Industry Focus

‘제3의 IT혁명 디바이스 시대’가 온다:

내 손 안의 AI, 온디바이스 AI(On-Device AI)

삼일PwC경영연구원

September 2024

pwc 삼일회계법인

들어가며

‘제3의 IT 혁명’을 탑재한 디바이스 시대 도래

- 기술의 발전은 인류의 삶에 수많은 변화를 가져왔음. 1800년 전기 배터리의 발명이나 1879년 백열전구의 발명과 함께, 대표적으로 세기의 발명이라고 불리는 것이 ‘인터넷’의 발명임. 한국은 미국 다음으로 두번째로 인터넷 연결에 성공한 국가로서, 1990년대 후반 부터는 한국 전역에 PC와 초고속 인터넷이 보급되었음. **인터넷은 첫번째 IT 혁명 디바이스라고 불릴 수 있는 PC(개인용 컴퓨터)를 대중화시켰음.**

- **2007년 이후 기술 혁신이 가져온 생활의 변화의 중심에는 ‘스마트폰’과 ‘IoT’가 있었음.** 스티브 잡스의 아이폰으로 시작된 스마트폰 세상은 2010년 들어 더욱 빛나게 되었음. 모바일 기기를 중심으로 세계가 하나의 네트워크를 형성했기 때문임. **따라서 스마트폰은 제2의 IT혁명 디바이스라고 할 수 있음.**

- **2022년말 이후 핫 이슈가 된 ‘생성형 AI’의 출현은 3번째 IT 혁명 디바이스라고 할 수 있는 AI를 탑재한 기기인 ‘온디바이스 AI’의 대중화 시대를 열어 나갈 것으로 기대되고 있음.** 온디바이스 AI는 AI 연산 칩을 기기에 내장하여 통신 연결 없이도 기기가 스스로 가벼운 AI를 학습, 연산을 수행하는 것을 말함 → **따라서 특정한 분야의 디바이스가 아닌 스마트 홈, 자동차산업, 특히 PC, 스마트폰, 웨어러블 기기까지 적용되고 있으며, 향후 의료, 금융서비스, 제조업 등 모든 산업분야로의 적용이 가능할 것으로 예상되고 있음. 그야말로 기존 디바이스 전체를 대체할 만한 파격적인 혁신으로 볼 수 있음.**

IT 혁명 디바이스: PC(인터넷) → 스마트폰(IoT) → 온디바이스 AI (AI 탑재 디바이스)

	인터넷 혁명기 (1980년대~)	모바일 혁명기 (2007~)	생성형 AI 혁명기 (2022~)
디바이스	PC	스마트폰	AI 탑재 ‘온디바이스’
기반기술 (개발내용)	인터넷 (Web 브라우저)	IoT (모바일 App)	생성형 AI (AI 채팅서비스 프로그램)
대표상품	인터넷 익스플로러 PC(개인용 컴퓨터)	아이폰	Chat GPT/ sLLM을 적용한 IT제품
영향	정보의 전달(확산)	모바일 시대	AI의 대중화/ 전산업의 AI화/ 기술과 일상의 융합
대표기업	IBM, 제록스 등	Apple, 삼성전자, Nokia, LG전자	Apple, 구글, 퀄컴, 삼성전자

Source: 언론종합, 삼일PwC경영연구원

Contents

1. 개요

- **2023년에 사용자의 요구에 맞춰 다양한 콘텐츠를 생성해주는 '생성형 AI(Generative AI)'가 뜨거운 화두를 이룬 것만큼, 이와 같은 최첨단 AI 기술을 본인의 손에서 직접 작동시켜 일상 생활에 더 깊이 적용시키고자 하는 소비자들의 수요가 점차 증가 중임**

- **이에 2024년은 개별 기기에서 AI 알고리즘이 작동하는 '온디바이스 AI(On-device AI)' 시대가 열리는 해가 될 것으로 예상 → 즉, '내 손안의 AI' 시대가 열리고 있음**

  *참고로, 온디바이스 AI는 최근 등장한 신기술은 아님 (미국 반도체 기업 퀄컴(Qualcomm)의 경우, 사용자의 AI 경험 개선을 위해 이미 10년 이상 온디바이스 AI 기술을 연구해 옴)

  인공지능(AI) 및 생성형 AI에 대한 상세 설명은 삼일PwC경영연구원의 '미리보는 CES 2024'와 '생성형 AI를 활용한 비즈니스의 현주소' 보고서 참고

