

지능형 반도체 산업 동향 및 시사점

권요안 정보통신기획평가원 산업분석팀 수석

1. 머리말

지능형 반도체는 순차연산 위주인 CPU의 병렬 연산 한계를 극복하고, 대용량 데이터 병렬연산을 요하는 AI 연산을 효율적으로 보조하는 반도체이다. AI 연산을 위해 CPU 내에 연산가속기를 탑재하거나, CPU와 병행하여 GPU, ASIC, FPGA 등을 활용하는 반도체가 활발히 개발되고 있다(<표 1> 참조).

AI 연산은 크게 학습과 추론으로 구분되며, 이들은 행렬에 의한 다중 병렬 연산 등에 기반한다. AI 연산의 데이터 흐름은 학습데이터를 통해 학습 모델을 도출하고, 도출된 모델을 추론 연산에 적용하는 방식으로 진행된다.

추론의 정확도는 도출된 학습 모델에 좌우되고 이에 대응하는 학습 모델 개발을 위한 지능형 반도체의 주요 성능 스펙은 연산처리 속도(FLOPS)와 소비 전력이다. 추론용 지능형 반도체는 정확도와 연산 레이턴시(저지연성)가 주요 성능 지표이다.

또한, AI 학습 모델 데이터 변수의 개수는 증가일로에 있다. GPT 자연어 학습 모델 변수의 경우 2018년 1억 1천만 개 → 2020년 1천 7백억 개로, 구글의 자연어 학습 모델(Switch-C)은 1조 6천억 개로 지수함수적으로 변수 증가하고 있다. 이러한 대용량 데이터에 대응하기 위해 GPGPU, 칩렛(Chiplet) 구조 등 다양한 폼팩터의 시도가 이루어지고 있다.

본고에서는 최근 지능형 반도체 관련 NVIDIA, Intel, AMD 등 주요 기업 동향과 RISC-V, Chiplet 등 산업 동향을 통해 지능형 반도체 산업의 지향점과 시사점을 제시한다.

2. 지능형 반도체 산업 동향

2.1 주요 기업 동향

AI 연산을 위한 반도체는 GPU 중심으로 형성되어 있고 대표적인 기업은 NVIDIA, Intel, AMD 등이다. NVIDIA는 강력한 GPGPU와 CUDA API의 구축으로 시장을 선도하고 있고,

<표 1> CPU/GPU/FPGA/ASIC 특징 및 학습 및 추론 성능 비교 [1]

	CPU	GPU	FPGA	ASIC
특징	<ul style="list-style-type: none"> 범용 및 다양한 AI 연산에 유리 고성능 연산코어 낮은 저지연성 소규모 추론 연산 다소 높은 소비전력 대용량 병렬연산에 한계 범용 및 다양한 AI 연산에 유리 고성능 연산코어 낮은 저지연성 소규모 추론 연산 다소 높은 소비전력 대용량 병렬연산에 한계 	<ul style="list-style-type: none"> 중/거대 AI 학습 모델 개발 풍부한 SW 호환성 다소 큰 칩 면적 소비 전력 면에서 단점 <ul style="list-style-type: none"> 개발기간의 단축, 빠른 리드타임 타 반도체에 비해 비싼 가격 주로 추론용 상대적으로 저전력, 연산유닛 면적 대비 코어간 배선연결 면적 비중이 큰 것이 단점 		<ul style="list-style-type: none"> 개발기간 장기 소요 호환성 부족 높은 전성비 학습 및 추론에 대응가능 효율적인 칩 구조 및 면적 개발기간 장기 소요 호환성 부족 높은 전성비 학습 및 추론에 대응가능 효율적인 칩 구조 및 면적
회사	Intel, AMD, ARM	Intel, NVIDIA, AMD	Intel(Stratix), Xilinx(Versal)	Intel(Gaudi, Goya), Google(TPU), Graphcore

	Training		Inference		Generality	Inference accuracy
	Efficiency	Speed	Efficiency	Speed		
CPU	1×baseline				Very High	~98-99.7%
GPU	~10-100×	~10-1,000×	~10-10×	~10-100×	High	~98-99.7%
FPGA	-	-	~10-100×	~10-100×	Medium	~95-99%
ASIC	~10-1,000×	~10-1,000×	~10-1,000×	~10-1,000×	Low	~90-98%

