www.iitp.kr
대한민국의 혁신과 도전, Global Digital Frontier
정보통신기획평가원

SPOT 2025-13 | 2025년 12월 23일

ICT SPOT ISSUE
ICT 산업동향

글로벌 AI반도체 기업 동향 분석

보고서 관련 문의처

작성자 탁영호 책임
소 속 정책기획팀
☎ 042-612-8214
✉ yhtak@iitp.kr

보고서 목차

주요 내용 요약

□ 개요

○ AI가 산업 전반의 혁신을 주도하며 AI반도체 시장이 빠르게 확대되는 상황에서, 기존 엔비디아의 독점 체제에 맞선 팹리스·빅테크·스타트업 간 주도권 경쟁 본격화

□ 글로벌 AI반도체 기업 동향

○ (글로벌 팹리스) 엔비디아가 GPU 시장 지배력을 공고히 하는 가운데, AMD와 화웨이는 자사 AI반도체를 개발·고도화하며 시장 점유율 확대 중

- 엔비디아는 압도적인 GPU 성능과 CUDA 기반의 독자적인 SW 생태계를 바탕으로 AI 학습·추론 시장 전반에서 지속적인 경쟁력 확보
- AMD('오픈소스')와 화웨이('AI 풀스택')는 각각 차별화된 전략으로 엔비디아 GPU 아성에 도전

○ (빅테크) 자사 맞춤형 AI반도체 개발을 통해 脫 엔비디아 시도하며, AI반도체의 성능과 효율을 동시에 최적화 하는 방향 모색

- 엔비디아 의존 구조를 벗어나기 위해 구글, MS, 아마존, 알리바바 등은 자사에 특화된 AI반도체 개발에 집중

○ (스타트업) GPU의 구조적 병목(지연·대역폭·비용)을 보완하기 위해 추론 특화 AI 반도체를 개발하고, 온디바이스 AI 산업 분야에서의 입지를 강화

- AI 추론 시장의 급성장을 기회로 삼아, 기존 GPU 구조와 차별화된 AI반도체를 개발하고 이를 기반으로 대규모 투자 유치와 시장 진입을 가속화

□ 시사점

○ R&D 전 주기 연계 체계 형성, 세제 혜택 확대, 신뢰성 검증 체계 구축, 미래 융합형 전문 인재 양성 등을 통해 AI반도체 독자적인 기술 경쟁력 확보

ISSN 2982-8279

1 개요

AI가 범용 기술로 자리 잡으며 경제·사회 전반으로 AX(AI 전환)가 빠르게 진행되며 생산성 향상, 노동의 변화 등 광범위한 혁신을 주도

* 경제활동 40~50%가 AI영향, 생산성 20% 향상 시 10년간 GDP 10% 증가(스탠퍼드大, '25.1)

- 이제 AI가 스스로 판단하고 실행하는 단계까지 발전하며, 물리적 환경과의 상호작용이 가능해짐에 따라 산업적 파급력은 더욱 확대될 전망

AI반도체는 모든 AI 서비스의 필수 구동 자원으로서 AI 기술 발전을 가속화하고, AI 가치사슬 최일선에서 시장 성장을 주도

- AI 모델의 학습·추론 성능 향상을 위한 필수 요소로, 대규모 데이터 처리와 복잡하고 정교한 AI 연산을 뒷받침
- 특히, 최근에는 AI 에이전트와 같은 AI 추론 서비스가 본격화* 되며, 학습 중심을 넘어 고효율·저전력 특성이 요구되는 추론용 AI반도체의 중요성이 부각

* 글로벌 추론용 AI반도체 시장 전망(옴디아, '23) : ('23) 60억 달러 → ('30) 1,430억 달러

하지만, AI반도체 시장을 장악한 엔비디아 GPU는 고비용·고전력 구조로 인해 AI 데이터센터의 전력 사용과 비용을 급증시키며 지속 가능성에 한계

- 방대한 양의 데이터 처리를 위해서는 연산과 메모리의 변화가 불가피하며, 이를 해결하는 기업이 미래 반도체의 승자가 될 예정

* "우리는 지금 AI가 모든 컴퓨팅 플랫폼에 융합되는 지각변동을 마주하고 있다"(젠슨 황, '25.5)