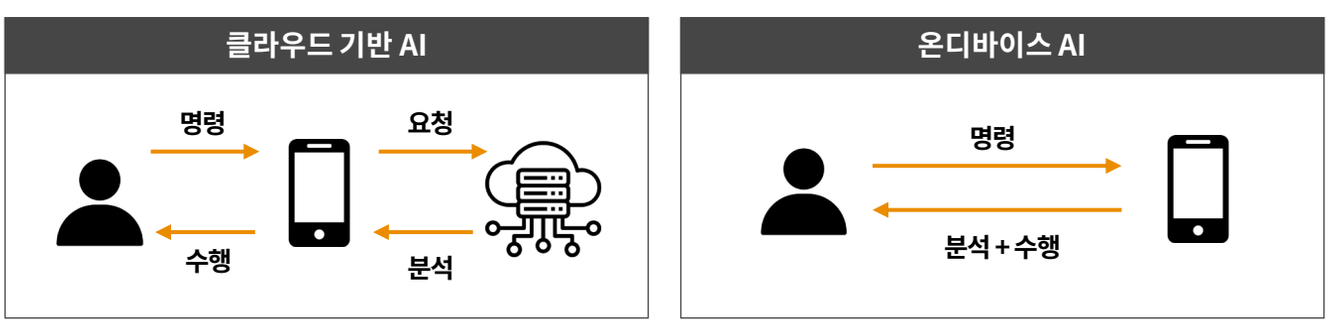
- **'엣지 AI(Edge AI)*주)', '타이니 AI(Tiny AI)'라고 불리기도 하는 온디바이스 AI는 말 그대로 스마트폰, 노트북, 자동차, 드론, 로봇 등 기기 자체에서 AI 기능을 수행하는 기술임**
  - 기존에 우리가 알고 있는 AI 기술은 기기에서 수집된 정보를 중앙 클라우드 서버(데이터 센터)로 전송해 정보를 분석하고 이를 다시 기기로 보내는 방식으로 진행되어 옴(=클라우드 기반 AI)
  - 이와 같이 데이터 센터를 중심으로 AI가 작동하는 방식이 보편적인 이유는, AI가 딥러닝(deep learning)을 구현하기 위해서는 ① 천문학적인 양의 데이터와 ② 이 데이터를 빠르게 처리할 수 있는 고성능의 컴퓨팅 역량이 필요하기 때문 → 이를 감당할 수 있는 대규모 네트워크가 필수적

    *주) 엣지 AI: 인터넷 연결 여부와 관계없이 AI를 엣지 컴퓨팅(사용자의 위치나 그 근처에서 컴퓨팅을 수행하는 것)과 결합해 사용하여 물리적 위치나 그 근처에서 데이터를 수집할 수 있도록 하는 것

- **온디바이스 AI는 멀리 떨어진 클라우드 서버를 거치지 않으며, 인터넷과 네트워크 연결 없이 기기 자체적으로 AI 모델을 수행하여 사용자 데이터를 수집하고 학습함**
  - '온디바이스 AI는 사물인터넷(IoT)에 생명을 불어넣는다' (앤드류 블레이크, Cambridge AI 센터장) → 기기들이 단순히 네트워크로 연결돼 AI 기능을 구현하는 데 그치지 않고, 자체 AI 수행을 통해 사용자 맞춤형 기능을 실행할 수 있음을 강조

- **온디바이스 AI는 향후 사용자의 일상 생활에 스며들어 사용자에 대한 정보를 직접적으로 파악할 수 있게 되면서 개별 기기의 사용자 맞춤형 서비스 강화에 기여할 것으로 예상**

클라우드 기반 AI vs. 온디바이스 AI 개념 이해



Source: 언론종합, 삼일PwC경영연구원

2. 특징

■ **생성형 AI**는 독창적 컨텐츠 생성 기능으로 많은 관심을 받았으나, **몇가지 단점이 존재: 클라우드 기반 데이터 송수신 과정에서 데이터 병목현상으로 인한 서비스 품질 저하, 고도화된 데이터 센터를 운영하는데 드는 막대한 비용, 완벽하지 않은 데이터 보안 기술, 에너지 소모량 심화 등이 있음**

→ 이러한 일부 단점 해결을 위해 최근에는 급증한 데이터 처리량 감소, 네트워크 지연 최소화, 데이터 보안 강화 등에 대한 대안으로 온디바이스 AI가 주목받게 됨

부상 배경

1	2	3	4	5
기술혁신(경량화&저전력 가능한 칩셋과 프로세서 개발) **과 인공지능 기술의 발달**	**데이터 보안**	**네트워크 연결 의존도 감소** (Offline 사용)	**엣지 컴퓨팅* 부상**	**하드웨어 발달** (웨어러블 기기, IoT 기기 등)

주*)엣지 컴퓨팅: 클라우드 컴퓨팅과 반대되는 개념으로, 인터넷이 아닌 로컬 장치(예: 스마트폰, 태블릿, IoT 장치 등)에서 데이터를 처리하는 기술

■ **온디바이스 AI의 주요 장점으로는 ① 빠른 서비스 제공, ② 데이터 보안 강화, ③ 자유로운 작동 환경, ④ 데이터 센터 운영비 및 에너지 소모량 절감, ⑤ 개인화된 AI 수요 대응 등이 있음**

- 이 외에도 온디바이스 AI는 대규모 하드웨어 인프라가 불필요해 **유지보수 비용이 저렴**한 등, **비용과 시간 측면에서 매우 효율적인 기술로 평가**받고 있음
- 다만, 클라우드 기반 AI 대비 **처리하는 데이터 양이 적어 산출되는 결과물의 완성도가 부족할 수 있다는 점이 단점**임

온디바이스 AI의 주요 장점

주요 장점	상세
빠른 서비스 제공	• 기기 자체적으로 AI를 실행할 수 있으므로 네트워크 지연이나 서버 부하의 영향을 받지 않아 응답시간이 빠르고 네트워크 지연 없이 빠른 서비스 수행이 가능
데이터 보안 강화	• 사용자의 데이터가 개인 기기 중심으로만 처리되기 때문에 데이터 유출의 리스크 감소 • 상황에 따라 필요한 데이터만 선별적으로 데이터센터에 전송하거나 민감한 부분을 사전에 제거하는 등 데이터 자체의 보안성 강화에 기여
자유로운 작동 환경	• 인터넷이 연결되어야만 작동하는 클라우드 기반 AI와는 달리, 오직 기기 내에서만 돌아가는 AI 기술이기 때문에 인터넷이 없는 환경에서도 실시간 번역 같은 작업이 가능
데이터 센터 운영비 및 에너지 소모량 절감	• 데이터 센터의 경우 조회 한 건당 비용이 인터넷 검색에 비해 10배 가량 높음 • 데이터 센터의 데이터 처리 부담을 크게 줄일 수 있어 운영 비용과 에너지 소모량을 효과적으로 줄일 수 있음
개인화된 AI 수요 대응	• 사용자와 접점을 이루는 기기(스마트폰, 태블릿 등)에서 AI 기술이 수행되기 때문에 기기가 작동하는 환경을 본질적으로 파악하여 사용자를 위한 보다 적절한 결정과 아웃풋을 내놓을 수 있음