<표 2> NVIDIA, AMD, Intel 지능형 반도체

제조사	칩	주요 특징
NVIDIA	H100	18,432개 쿼다코어, NVIDIA고유의 900GB대역의 NVlink등으로 120TFLOPS(FP16)
	Blufield	데이터 센터 스토리지 액세스 성능 향상 등을 위해 CPU, GPU, NIC, DDR5, 가속기 등 여러 기능의 반도체들을 집적
	ORIN	칩 내에 4개의 기능(자율주행, 인포테인먼트, 계기판 그래픽, 운전자 모니터링 등)을 구현, Volvo XC 90, Faraday Future FF91, TuSimple, JiDu 등의 자율주행에 활용
AMD	MI200	듀얼 칩렛(Chiplet) 기반, CPU와 AI 가속기간 Infinity Fabric Link 최대 800GB/s, 383 TFLOPS/500~600W
Intel	Gaudi	학습용, ResNet50 이미지 1,590/s
	Goya	추론용, ResNet50(이미지 분류용 벤치마크 테스트) 기준 7ms/15,488fps의 레이턴시 구현
	Ponte Vecchio	칩렛 기반, 128개 코어, ResNet 추론 성능 테스트 결과 43,000image/s(소비전력 600W추정)
	Saphire Rapid	DDR5, HBM, 옵테인 메모리, 학습 및 추론을 위한 AI 가속기 등 탑재한 데이터 센터용 CPU


Intel은 데이터 센터용 AI 가속기를 출시하고 최근 개발된 GPU(Vonte Vechhio)로 시장의 요구에 대응하고 있다. AMD는 NVIDIA의 CUDA에 해당하는 ROCm으로 GPU 생태계 확장에 동참하고 있다.

Google, Meta, Apple, Baidu 등 주요 SW 빅테크 기업은 AI 연산, AR/VR 등에 의한 데이터 센터 연산 로드 증가에 대응하기 위해 지능형 반도체를 자체 개발하거나 경쟁력 있는 스타트업의 가속기 도입을 추진하고 있다.

<표 3> Google, Amazone, Meta 등 SW 플랫폼 기업의 인공지능 반도체 개발 사례

기관명	국가	chip	비고
Google	미국	TPU v4	• v4 4,096개×4set = 1 exa FLOPS 구현
Amazon	미국	Graviton3	• ARM기반의 서버 CPU, 5nm • 7개의 칩으로 구성된 칩렛(ARM코어, DDR5 등) 기반
Meta	미국	-	• NVIDIA GPU A100을 765개로 슈퍼컴 구성
Apple	미국	A15	• 뉴럴엔진 15조8천억/s • 아이폰 및 아이패드 탑재
MS	미국	-	• Azure 서버에 그래픽코어 AI가속기 도입 추진 • NVIDIA A100 서버에 활용
Baidu	중국	Kunlun II (21.8)	• 128 TFLOPS(FP16), 7nm • 최대 소비전력 120W
Alibaba	중국	Yitian 710	• ARM기반의 서버 CPU, 5nm
		XuanTie	• RISC-V기반 서버 CPU, 독자적인 OS AliOS지원

<표 4> 주요 지능형 반도체 개발 스타트업

기업명/ 칩명	주요 특징
Cerebrass WSE2 	<ul style="list-style-type: none"> • 850,000개 AI Core, 50GB on-chip SRAM Memory, core to core bandwidth 20 PB/s, 7nm • CS-2 : 15개의 WSE-2를 장착한 랙으로, 소비 전력 23kW, 120조(Trillion)개 변수 모델 연산 가능 • 웨이퍼 스케일 지능형 반도체 WSE-2와 외부 메모리 연결을 위한 MemoryX 기술 • 15개의 WSE-2를 장착한 CS-2의 병렬 연결을 위한 자체 개발의 SwarmX Interconnection 
Samabanova SN 10 RDU 	<ul style="list-style-type: none"> • TSMC 7nm 공정, On-Chip memory: >300MB, 연산속도: >300 TFLOPS(BF16) • AI학습과 추론 연산을 지원하고, AI 연산시 칩 내 데이터 흐름을 개선, 학습 모델에 따라 재구성 가능하고 복합 작업 지원 <div> 1) High Performance Mixed Workloads  2) Efficient Concurrent Applications  3) Secure Multi-Tenancy  </div> <ul style="list-style-type: none"> • 8개의 SN10을 칩을 적용한 랙시스템으로 수 천개의 GPU가 필요한 거대 모델(고해상도 영상 판독, 자연어 학습 등) 학습 데이터를 파편화하지 않고 연산 가능한 시스템
Lightelligence Pace (Photonic Arithmetic Computing Engine) 	 <ul style="list-style-type: none"> • 광 신호 기반으로 NP완전 문제에서 RTX3080 보다 100배 빠른 연산 • 행렬 연산을 마하젠더 간섭기로 실행, 별도의 소비전력없이 고효율 실행이 가능 • Phase Shifter에 의한 위상 조절로 출력값 조절