이에 엔비디아가 주도하는 시장에 팹리스(AMD·화웨이), 빅테크(구글, MS, 알리바바 등), 스타트업(그록, 캠브리콘 등)이 가세하며 치열한 경쟁 전망

2 글로벌 AI반도체 기업 동향

가 글로벌 팹리스(Fabless)

▣ 엔비디아가 GPU 시장 지배력을 공고히 하는 가운데, AMD와 화웨이는 자사 AI반도체를 개발·고도화하며 시장 점유율 확대 중

* '24년 1~3분기 GPU 글로벌 시장 점유율(IoT Analytics, '25) : (엔비디아) 92%, (AMD) 4%, (화웨이) 2.2% 등

▣ (엔비디아) 압도적인 GPU 성능 우위*와 독자적인 SW 플랫폼인 CUDA를 기반으로 글로벌 AI반도체 생태계를 장악하여 막강한 영향력 지속

* 성능(FP8기준, PFLOPS) : (엔비디아 '블랙웰 울트라') 10 vs. (AMD 'MI325X') 5.2

- 고성능 HW와 강력한 SW 생태계를 결합하여, AI 학습·추론 시장 전반에서 주도권 확보
  - 엔비디아 GPU는 AI 학습용 시장 중심으로 성장하였으며, CUDA 플랫폼을 기반으로 사실상 표준화된 개발 환경을 구축함으로써 AI반도체 생태계 지배력을 강화
- 기존 GPU에만 의존하지 않고, 차세대 AI 칩(B200*, GB300** 등) 개발을 통해 AI 반도체 포트폴리오를 확대하며 기술 경쟁력을 지속적으로 고도화

  * 이전 세대(호퍼)의 성능을 뛰어넘는 차세대 GPU(블랙웰)

  ** 하나의 칩 안에 CPU와 2개의 GPU를 결합한 통합 프로세서

  - 이러한 차세대 AI 칩 개발을 기반으로 다양한 산업·응용 분야에 적용 가능한 범용 AI반도체 전략을 구사하고, 제품 경쟁력과 활용 확산 기반을 강화
- 또한, 오픈AI와의 전략적 파트너십을 체결하여 차세대 AI 모델 학습·서비스를 위한 최소 10GW 규모의 AI 데이터센터 구축 지원('25.9)[1)]

  * 첫 번째 1GW는 2026년 하반기에 NVIDIA 'Vera Rubin' 플랫폼으로 가동 예정(최대 1,000억달러 투자 계획)

  - 이를 통해, 단순한 AI 칩 공급사를 넘어 AI 인프라 생태계에서 영향력과 시장 지배력 강화

1) https://nvidianews.nvidia.com/news/openai-and-nvidia-announce-strategic-partnership-to-deploy-10gw-of-nvidia-systems

(AMD) 개방성·효율성(CPU-GPU-메모리 단일 패키지)·확장성(고용량 메모리 탑재*)과 가격 경쟁력을 기반으로 시장 확산 노력 경주

* (AMD 'MI350X') 288GB HBM3E vs. (엔비디아 'B200') 192GB HBM3E

- 특히, 오픈소스 기반의 개방형 SW 생태계(ROCm*) 구축과 함께, 호환성(HIP**)·신규 API 제공까지 갖추며 非CUDA 생태계의 대안으로 대두

* Radeon Open Compute : AMD의 GPU 프로그래밍을 위한 개방형 SW 플랫폼

** Heterogeneous Compute Interface for Portability : 엔비디아 CUDA와 유사한 GPU 개발 환경을 제공하는 인터페이스

  - ROCm을 통해 개방성, 유연성, 비용 효율성 등을 내세워 개발자·연구소·기업들의 참여를 확대하는 전략을 추진하며,
  - 단순히 엔비디아의 대체 플랫폼이 아닌 AI 개발·배포를 위한 실질적인 SW 생태계로 성장 중

- 또한, 자사 최신 GPU(인스팅트 MI350 등)와 결합하여 대규모 AI 모델 학습·추론에 적합한 인프라 환경 제공

- 최근에는 AI반도체 시장 점유율 확대를 위해 오라클, 오픈AI 등 글로벌 AI 기업*과의 파트너십 전략을 강화

* △오라클 클라우드 인프라스트럭처(OCI)에 AMD 'Instinct MI450' GPU 탑재 예정('25.10), △오픈AI 6GW급 데이터센터에 GPU 공급('25.10) 등