Source: TRI, KETI, 삼일PwC경영연구원

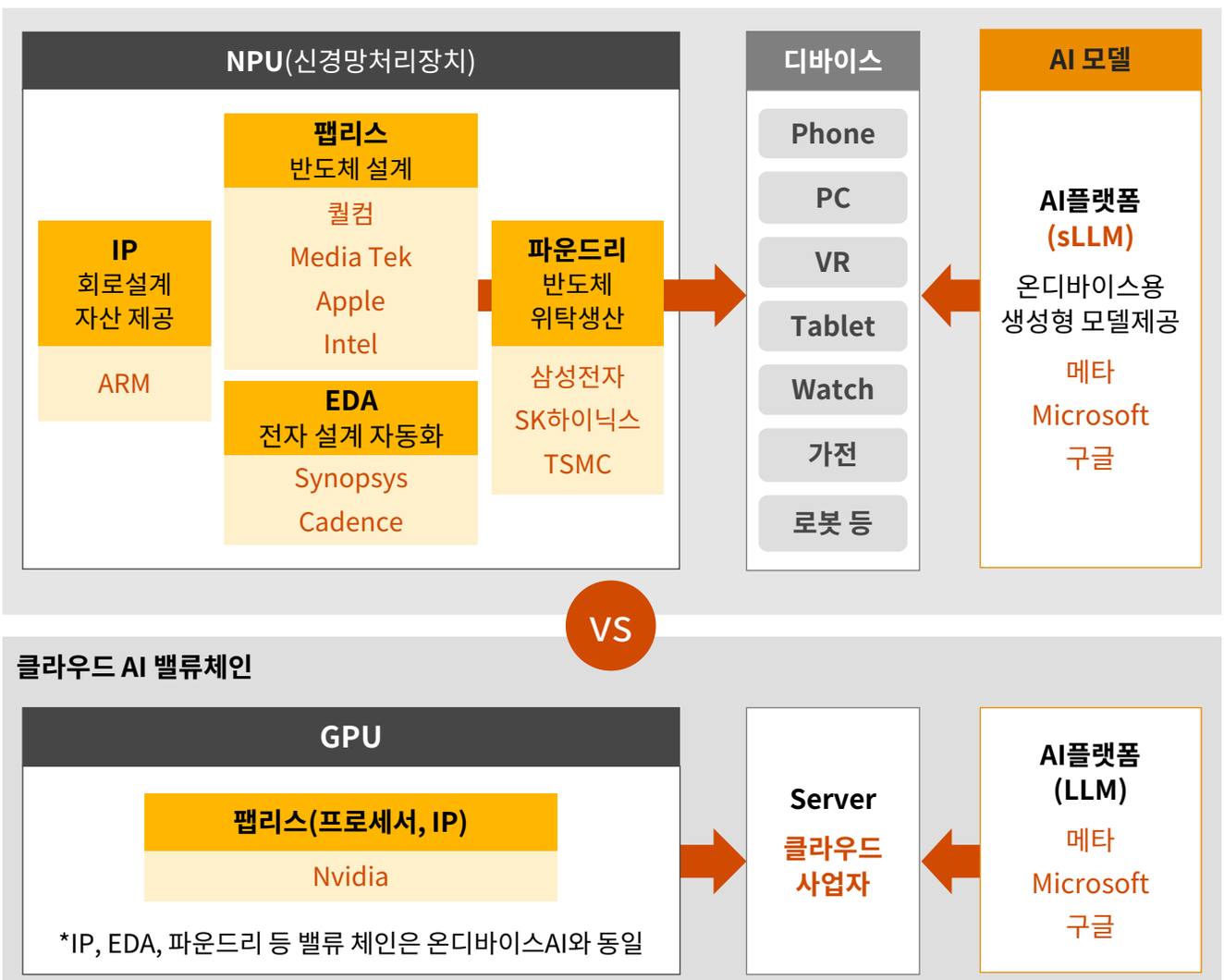
3. 온디바이스 AI 밸류 체인

- 앞에서 언급했듯이 2024년 이후 온디바이스 AI 시장은 급속도로 성장할 것으로 예상되고 있음. 스마트폰을 필두로 PC, 가전, 자동차, 보안, 헬스케어 등 실생활의 다양한 분야로 확산되며 **커스터마이징(Customizing)된 AI 칩 수요도 동시에 급증할 것으로 예상되는 등 온디바이스 AI 시장의 급성장은 관련 생태계 확장과 도약으로 이어질 전망**임.
  - **AI의 스마트폰 침투율 24년 9% → 27년 39%*, AI의 PC 침투율 24년 11% → 27년 53%*으로 확대 전망 등 24~25년에 시작된 온디바이스 AI 관련 제품들이 26~27년부터는 본격 상용화 될 전망임.**

    주*) 위 전망은 Techinsights (SA), 가트너, 카운터포인트, Canalys 4기관의 평균치

    → 따라서, 온디바이스 AI 생태계를 둘러싼 밸류체인의 확장은 당연한 결과임.
- 그 동안 AI 학습에 대한 기업 투자가 집중되어 엔비디아를 중심으로 한 상품들만 대거 출시되었다면, **향후에는 기술과 자본적 우위를 점하고 있는 클라우드 AI에 대한 투자와 더불어, 이제 막 개화하기 시작한 온디바이스 AI 기업에 대한 투자도 커질 것으로 예상됨. 대표적으로 AI반도체(Processing-in-Memory(PIM), 주문형 반도체 (ASIC: NPU))관련 회사에 대한 투자가 이에 해당함.**