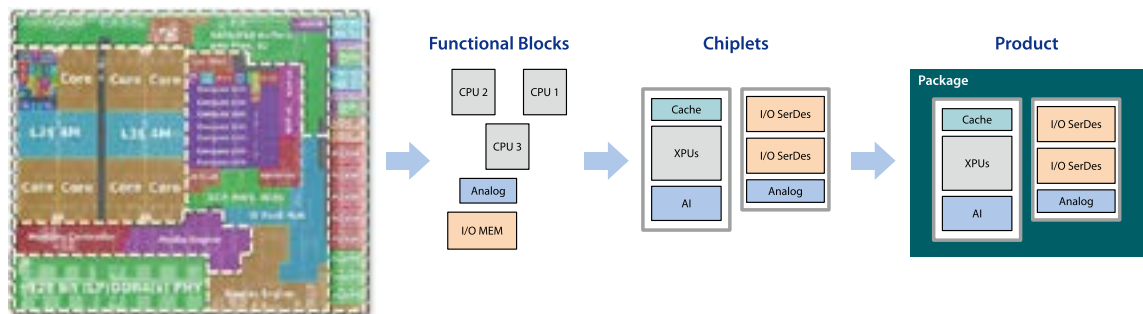
알파고 이후, 지능형 반도체 개발 스타트업이 급증하는 추세[2]이며, 지능형 반도체 초기시장을 선점하기 위해 Cerebrass[3], Sambanova[4], Lightelligence 등은 웨이퍼스케일, 재구성, 광가속기 등 도전적 기술로 시장을 공략하고 있다.

2.2 칩렛(Chiplet)

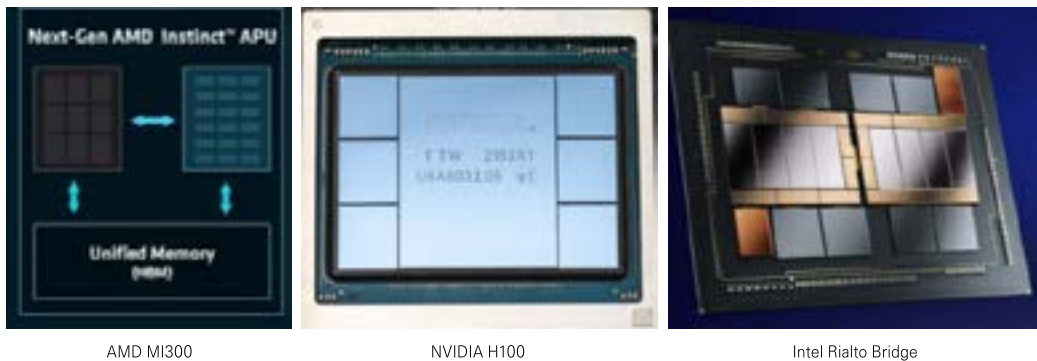
칩렛(Chiplet)은 반도체 패키징의 한 종류다. 다양한 기능을 집적한 단일칩(Monolithic) 구성이 아니라 다이(DIE) 내 기능을 단순화 및 분리하고, 개별 다이 간 인터커넥션을 구성하여 패키징을 적용하는 방식이다.