(화웨이) 중국 AI반도체 기술 자립을 이끄는 선두 주자로, 자사 GPU를 지속적으로 개발·고도화하고 HW부터 SW까지 아우르는 독자 생태계* 구축

* (AI칩) 어센드 - (아키텍처) CANN - (AI프레임워크) 마인드스포어, - (AI모델) 판구

- '어센드'칩을 중심으로 빠르게 기술 자립을 이루며, 거대한 내수시장을 기반으로 성장
  - 다만, 아직 엔비디아 최신 GPU 대비 절대 성능*은 부족하나 향후 자체 HBM 탑재 계획을 공개*하는 등 핵심 기술 내재화와 성능 고도화를 본격화하며 기술 격차를 축소할 전망

* (어센드 910B) A100 대비 80% 효율, (어센트 910C) H100 대비 추론 성능 60% 수준 등

** 3개년 AI 칩 로드맵(화웨이커넥트, '25.9) : 어센드 950PR('26.1분기), 950DT('26.4분기), 960('27.4분기), 970('28.4분기)

- AI칩과 SW를 풀스택으로 통합하여 독자적인 생태계를 구축하고, CANN(CUDA 역할)을 오픈소스로 전환하여 자사 AI 칩 활용을 유도
  - SW 최적화 전략을 통해 클라우드·데이터센터 환경에서의 활용성을 높이고, 마인드스포어와의 통합으로 AI 모델 개발·배포 생태계를 한층 강화

- 또한, 딥시크와의 협력을 바탕으로 '칩(어센드) → 클라우드(화웨이 클라우드) → AI모델'로 이어지는 통합 설계를 구축하며, 시장 확대와 비용 효율화 전략을 함께 추진

  * (딥시크) 차세대 AI모델 'R2' 훈련에 화웨이 '어센드' 칩 일부 적용 중('25.8)

표1 글로벌 팹리스 기업의 AI반도체 성능

기업명	엔비디아	AMD	화웨이
국가			
주요 제품	GB300	Instinct MI350x	어센드 910C
목적	학습·추론	학습·추론	학습·추론
전력소모(W)	1,400	1,000	310
공정·패키징	Custom-built TSMC 4nm	TSMC 3nm + 6nm dual	SMIC 7nm N+2
메모리 용량	288GB HBM3E	288GB HBM3E	128GB HBM3
대역폭	8 TB/s	8 TB/s	3.2 TB/s

| 자료 : 각 기업 종합

나 빅테크(Big Tech)

▣ 자사 맞춤형 AI반도체 개발을 통해 脫 엔비디아 전략을 가속화하고, AI 워크로드별 특성에 최적화된 성능과 효율을 동시에 확보하려는 방향 모색

▣ (구글) '15년부터 자사 전용 AI 칩(TPU) 개발하여 AI 연산 성능과 전력 효율을 지속적으로 고도화*하였으며, 최근에는 외부 데이터센터까지 적용 범위를 확대

* TPU v1('15)→TPU v2('18)→TPU v3('20)→TPU v4('22)→TPU v5('23)→TPU v6('24)→TPU v7(아이언우드, '25)

- 경쟁사에 대한 AI 칩 의존도를 낮추고, AI 클라우드 시장을 선점하기 위해 TPU를 개발·고도화하여 연산 성능과 전력 효율을 동시에 확보

  – 제미나이, 이마젠을 포함한 AI 모델과 클라우드 서비스를 구동하고 있으며, 강화학습을 사용하여 칩 레이아웃을 최적화하고 설계 시간 단축하여 성능을 향상

  – 특히, 7세대(아이언우드)는 초거대 AI 모델의 추론 성능 향상을 위해 메모리·대역폭·전력 효율성 등 극대화, 엔비디아의 최신 GPU와 경쟁 가능한 수준에 도달

* TPU 6세대보다 고대역폭 메모리(HBM) 용량 6배·대역폭 4.5배 증가, 전력효율 2배 향상

- 자사 초거대 언어모델(제미나이3)에서 TPU를 활용하면서 성능과 안전성을 검증하였고, 빅테크와 클라우드 서비스 기업(CSP)에 공급하는 전략을 통해 AI 시장에서의 입지 확대