온디바이스 AI 밸류 체인과 주요기업들



주) NPU: 전력 소모가 적고 AI 추론을 효율적으로 수행하는 전용 프로세서

Source: 미래에셋증권, 삼일PwC경영연구원

4-1. 온디바이스 AI 기술 구현의 필수 요건 – 'AI 반도체' 현황

■ **AI 기술이 확대되면서 함께 변화하고 있는 대표적 산업이 '반도체'임. 특히 AI 반도체 분야는 기존의 반도체 강자들 뿐 아니라 글로벌 빅테크 기업들까지 엄청난 투자와 M&A를 통해 경쟁력 확보를 하려고 노력 중임.**

정의 및 종류

■ **AI반도체는 인공지능 서비스 구현에 필요한 대규모 연산을 초고속·저전력으로 실행하는 비메모리 반도체를 말하며, AI의 핵심 두뇌에 해당함. 대표적인 AI 반도체로는 NPU**(신경망처리장치, Neural Processing Unit), **PIM**(processing in memory), **뉴로모픽**(neuromorphic, 인간 뇌의 구조와 기능을 모방해 설계된 기술 또는 시스템)등이 있음. 다만 뉴로모픽 반도체는 성능과 효율성은 뛰어나지만 범용성이 낮고 아직은 개발 중인 차세대 AI 반도체임.

- NPU의 형태: FPGA과 ASIC

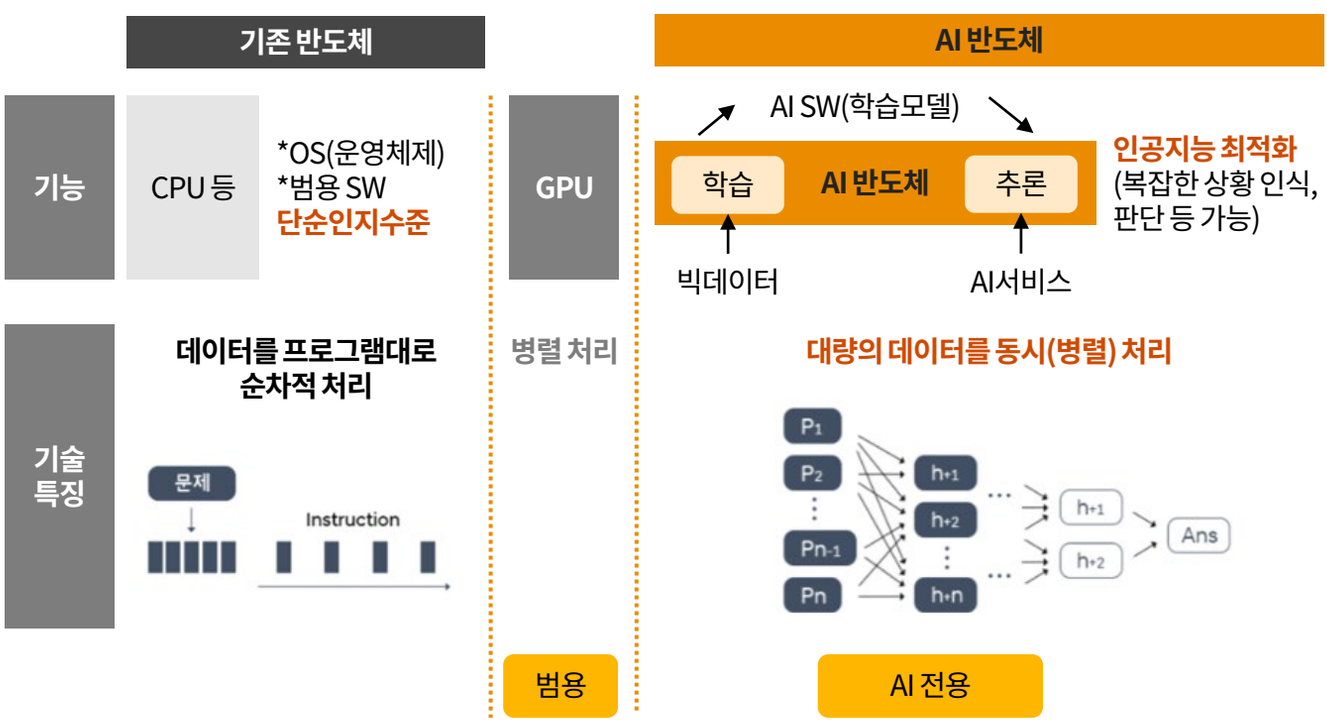
① FPGA(Field Programmable Gate Array): 칩 내부의 하드웨어를 목적에 따라 재프로그래밍이 가능해 유연성이 높음

② ASIC(Application Specific Integrated Circuit) 특정 목적에 맞춰 제작된 주문형 반도체로 효율이 높음.

등장 배경

■ AI 반도체가 개발되기 전에는 CPU(중앙처리장치)와 GPU(그래픽처리장치)가 핵심두뇌 역할을 담당했음. 다만 이 두가지는 AI를 처리할 수 있는 성능은 지녔으나 AI용으로 개발된 것이 아니기 때문에 AI 연산 외의 부분에 성능이 낭비되는 등 비용이나 전력소모 면에서 비효율적이었음. 높은 전략과 빠른 속도가 필수적이기 때문에 CPU.GPU 대비 범용성은 낮지만 AI 알고리즘에 최적화된 AI 전용 반도체가 등장하게 됨.