고성능 다기능을 단일 칩으로 집적하는 것보

다 성능과 수율 간의 트레이드오프나 확장성 등에서 유리하여 단(單) 기능 Multi Die 패키징으로 전환되고 있다. AMD MI300(CPU, GPU, HBM), NVIDIA H100(GPU, HBM), Intel Rialto Bridge(GPU,HBM) 등이 칩렛 구조를 적용한 대표적인 사례이다. 거대 AI 모델 학습을 위해 웨이퍼 스케일의 지능형 반도체 구조를 개발하는 Cerebrass의 사례처럼 웨이퍼 스케일의 칩렛 구조를 적용한 연구도 진행 중이다. Sapatadeep Pal¹ 등은 이종(heterogeneous) 다이를 집적한 2,048개의 칩으로 구성된 칩렛 기반의 웨이퍼스케일 지능형 반도체를 개발, 발표하였다.[6]



[그림 1] 칩렛(Chiplet) 개념도[5]



[그림 2] 칩렛(Chiplet) 적용 사례

¹ 4.3TOPS/725W

한편, 이중 다이 간 데이터 연결을 위한 고속 인터커넥션을 위한 표준이 필요하여, 칩 영역뿐만 아니라 엣지와 데이터센터에서 활용 가능한 개방형 인터페이스 칩렛 표준 UCle²가 AMD, INTEL, 삼성, TSMC, MS, Meta 등을 중심으로 추진되고 있다. UCle의 출현으로 컴퓨터 메인보드가 한 개의 칩렛 구조로 전환되어 스틱 PC보다 진일보한 형태로 컴퓨터 폼팩터에 혁신을 불러올 것으로 예상된다.[7]

2.3 RISC-V

기본 ISA(Instruction Set Architecture) 수가 CISC(Intel, AMD) 1,500여개, RISC(ARM) 200여개인 데 비해 RISC-V는 47개이다 이에 따라 명령어 구조, 확장성, 활용 목적에 따라 ISA를 유연하게 적용해 저전력화가 가능하다는 장점이 있다. 이 뿐만 아니라 반도체 IP 무료 활용, 보안성, 비정치성 등의 강점으로 스타트업 중심으로 최근 활용이 증가하고 있다. 또한 Intel은 RISC-V 프로세서 개발에 4억유로를 투자하고 있고, 대표적인 컴퓨터 저장장치 기업 웨스턴 디지털(Western Digital)과 씨게이트(Seagate)는 SSD/HDD 컨트롤러를 ARM 기반에서 RISC-V로 대체하는 방안을 추진하고 있다.

RISC-V International³은 2025년 반도체 IP 시장이 80억 달러를 상회할 것으로 전망했다. RISC-V의 시장 점유율은 IoT에서 28%, 제조산업 12%, 자동차 부분 10%에 이르고, 약 800억 개의 RISC-V 기반 CPU코어가 제조될 것으로 전망하고 있다.[8]

그러나 반도체 미세 공정에 따른 칩 제작 단가는 증가하나, 칩 제작에 투입되는 비용 중 실제 설계 IP 비용은 15% 이내라는 점을 고려하면, 안정적인 칩 성능 확보를 위해 RISC-V 단독의 전면적인 사용보다는 상업적으로 완성된 IP와의 혼용해 하이브리드 형태로 활용되는 과도기 기간을 거친 이후 시장에 안착될 것으로 판단된다. 테슬라의 D1칩, 애플의 ARM 프로세서에 RISC-V 일부를 도입하고 있는 것이 대표적인 사례이다.

RISC-V의 대표적 스타트업인 Esperanto Technology는 RISC-V 기반으로 데이터센터용 추론 칩을 개발 발표하였고, 이는 RISC-V 기반으로 데이터센터 시장을 공략하고 있는 의미 있는 사례이다.[9]

현재 RISC-V에 가장 적극적인 국가는 중국이다. 중국 과학원은 RISC-V 기반 CPU 프로젝트 장산(Xianshan)⁴을 위해 베이징 카이신연구소(Kaixin Institute)⁵를 개소하고, 중국의 StarFive는 중국 OS 독립을 위해 Kylin Ubuntu를 지원하는 RISC-V CPU 개발을 추진하고 있다.

