-750-2728 E-mail: dooyupjung@kistep.re.kr

※ 본 KISTEP 기술동향브리프의 내용은 필자의 개인적 견해이며, 기관의 공식적인 의견이 아님을 알려드립니다.

## [ KISTEP 브리프 발간 현황 ]

발간호 (발행일)	제목	저자 및 소속	비고
57 (23.01.06.)	MZ세대를 위한 미래 기술	지수영·안지현 (KISTEP)	미래예측
- (23.01.20.)	KISTEP Think 2023, 10대 과학기술혁신정책 아젠다	강현규·최대승 (KISTEP)	이슈페이퍼 (제341호)
58 (23.02.02.)	세계경제포럼(WEF) Global Risks 2023 주요내용 및 시사점	김다은·김유신 (KISTEP)	혁신정책
59 (23.02.07.)	미국의 「오픈사이언스의 해」 선포와 정책적 시사점	이민정 (KISTEP)	혁신정책
- (23.02.21.)	‘데이터 보안’ 시대의 10대 미래유망기술	박창현·임현 (KISTEP)	이슈페이퍼 (제342호)
60 (23.03.06.)	연구자산 보호 관련 주요국 정책 동향 및 시사점	유지은·김보경 (KISTEP)	혁신정책
61 (23.03.20.)	美 「과학적 진실성 정책 및 실행을 위한 프레임워크」의 주요 내용 및 시사점	정동덕 (KISTEP)	혁신정책
- (23.03.29.)	우리나라 바이오헬스 산업의 주력산업화를 위한 정부 역할 및 지원방안	홍미영·김주원 안지현·김종란 (KISTEP)	이슈페이퍼 (제343호)
62 (23.03.30.)	2021년 한국의 과학기술논문 발표 및 피인용 현황	한혁 (KISTEP)	통계분석
63 (23.03.30.)	2021년 신약개발 정부 R&D 투자 포트폴리오 분석	강유진·김종란 (KISTEP)	통계분석
- (23.04.03.)	국방연구개발 예산 체계 진단과 제언	임승혁·안광수 (KISTEP)	이슈페이퍼 (제344호)
64 (23.04.06.)	2023년 중국 양화의 주요 내용 및 과학기술외교 시사점	강진원·장지원 (KISTEP)	혁신정책
65 (23.04.10.)	2023 인공지능 반도체	채명식·이호윤 (KISTEP)	기술동향