* '메타'의 데이터센터에 TPU 대규모 공급 논의('25.11), '앤트로픽'에 최대 100만개의 TPU 공급계약 체결('25.10), '플루이드스택(英 CSP 기업)'의 뉴욕 데이터센터에 TPU 설치 추진('25.9)

– 기존에는 구글 클라우드 인프라로만 제공되던 TPU가 고객사 요청에 따라 이제는 외부 기업에 온프레미스* 방식으로 직접 판매하는 전략도 함께 추진

* 고객사의 데이터센터에 AI반도체를 직접 설치하는 방식

(마이크로소프트) AI 학습·추론 비용을 절감하며, 장기적으로 자사 클라우드 플랫폼(애저) 중심의 독자 AI 인프라를 구축하기 위한 자체 AI 칩 개발 본격화

- 대규모 클라우드 기반 AI 학습과 추론에 최적화된 '마이아100' 칩은 '애저' AI 인프라에 직접 통합되어 학습·추론 성능을 개선

– HW·SW 공동 설계를 통해 효율성을 극대화하고 GPU 공급망 리스크와 비용을 절감하는 핵심 전략으로 활용

- 클라우드 인프라의 핵심 부품을 외부 공급망에 의존하지 않고 시스템 설계 자율성을 확보하기 위해 AI 인프라 전반을 직접 통제하려는 전략 추진

– 데이터센터 내에서 사용 중인 AMD·엔비디아 AI 칩을 단계적으로 대체할 계획을 공식화하고, '26년 상용화를 목표로 차세대 칩 '마이아200' 개발에 착수('25.10)

* '25~27년에 코드명 브라가(Braga), 브라가-R(Braga-R), 클레아(Clea)의 세 가지 추론 칩 생산 로드맵 수립(더인포메이션)

- 다만, '마이아200'의 성능은 엔비디아 블랙웰 대비 다소 낮을 것으로 예상되어 CUDA 기반 생태계를 단기간에 대체하기에는 다소 시간이 소요될 예정

– 그럼에도 자체 AI 칩 확보를 통해 클라우드 고객에게 더 다양한 선택지를 제공할 수 있으며, 이는 장기적으로 따라 마이크로소프트의 서비스 경쟁력 강화로 이어질 전망

(아마존) 자사 맞춤형 AI반도체 개발을 통해 비용·성능 경쟁력을 확보하고 AI 풀스택 전략, 파트너십 체결 등을 통해 시장 확대 모색

- 생성형 AI 모델 학습('트레이니엄3', '25)과 추론('인퍼런티아2', '24)에 특화된 AI 칩을 각각 설계·개발하여 맞춤형 클라우드 AI 환경 제공

– 엔비디아 등 범용 GPU에 대한 의존도를 낮추는 한편, 비용 효율성과 성능을 극대화하여 경쟁 우위 확보

- AI 칩부터 SW·클라우드·AI 서비스까지 이어지는 AI 풀스택* 경쟁력을 확보하여 글로벌 AI 인프라 시장 영향력 강화

  * AI 칩 → SW(Neuron SDK) → 클라우드 플랫폼(EC2) → AI 서비스(베드록)

- 자체 AI 칩의 실사용 레퍼런스를 확보하기 위해 앤트로픽과의 전략적 파트너십 체결('23)

  * 아마존은 앤트로픽에 총 80억 달러 투자(약 11.7조 원, '24년 기준)하고, 앤트로픽과 협력하여 AI 칩 개발·최적화 진행

  – 앤트로픽의 '클로드(생성형 AI)'에 '트레이니엄'과 '인퍼런티아' AI 칩을 이용해 학습·추론 수행

(메타) 페이스북, 인스타그램 등 자사 AI 서비스 비용을 절감하기 위해 자체 AI 칩 개발, 이를 기반으로 독자적인 추론 생태계 구축 모색

- AI 칩 'MTIA*'는 GPU 대비 비용 절감 효과와 낮은 전력 소비량을 강점으로 내세우며, AI 플랫폼 생산성 향상에 기여

  * Meta Training and Inference Accelerator

  – 대규모 AI 서비스(추천, 광고, 랭킹 등)에 적합한 맞춤형 추론 플랫폼을 확보하는 기반으로 활용되며, 실시간 추론 처리 역량을 강화

- 현재 'MTIA'는 특정 추론 처리 영역에 특화 되어있지만, 장기적으로는 생성형 AI(라마)까지 확장하여 자체 AI 인프라를 완성하는 방향으로 진화할 전망