기존 반도체와 AI 반도체의 비교



Source: 과학기술정보통신부, 삼일PwC경영연구원

전망

- 21세기의 산업혁명이라고 불리우는 생성형 AI 시대의 필수품인 **AI 반도체는 생산성 개선과 비용 효율성 측면에서 전 산업 지형을 크게 변화시킬 것으로 예상되며 중장기적으로 반도체 수요의 구조적 상승을 견인할 것으로 예상**되고 있음. AI에 사용되는 반도체 중 CPU, GPU 시장은 이미 기술 성숙 단계 진입했으며, 최적화된 **저전력·고효율 ASIC 중심의 추론형 AI 반도체(NPU) 시장이 성장 중임.**
  - 그동안 메모리 반도체 비즈니스는 '같은 용량을 더 싸게 파는 비즈니스'가 핵심이었다면, **향후에는 용량 대비 가격이 비싸더라도, 더 큰 용량과 더 큰 대역폭을 제공하는 메모리가 중요해짐에 따라 HBM**(고대역 메모리, High Bandwidth Memory, 한마디로 고성능 D램)**의 수요는 더욱 증가할 것으로 예상**됨.
  - AI 반도체 발전 단계: 1세대(CPU, GPU) → **2세대 NPU(FPGA, ASIC)** → 3세대(뉴로모픽)
- **AI 반도체 수요가 급증하면서 24년 하반기 이후 반도체 공급부족 전망이 나오고 있음.** 최근 AI는 클라우드 서비스 제공사 뿐 아니라 제약, 금융, 법률, 유통, 제조 등 다양한 분야로 확산되며 각 산업마다 다양한 AI 모델 학습과 추론할 AI 반도체 수요를 급증시키고 있기 때문임.
  - 특히 메모리 반도체업체들은 HBM, DDR5 등 AI 반도체 중심의 신규 투자와 범용 반도체 생산라인의 선단 공정 전환 등 레거시 DRAM과 NAND 생산능력이 자연스럽게 축소되면서 공급부족 현상을 야기하고 있음.
- 또한 **엔비디아 vs. 反 엔비디아의 경쟁이 확대되면서 반도체 수요 급증을 가속화**하고 있음. 최근 구글, 인텔, 퀄컴, 삼성전자, ARM 등은 엔비디아 의존도를 줄이기 위해 S/W 기술 컨소시엄인 UXL(Unified Acceleration Foundation)을 구성해 One API라는 오픈소스 프로젝트 추진 중임.
  - 최근 MS와 아마존은 AI 전용 데이터 센터에 총 2,500억 달러(337조 원) 투자를 결정한 바 있음. 특히 전체 프로젝트 투자금액 중 50% 이상이 GPU, NPU, HBM, DRAM, NAND 구매에 사용될 전망임. MS나 아마존을 비롯한 구글, 메타, 애플 등 모든 글로벌 빅테크 기업들은 AI 전용 데이터센터 구축을 추진하고 있어 향후 원하는 시기에 충분한 AI 반도체를 확보하는 것이 무엇보다도 중요한 과제가 됨.
- 옴디아에 따르면, **AI반도체 시장은 데이터센터용과 온디바이스용으로** 2022년 411억 달러에서 2028년 1330억 달러 규모로 **CAGR 21.6% 성장**할 것으로 전망되고 있음. 또한 온디바이스용 AI 반도체는 반도체 기업과 수요기업들이 참여하며 스마트폰, 자동차 등 수요처별로 상이한 경쟁구도를 형성할 것으로 예상됨.
- AI 반도체는 일시적 유행이 아니라, 글로벌 산업 패러다임의 변화 과정이기 때문에 비교적 빠른 시일내에 미래 준비를 위해 급성장할 가능성이 높을 전망임. **→ AI 반도체는 최소 50년간 미래 먹거리가 될 전망**

수요처별 온디바이스 AI 반도체 주요 기업

구분(수요처)	AI 반도체 (제조) 주요 기업
스마트폰	애플, 삼성전자, 퀄컴, 엔비디아
PC	인텔, AMD, 퀄컴
자동차	인텔(모빌아이), 엔비디아, 퀄컴, 테슬라, Horizon Robotics(중), NXP(네), 인피니언(독)
스마트 TV	삼성전자, LG전자, 미디어텍
가정용 보안카메라	Ambarella(미), 퀄컴

Source: 언론 기사, 삼일PwC경영연구원

4-2. AI 반도체 하드웨어(HW) 기술 현황 – NPU

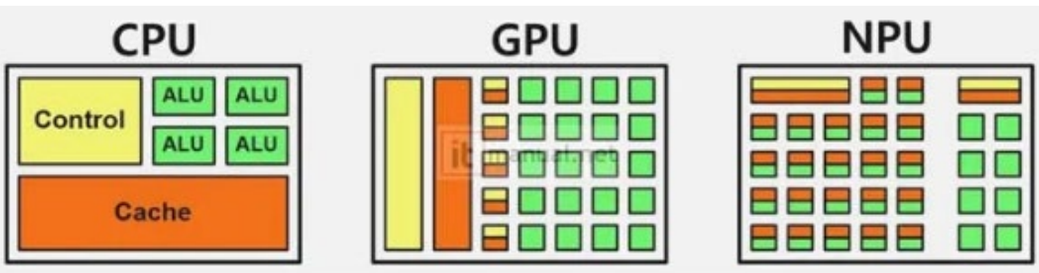
■ **NPU(Neural Processing Unit)은 추론용 AI 반도체로, 딥러닝에 적합한 연산들을 가속해주는 칩임. 한마디로 GPU에서 불필요한 기능을 모두 제거하고 인공지능 연산에 최적화된 기능만 집약한 칩**