2.4 뉴로모픽(Neuromorphic) 지능형 반도체

폰 노이만 구조는 연산을 담당하는 연산 유닛(CPU, APU, GPU 등)과 학습데이터를 저장하는 메모리가 물리적으로 분리되어 있는 현대 컴퓨터의 대표적인 구조이다. 이러한 방식은 AI 연산과 같이 대규모 데이터와 복잡한 연산이 필요할수록 연산유닛과 메모리 사이의 병목 현상(bottleneck)으로 많은 시간이 소요된다.

반면, 인간의 뇌는 1,000억 개의 뉴런과 100조

2 UCle : Universal Chiplet Interconnect Express

3 2015년 설립된 RISC-V 관련 비영리 단체

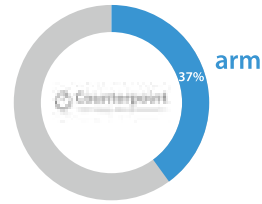
4 SMIC 14nm, 동작 클럭 2GHz 추정, Linux에서 동작 성공

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

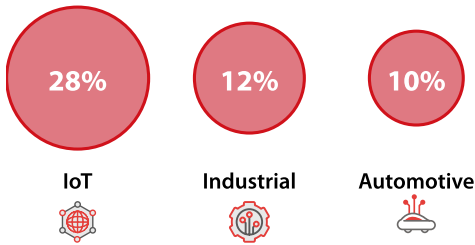
Semiconductor IP market size, 2020 vs 2025



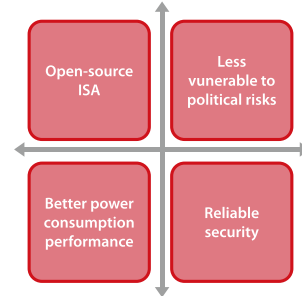
ARM dominates global pure play IP market with 37% share



RISC-V Penetration Rate by 2025



Advantages RISC-V offers



[그림 3] RISC-V 시장 전망[8]

개 이상의 시냅스가 병렬 연결되어 있고, 뇌 신경 전달 신호는 event 기반의 스파이크 신호로 구성, 단 20W 수준의 저전력으로 기억/연산/추론/학습 등 고도의 연산을 동시에 수행한다.

뉴로모픽 반도체는 인간 뇌의 동작을 모방하여 초저전력으로 구동되며, 메모리 병목 해결과 저전력 구현을 위해 스파이킹 신호, 시냅스 및 뉴런 등 인간의 뇌 신호처리를 모사한 스파이킹 뉴런 네트워크(SNN)에 대응 가능한 구조의 AI 연산 처리용 지능형 반도체이다.

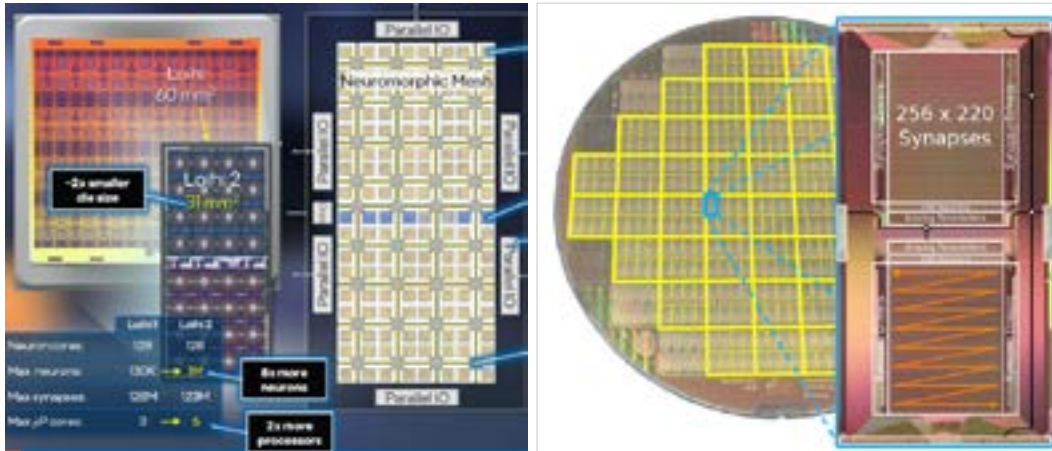
뉴로모픽 반도체 분야에서는 뇌신경 모사를 위해 3차원 연결 신경망, 신호처리를 담당하는 뉴런, 메모리와 신호 전달함수 기능 등을 수행하는 시냅스(Synapse)로 모델링하고, 이의 구현을 위해 칩 내에 라우팅, 멤리스터(Memristor), NVM(RRAM, PRAM, MRAM 등) 등을 이식한

구조 연구가 활발하다. 그러나 뇌의 기전과 SNN 이론은 정립되지 않은 상태라 기초 선행 연구가 더 필요하다.