(알리바바) 자사 클라우드 운영에 필요한 맞춤형 AI 칩을 개발하고, 칩을 직접 판매하기보다는 AI 칩 기반 컴퓨팅 자원을 '서비스 형태*'로 제공하는 전략 추진

  * Chip-as-a-Service 전략

- 자회사(T-head)를 통해 추론용 AI 칩 '한광 800'을 출시('19) 후 미국의 반도체 수출 규제에 대응하기 위해 새로운 AI 칩 'PPU' 개발('25.9)

  – 자사 클라우드(알리윤)을 통해 고객이 'PPU' 자원을 서비스 형태로 이용하게 함으로써 리스크를 줄이고 효율 높이는 방향 추구

  – 이러한 'PPU'는 엔비디아 H20 대안으로 부각되고 있으며, 중국 내 대규모 데이터센터에 투입 예정

- 또한, 클라우드 기반 AI 연산을 지원하는 '진우(Zhenwu)' 칩을 개발하여 AI반도체 자립을 추진

  – 추후 알리바바가 자체 개발한 AI 모델들에 탑재하여 테스트 예정

(바이두) 자사의 검색엔진과 LLM(어니봇) 등 대규모 AI 연산을 위해 '쿤룬'칩을 개발하였으며, 이를 기반으로 초대형 AI 컴퓨팅 클러스터 '완카*'를 구축·운영[2)]

* '1만개 카드'라는 의미로 바이두 AI컴퓨팅 인프라 브랜드('쿤룬 P800'칩 3만개로 구성)

- AI 칩 설계 → AI 클러스터 구축 → AI 모델 학습·추론 서비스로 이어지는 AI 풀스택 전략을 추진
  - 초기 '쿤룬' AI 칩이 추론 중심이였다면, 차세대 '쿤룬 P800'칩은 LLM 학습 성능까지 강화하여 학습과 추론을 모두 지원하는 범용 AI반도체로 발전시키는 로드맵을 진행 중
- 또한, 차세대 AI칩 'M100'과 'M300'을 공개하며 자사 에이전트와 옴니모달* 생태계 전반의 AI 연산 기반을 한층 강화('25.11)

* 멀티모달보다 한층 진화한 개념으로 텍스트, 이미지, 음성 등 모든 데이터를 동시에 통합하여 인간처럼 추론하는 AI 개념

  - 두 AI 칩 모두 대규모 멀티모달 처리와 고속 추론을 위해 설계되었으며, 'M100'은 MoE*기법을 사용하여 AI 모델 추론 효율성을 향상

* 전문가 혼합 방식(Mixture of Experts) : 하나의 거대 단일 모델 대신, 데이터 입력시 특정 모델만 선택해서 처리하는 방식

  - 'M300'은 초대형 멀티모달 모델의 학습과 AI 에이전트 기반 실사용 환경에서의 효율성과 실시간 반응 속도를 강화하면서도 전력 소모는 기존 대비 최대 30% 절감

표2 빅테크 기업의 AI반도체 성능

기업명	구글	마이크로소프트	아마존	메타	알리바바	바이두
국가						
주요 제품	TPU v7 아이언우드	마이아 100	트레이니엄 2	MTIA 2	한광 800	쿤룬 P800
목적	학습·추론	학습·추론	학습	추론	추론	추론
전력소모(W)	-	700	500	90	276	150
공정·패키징	-	TSMC 5nm CoWoS	TSMC 5nm CoWoS	TSMC 5nm	TSMC 12nm	TSMC 7nm
메모리 용량	192GB HBM3E	64GB HBM2E	96GB HBM3	128GB LPDDR5	192MB SRAM	32GB GDDR6
대역폭	7.4 TB/s	1.8 TB/s	2.9 TB/s	204.8 GB/s	-	512 GB/s

| 자료 : 각 기업 종합

2) https://www.reuters.com/world/china/chinas-baidu-says-its-kunlun-chip-cluster-can-train-deepseek-like-models-2025-04-25/

다 스타트업(Startups)