- AI 모델에 광범위한 데이터를 학습시켜 범용성이 뛰어난 GPGPU(General Purpose Graphic Processing Unit) 칩 대비 '특수목적성'을 지님 → 자율주행, 이미지 생성 등 목적성이 있는 개별 AI 서비스에 최적화된 AI 학습이 가능
- 기존 데이터센터용 NPU 이외에 최근 들어 온디바이스 AI로 인해 더욱 각광을 받고 있음. 스마트폰, PC, AR/VR, 자동차 등의 소비재에 가속기를 장착해 AI 모델을 더욱 효율적으로 구동시킬 수 있도록 하는데, 이에 가장 적합한 반도체가 NPU이기 때문임.
- **다만 아직까지는 성능이 좋은 NPU를 디바이스에 탑재한다고 하여도 챗GPT와 같은 거대 AI 모델을 그 NPU 하나로 구동 시킬 수 없음.** 온디바이스 AI는 연산의 규모가 크지 않은 엣지 디바이스용 AI 서비스(동시 통역, 번역, 문서 작업 등)정도가 구동될 것임. 스마트폰 AP, PC용 칩에서 NPU의 기능이 강화되며 연산의 효율성이 극대화되기 때문에 각자가 휴대하고 있는 디바이스에서 더욱 빠른 AI서비스를 이용할 수 있게 되며 디바이스에 특화된 서비스를 제공받을 수 있게 되는 것임. 또한 면적 측면에서도 AI 가속기는 장점을 가지는데 GPU 대비 차지하는 면적이 훨씬 작고, 전력소모도 적기 때문에 배터리를 사용하는 엣지 디바이스에 더욱 적합함.

반도체 – CPU, GPU, GPGPU, NPU 구분

구분	특징 설명
CPU	• 컴퓨터의 중앙처리장치로 데이터를 고속으로 연산하고 처리 (=두뇌 역할) • 일반적인 계산, 시스템 작업, 사용자 인터페이스 등 다양한 작업을 순차적으로 처리
GPU	• **그래픽 처리 장치로 멀티미디어를 화면으로 출력하기 위한 그래픽 카드의 핵심 부품** • 게임, 영상 편집, 인공지능(AI), 데이터 마이닝 등 병렬 처리가 중요한 작업에 특화 • 다만, 애초에 AI 연산을 위해 만들어진 칩이 아니었기에 AI에 필요한 단순 연산만 수행
GPGPU	• CPU를 대신해 모든 데이터 연산 및 처리를 GPU를 통해 하는 범용 계산 장치 • GPU가 연산 처리 용도로 쓰이면서 기존에 CPU로 진행하던 대규모 데이터 처리와 AI 개발을 위한 기계학습, 딥러닝 분야 등에서까지 그래픽 카드가 쓰이기 시작
NPU	• GPU처럼 병렬 처리에 최적화된 구조로 되어있으나, AI를 개발하는데 필요한 제어 및 산술 논리 구성 요소를 갖춰 기계 학습, 심화 학습 알고리즘을 실행하는데 최적화 • **GPU에서 불필요한 기능을 모두 제거하고 AI 연산에 최적화된 기능만 집약한 칩** • 이에 GPU와 동일한 AI 작업 시, **GPU 대비 전력 소모는 적고 더 많은 결과물을 산출**

CPU, GPU, NPU의 구조



주) ALU: arithmetic and logical unit, 산술 논리 장치 / Cahe: 자주 사용하는 데이터나 값을 미리 복사해 놓는 임시 장소

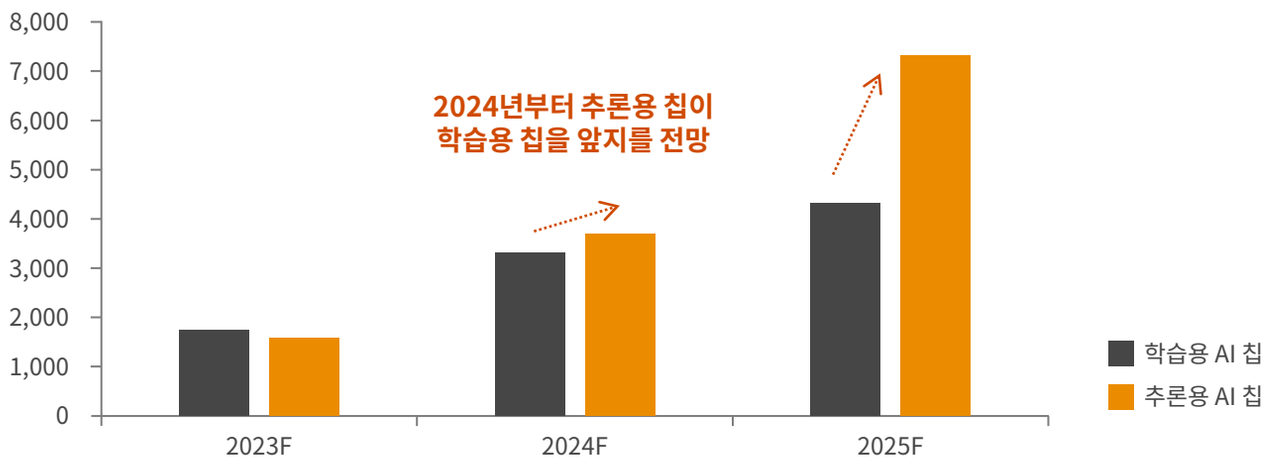
- **NPU 개발 역사는 GPGPU 대비 짧기 때문에 고도의 완성도를 지닌 NPU가 현재까지는 많지 않으나, NPU를 개별 AI서비스에 최적화하여 만든다면 학습과 추론면에서 GPGPU보다 우수할 것**