최초의 뉴로모픽 반도체는 IBM의 TrueNorth (2014년)로 64×64 시냅틱 코어를 구성하고 코어당 라우터를 집적하여 시냅틱 코어간 연결망을 형성했다. 신경망의 가중치, 임계값, 출력값 등을 저장하기 위해 코어 내 SRAM을 구성하여 1W 미만(25~275mW) 수준의 전력으로 초당 1,200~2,600 프레임 이미지를 분류 가능하게 구현하였다.

인텔도 뉴로모픽 반도체 개발에 동참, Loihi2⁶ (2021년)를 개발하고 스파이크 신경망을 지원하는 텐서플로우와 유사한 SW 플랫폼 LAVA를 출시했다. 최근 하이델베르크 대학과 드레스덴 공대의 협업으로 1개의 칩당 256×220 시

6 IBM에 비해 4~22배 빠른 성능으로 알려져 있으나 소비전력은 미공개



[그림 4] 인텔의 Loihi2(좌)와 하이델베르크 대학의 웨이퍼 스케일 뉴로모픽 반도체(우)

냅스 코어와 FPGA를 집적하고 1개의 칩당 1만 6,000개의 동시입력과 초당 164 Giga-events의 처리가 가능한 웨이퍼 스케일 뉴로모픽 반도체를 개발하였다.[10] 현재 까지 뉴로모픽 반도체의 상용화는 소자 레벨(메모리)에서의 신경 모사보다 신경의 기능을 구현한 회로 설계에 기반하고 있다.

3. 맺음말

전 산업의 AI 확산과 거대 학습 모델의 수요는 데이터센터의 부담으로 직결되며, 이에 따라 데

이터센터 연산 칩은 인텔 중심의 x86에서 x86 + (ARM, GPU, AI 가속기)로 전환할 전망이다. 또한 지능형 반도체는 관련 반도체 산업인 SoC, 메모리, 스토리지 등에 대한 수요를 더욱 증가시켜 시스템 반도체 시장 성장을 견인할 것이다.

RISC-V는 Intel, AMD, ARM, Qualcomm 등 CPU, APU 코어 시장 지배구조에 지각변동을 유발하고 결국 현 ARM, x86과 더불어 또 하나의 코어 시장을 형성할 것으로 전망된다.

아직 초기 단계인 지능형 반도체는 국내 시스템 반도체 산업 육성에 좋은 기회이고 이를 위해서는 보다 적극적인 정책 개발이 필요하다. TTA

참고문헌

- [1] CSET, “What They Are and Why They Matter”, 2020
- [2] Hiroshi M. et al, “ Systems and circuits for AI chips and their trends,”Japanese Journal of Applied Physics, 59 050502, 2020
- [3] Sean L., “The Multi-Million Core, Multi-Wafer AI Cluster”, Hotchips33, AUG, 2021
- [4] Raghu P., “SambaNova SN10 RDU: Accelerating Software 2.0 with Dataflow”, Hotchips33, AUG, 2021
- [5] David C. “Semiconductor packaging trends: an OSAT perspective”, Chip Scale Review Ja/Feb. 2022
- [6] Saptadeep Pal et al, “Designing a 2048-Chiplet, 14336-Core Waferscale Processor,” 58th ACM/IEEE Design Automation Conference (DAC), Dec, 2021
- [7] Debendra D. S., “Universal Chiplet Interconnect Express (UCIe)®: Building an open chiplet ecosystem”, 2022(<http://www.uciexpress.org>)
- [8] RISC V Foundation
- [9] David D., “Accelerating ML Recommendation with over a Thousand RISC-V/Tensor Processors on Esperanto’s ET-SoC-1 Chip”, Hotchips33, AUG, 2021
- [10] Eric M et al., “The Operating System of the Neuromorphic BrainScaleS-1 System,” arXiv:2003.13749, Feb, 2022