▣ GPU의 구조적 병목(지연·대역폭·비용)을 보완하기 위해 추론 특화 AI 칩을 개발하고, 데이터센터의 특정 연산 수요와 온디바이스 AI 산업 분야 집중 공략

▣ (美, 그록) GPU 대비 빠른 응답 속도를 구현하는 초저지연·고처리량 추론 전용 AI 칩(LPU*)을 개발하며, 실시간 LLM 추론 시장에 역량을 집중

* Language Processing Unit

- 추론 영역에서의 핵심인 높은 AI 연산 속도*와 효율성 면에서 강점을 확보

  * 메모리 대역폭 : (엔비디아 'GB300') 8 TB/s vs. (그록 'LPU') 80 TB/s

  – 실시간 챗봇, 음성 인터랙션, 스트리밍 모델 등 지연 민감 서비스에서 GPU 대비 압도적인 응답 속도를 제공

- 이러한 기술 성과를 바탕으로 대규모 투자*를 유치하여 AI 추론 전용 칩 시장의 선도 스타트업 중 하나로 부상

* 신규 투자 유치(단회 기준, '25) : 7억 5천만 달러(약 1.1조원), 기업가치('25) : 69억 달러(약 10조원)로 평가

▣ (美, 삼바노바) HW부터 SW까지 아우르는 'AI 풀스택 통합 전략'을 추진하여 대규모 투자 유치*에 성공하였으며, 글로벌 AI반도체 시장에서 활로 모색

* 누적 투자 유치(~'21) : 총 11억 3천만 달러(약 1.6조원)

- 기존 CPU·GPU와 달리 추론 단계에서 발생하는 데이터 흐름을 효율적으로 관리하기 위해 데이터플로우* 방식(RDU)을 채택

  * Recongifurable Dataflow Unit : 칩 구조를 실시간으로 재구성하여 데이터 처리의 효율성을 향상시키는 기술

  – 데이터 이동을 최소화하여 GPU 대비 6.6배 빠른 속도를 구현하였고, 5배 이상 높은 에너지 효율로 대규모 언어 모델(LLM) 추론에서 뛰어난 성능을 발휘

- AI 풀스택 전략의 일환으로, AI 칩(SN40L)부터 시스템(삼바랙), 클라우드(삼바클라우드), SW(삼바노바 스위트)까지 완벽하게 통합된 솔루션 제공

  – 또한, 다양한 AI 도입 환경에 대응하기 위해 높은 확장성을 가진 클라우드 기반 서비스와 보안에 용이한 온프레미스 구축 방식을 모두 지원

- 최근에는 시장 진출 지연으로 수익 창출, 추가 자금 확보 등에 한계가 발생하여 매각* 모색

* 인텔에 16억 달러(약 2조 3천억원) 규모로 최종 인수 협상 논의 중('25.12)[3]

(美, 세레브레스) 웨이퍼스케일 아키텍처(WSE-3*)로 GPU 구조의 물리적 한계를 돌파하고, LLM·LMM(멀티모달) 모델 학습에 최적화된 독자 시스템 구축

* Wafer Scale Engine 3

- 단일 웨이퍼 전체를 거대한 하나의 칩으로 사용하여 기존 GPU에서 나타나는 병목현상을 해소하며, 극단적 확장성을 구현하는 전략 추진
  - 칩 간 통신 지연과 대역폭 병목을 제거하여, 설계 단순성·에너지 효율성 등을 극대화함으로써 AI 모델 학습에 최적화된 성능을 구현
- 'WSE-3' AI 칩을 기반으로 한 'CS-3' AI 슈퍼컴퓨터 시스템을 통해 AI 모델 학습에 특화된 플랫폼을 제공하며 GPU 대비 학습 속도와 확장성에서 경쟁력 보유
  - 최대 24조 개의 파라미터를 가진 대규모 AI 모델 훈련을 지원하며, 데이터 병렬 처리 기술을 통해 학습 훈련을 가속화
- 최첨단 AI 모델인 'Qwen3-235B'를 출시('25.7)하여 실시간 AI 추론 속도를 향상* 시켰으며, GPU 클러스터의 대안·보완 솔루션으로 자리매김 중[4]