- 이제는 기업들이 너도나도 AI 서비스를 출시하는 시점에서 이들에게 중요한 것은 더 이상 학습용 반도체가 아닌 추론용 반도체임
  - 다양한 응용처에 대한 맞춤형 추론용 칩을 사용할 수록 서비스의 퀄리티가 좋아지기 때문

- **글로벌 빅테크 및 팹리스(반도체 설계) 회사들은 이미 자체적으로 NPU를 개발 및 출시하고 있으며, 전문 스타트업들의 경우 다른 AI 서비스 기업들의 NPU 수요에 대응하기 위해 NPU 설계 중**

추론용 AI 칩 vs. 학습용 AI 칩 출하 추이

(단위: K unit)



주) **학습용 AI 칩** – 인공지능 모델을 학습시키기 위해 방대한 데이터와 복잡한 수학적 연산을 빠르고 효율적으로 처리할 수 있는 성능이 필요하기 때문에 주로 범용성이 높은 형태로 설계되어 다양한 AI 모델에 적용 가능한 칩 (예: GPU, TPU)

추론용 AI 칩 – 학습된 인공지능 모델을 사용하여 새로운 입력에 대한 출력을 생성하기 위해 적은 데이터와 간단한 수학적 연산을 빠르고 저전력으로 처리할 수 있는 성능이 필요하기 때문에 특정 AI 모델에 특화된 칩 (예: 추론용 NPU)

Source: 현대차증권, 삼일PwC경영연구원

■ 국내외 주요 빅테크 및 반도체 기업들은 풍부한 자금을 기반으로 자체 NPU를 개발 및 출시 중

국내외 주요 빅테크 · 팹리스 기업들의 NPU 개발 현황

기업명	국가	AI 칩	상세
삼성전자	한국	엑시노스 2400	• ’23년 10월 기존 대비 NPU 성능이 14.7배, CPU 성능은 1.7배 향상된 스마트폰용 AI 칩 공개 • 향상된 AP(Application Processor)를 통해 갤럭시 시리즈에 탑재되는 ‘갤럭시 AI’는 실시간 통역 기능을 제공 • 새로운 AI TV ‘Neo QLED 8K’에 스크린 전용 NPU 탑재
LG전자	한국	LG8111	• 가전제품용 AI 칩으로, ’23년 9월, 유럽 최대 가전 박람회 IFA에서 처음으로 소개된 ‘LG 디오스 오브제컬렉션 무드업’에 탑재된 바 있음 • LG전자 시스템 반도체 조직인 SIC 센터는 현재 온디바이스 2세대 NPU 반도체 설계자산(IP)을 개발 중
애플 (Apple)	미국	M2	• CPU와 GPU, NPU, 메모리(RAM)등을 하나의 칩으로 통합한 고성능 시스템반도체 ‘M1’을 독자적으로 개발 • ’22년 2세대 5나노미터 기술과 200억 개 트랜지스터를 사용해 제작한 차기 버전 ‘M2’ 선보임(전작 M1 대비 40% 향상된 성능 보유)
구글 (Google)	미국	TPU v4	• ’23년 4월 4세대 AI 반도체인 ‘TPU(Tensor Processing Units) v4’를 공개 (*TPU도 일종의 NPU) • 기계학습 성능에 있어 종전 3세대보다 10배 이상 뛰어나고 에너지 효율 2~3배 높은 것으로 알려짐
메타 (Meta)	미국	MTIA	• ’23년 5월 AI 프로그램 구동을 위해 설계된 1세대 맞춤형 실리콘 '메타 트레이닝 및 추론 가속기(MTIA)'를 발표 • 메타의 AI 프레임워크인 '파이토치(PyTorch)'를 사용해 최적화된 소프트웨어를 실행
퀄컴 (Qualcomm)	미국	Snapdragon8 Gen 3	• 직전 2세대 제품보다 NPU 성능을 98% 개선한 스마트폰용 AI 칩으로, 메타(Meta)의 생성형 AI인’ Llama-2’를 지원할 계획 • ’23년 10월 새로운 AI PC용 칩인 ‘XElite’를 소개했으며, 탑재한 코파일럿+ PC 22종 공개
인텔 (Intel)	미국	Meteor Lake	• ’23년 12월 최초로 NPU 탑재된 14세대 노트북용 코어 울트라 칩 ‘메테오레이크(Meteor Lake)’ 출시 • Meteor Lake에는 NPU 외에도 인텔의 ‘인텔 4’(7nm급 공정)와 ‘FOVEROS’ (3D 반도체 적층 기술 등이 사용되어 전력 효율성과 그래픽 성능을 대폭 향상시킴
AMD	영국	Ryzen 8040	• 노트북용 프로세서로, 이전 세대 모델 대비 64% 더 빠른 비디오 편집과, 37% 빠른 3D 렌더링을 지원 • 에이수스, 델테크놀로지스, HP, 레노버 등 주요 노트북 제조사들이 ‘Ryzen 8040’ 탑재 노트북 공급 중
엔비디아 (Nvidia)	미국	GeForce RTX 40 (고성능 GPU)	• ’22년 10월 RTX 40 시리즈 3종 공개(RTX 4070 슈퍼, RTX 4070 Ti 슈퍼, RTX 4080), 코어 수 증가 및 메모리 입출력 속도 강화 • RTX GPU를 내장한 노트북이 레노버, HP, 델, 에이수스, 삼성전자 등 주요 제조사를 통해 출시
테슬라 (Tesla)	미국	D1	• 자율주행 수준을 높이기 위해 최상의 AI 학습 성능을 보유한 슈퍼-컴퓨터 ‘Dojo’에서 대용량 정보처리를 담당할 NPU 칩 ‘D1’ 개발
미디어텍 (MediaTek)	대만	Dimensity 9300	• ’23년 11월 자체 개발한 NPU ‘APU(AI Processing Unit) 790’을 적용한 ‘Dimensity 9300’ 발표 • 최대 330억 개의 매개변수 처리가 가능, 생성형 AI 성능을 8배 개선함