* 초당 1,500개의 토큰 처리 속도를 가속화하여 응답 시간은 1~2분에서 0.6초로 단축

(美, 신티언트) 초저전력 특화 온디바이스 AI 칩*을 개발하여 웨어러블·IoT의 실시간 처리 성능과 에너지 효율 향상시켜 음성·센서 인식용 AI반도체 시장 선도

* NDP 200 : 1mK 미만의 전략 소비로 고정밀 추론 수행

- 클라우드가 아닌 기기 내 AI 추론 실행을 통한 개인정보 보호와 저지연 응답 구현
  - 전력 효율이 높아 배터리로 구동되는 기기에서도 상시 작용(Always-On)이 가능하며, 음성 명령 인식 등 실시간 AI 기능을 안정적으로 제공
- '놀즈(Knowles)'기업의 마이크로폰 사업부를 인수('24)하는 등 온디바이스 음성·오디오 AI 시장에서 리더십 확보에 주력[5]
  - 기존 주력 분야인 모바일, 웨어러블 분야를 넘어 자동차(자율주행) 등 차세대 음성 AI 인터페이스 수요가 급증하는 고부가가치 시장으로의 영역 확장 가속화

3) https://www.bloomberg.com/news/articles/2025-12-12/intel-nears-1-6-billion-deal-for-ai-chip-startup-sambanova
4) https://www.unite.ai/cerebras-unveils-qwen3%E2%80%91235b-a-new-era-for-ai-speed-scale-and-cost/
5) https://finance.yahoo.com/news/syntiant-completes-acquisition-knowles-consumer-122512306.html

(加, 텐스토렌트) 오픈소스(RISC-V*) 기반의 아키텍처를 활용하여 AI 칩 설계의 유연성을 확보하고 자동차·모빌리티, 온디바이스 AI 등 다양한 산업 진출 추진

* Reduced Instruction Set Computer Fifth generation : 오픈소스 기반으로 확장성·유연성을 갖춘 차세대 프로세서 기술

- 개방형 전략을 통해 자동차, 로봇 등 다양한 분야의 고객 맞춤형 AI 칩 설계가 가능
  - 특히, 칩렛* 구조를 적용하여 비용을 절감하고 전력효율을 향상시키는 효과 제공

    * 여러 개의 작은 기능별 칩을 연결하는 방식
- 6억 9,300만 달러(약 1조원) 규모의 시리즈 D 펀딩을 유치하며, 오픈 생태계 기반 AI반도체 기업으로의 성장 가능성을 입증('24)

* 기업가치는 26억 달러(약 3조 8천억원)로 평가('24)

  - 가전·스마트홈(LG전자), 온디바이스 AI반도체 설계 IP(라피더스), 자동차용 AI 칩렛(보스) 등 광범위한 기술 파트너십 전략도 함께 추진
- '블루 치타 아날로그 디자인*' 기업을 인수('25.7)하여 칩렛 간의 연결 기술을 내재화하고 오픈 칩렛 아키텍처(OCA)와의 연계를 통해 개방형 생태계 구축 가속화

  * 칩렛 간의 통신을 위한 첨단 인터커넥스 기술 전문 기업

(이스라엘, 헤일로) 온디바이스 AI를 위한 초저전력·고성능 AI 칩을 개발하여 실시간 추론을 강화하고, 투자·상용화를 발판으로 온디바이스 AI 대중화를 주도

- 온디바이스용 추론·경량 LLM에 특화된 아키텍처를 기반으로, 저전력 환경에서 고효율 연산과 초저지연 AI 서비스를 구현할 수 있도록 설계
  - 클라우드 통신 지연·대역폭 부담을 최소화하여 스마트카메라, 로보틱스 등 온디바이스 AI 활용 분야에서 경쟁력 확보
- 지속적인 대규모 투자유치와 글로벌 수요처를 통해 온디바이스 AI 시장 확대 추진
  - 1억 2천만 달러('24)의 신규 투자를 포함하여 현재까지 3억 4천만달러(약 5천억원) 이상 조달하였으며 한국, 중국, 일본, 미국, 유럽 등 여러 고객사와 300여개의 프로젝트 진행 중

(中, 캠브리콘) 중국 최초의 AI 전용 칩 스타트업으로, 자사 AI 칩(시위완)과 SW(뉴웨어) 동시 공략을 통해 중국 AI반도체 생태계의 핵심 축으로 대두

- 엔비디아의 A100 GPU를 겨냥하여 설계된 AI 추론용 칩*을 통해 데이터 센터부터 온디비아스 AI까지 아우르는 AI반도체 생태계 구축 추진

  * '시위완 590'은 총연산성능(TPP) 기준으로 엔비디아 A100 성능의 90% 정도 되는 제품으로 평가

  – 클라우드, 자율주행·로보틱스, 스마트기기·온디바이스 등 전 분야의 AI칩을 개발하고 있으며 데이터센터-엣지·온디바이스 AI로 이어지는 추론 전용 아키텍처 확장 전략 추구

- 알리바바, 텐센트 등 중국 빅테크와 딥시크와 같은 AI 기업을 주요 고객으로 확보하며 데이터센터용 AI반도체로 사업을 빠르게 확대

  – 특히, '25년 AI 칩 수요 급증으로 상반기 매출이 전년 동기 대비 4,438% 급증(약 126억원 → 5,600억원)하는 등 최근 가파르게 성장 중

  * 캠브리콘 순이익 : ('24.上) 1,000억원 적자 → ('25.上) 2,000억원 흑자

- AI 칩 산업의 경쟁 구도가 단순한 컴퓨터 성능 중심에서 생태계 독립으로 이동함에 따라, 중국형 CUDA인 '뉴웨어'를 완성하여 중국 AI반도체 자립화의 대표적인 기업으로 부상

  – 중국 내 공급망 자립과 소프트웨어 생태계 강화에 핵심적인 역할을 수행 할 예정

3 시사점

AI반도체 분야의 핵심 역량 확보를 위해 민·관 협력 강화와 R&D 투자를 확대하고, 실질적인 세제 혜택을 제공하는 등 다각적인 지원 정책 필요

- AI반도체 경쟁에서 기술 우위를 확보하기 위해 기초연구부터 상용화까지 전 주기를 아우르는 R&D 연계 체계를 형성하여 지속 가능한 성장 동력 마련
  - 동시에 고성능 컴퓨팅 인프라를 확충하여 민간의 R&D 속도를 극대화하고, 국산 AI반도체 실증과 양산을 지원하여 글로벌 시장 진출을 위한 발판 마련
- AI 인프라 확충을 위해 데이터센터, 고성능 서버·네트워크 등 핵심 설비에 대한 세제 지원 대상 확대 모색

차세대 AI반도체 경쟁력 강화를 위해 신뢰받을 수 있는 수준의 실증 기반과 고도화된 신뢰성 검증 체계 구축 방안 모색

- 실제 산업 환경 기반의 테스트베드와 실증 인프라를 구축하여 성능·안정성 등을 체계적으로 평가·검증하고, 실증-인증-사업화로 이어지는 전 주기 지원 체계 확립
  - 국내·외 시장 진출에 요구되는 신뢰성 있는 레퍼런스 확보를 지원하고, 이를 활용한 사업화 연계 추진
- 글로벌 팹리스·빅테크 기업들과의 협력 네트워크를 공고히 하여 실질적인 수출 판로를 개척하고, 글로벌 AI반도체 공급망 내 핵심 국가로 도약

AI반도체 독자적 생태계 강화를 위한 미래 융합형 전문 인재 양성 체계 마련

- 핵심 인재 양성(석·박사 등)을 위한 AI반도체 특화 대학원 확대, 국제 공동연구 강화 등 실효성 있는 지원으로 연구 역량 향상 도모
- 기업 연계형 프로그램, 산·학 공동 랩 등을 확대 구축하여 산업 현장과 교육·연구 간의 연계를 강화하고, 차세대 AI반도체 산업을 선도할 전문 인재 배출 기반 구축
  - 실제 산업 현장에 즉시 투입 가능한 역량을 확보하고 국가 차원의 AI반도체 초격차 기술 경쟁력 향상

ICT SPOT ISSUE

☑ 발 행 일 : 2025년 12월 23일

☑ 저　　자 : 정보통신기획평가원 정책기획팀

☑ 발 행 인 : 홍진배(정보통신기획평가원장)

☑ 발 행 처 : 정보통신기획평가원

☑ 주　　소 : 대전광역시 유성구 유성대로 1548(화암동)

☑ 전　　화 : 042) 612-8001

☑ 홈페이지 : www.iitp.kr