Source: 언론종합, 각 사, 삼일PwC경영연구원

■ **해외에서는 미국을 필두로 유럽, 중국 등 여러 반도체 칩 전문 스타트업에서 NPU 또는 그에 상응하는 AI 반도체 개발에 나서는 중**

■ **국내 중소기업 및 스타트업의 경우 자율주행, IoT, 금융 등 응용 분야에 특화된 AI 반도체를 개발 및 상용화 추진 중**

해외 주요 스타트업들의 NPU 개발 현황

기업명	국가	상세
Cerebras Systems	미국	• '21년 1세대보다 두 배 이상의 성능을 자랑하는 2세대 AI 프로세서인 'WSE-2'를 공개 (85만개의 AI 연산유닛, 20 PB/s에 달하는 메모리 대역폭 등 보유) • 이를 기반으로 한 'CS"라는 AI PC를 공개한 바 있음
Kneron	미국	• '19년에 자체 개발한 NPU를 포함한 엣지 AI 칩셋인 'KL520' 발표 • '21년에는 트랜스포머 네트워크와 4-bit INT를 추가로 지원하는 'KL530'을 발표
Groq	미국	• '20년 높은 플랫폼 유연성을 지닌 AI 반도체인 'TSP'를 발표 • 기존 경쟁사들 대비 최대 10배 이상 빠른 '820 TOPS(Trillion Operation Per Second, 초당 수백 테라 작업)' 성능을 확보
Sambanova	미국	• TSMC 7nm 공정 기반 300mb 이상 On-Chip 메모리, 300 TFLOPS(BF16) 이상의 연산속도를 갖는 데이터센터용 AI 반도체 생산 • AI 학습과 추론 연산을 지원하고, AI 연산 시 칩 내 데이터 흐름을 개선, 학습 모델에 따른 재구성이 가능하며, 복합 작업 지원 • 소프트뱅크, 삼성전자, 블랙락, 구글, 인텔, SKT 등으로부터 누적투자금액 20억 달러(2조 4천억 원)를 돌파하며 동종 업계 최고 수준의 투자유치 달성한 바 있음
GraphCore	영국	• 여타 AI 반도체와 달리 MIMD(mutliple instruction multiple data) 명령어에 기반 • 3차원 패키징 기술을 도입하여 더 높은 성능 및 전력효율 향상을 도모하는 AI 연산 프로세서인 'BOW IPU(Intelligence Processing Unit, 지능형처리장치)'를 독자적으로 설계 • IPU는 CPU나 GPU, NPU와 조금 다르게 프로세서에 직접 메모리 블록을 대칭한 것이 가장 큰 특징; 코어 당 메모리를 타일형태로 배치해서 지연을 줄임
Syntiant	미국	• '21년 IoT나 스마트폰을 대상으로 한 AI 반도체인 'NDP120'를 발표 • 이는 대표적인 딥러닝 모델인 CNN(Convolutional Neural Network, 합성곱 신경망)과 RNN(Recurrent Neural Network, 순환신경망)을 구동
Mythic	미국	• '20년 제조, 영상, 스마트 홈, AR/VR, 드론 등 다양한 영역에 온디바이스 AI를 배치하는 걸 목표로 'M1108' 아날로그 매트릭스 프로세서(AMP)를 발표
Blaize	미국	• '21년 엣지 컴퓨팅 프로세서인 'Pathfinder P1600'을 발표 • 16개의 인공지능 연산프로세서를 이용해 16 TOPS의 성능을 구현
Enflame Technology	중국	• '18년에 설립된 스타트업으로, AI 반도체 'DTU(Deep Thinking Unit)'를 개발 • DTU 기반의 20TFLOPS 성능 AI 가속기 솔루션 '클라우드블레이져(CloudBlazer) T10'을 출시
Horizon Robotics	중국	• 스마트 모빌리티, 감시 카메라 등 스마트 디바이스에 탑재되는 AI 반도체를 개발하고, 이를 통한 자율주행, 딜리버리 로봇 등 솔루션 개발

Source: 언론종합, KISTEP, 삼일PwC경영연구원

국내 주요 중소기업 및 스타트업들의 NPU 개발 현황

기업명	상세
사피온	• SKT로부터 분사梗립하였으며, ’20년 6.7kFPS/60W의 추론용 AI 반도체 ‘Sapeon X220’을 개발; MLPerf(Maching Learning Performance) 벤치마킹 테스트에서 상용화 등급(Available) 인정 • SKT의 5G 인공지능 서비스 및 자율주행차 등 상용화 추진 중
퓨리오사 AI	• ’21년 엔비디아(Nvidia)의 ‘T4’(GPU) 대비 4배 성능의 영상인식에 특화된 NPU ‘워보이(Warboy)’를 개발했으며, 해당 NPU가 탑재된 AI가속카드를 본격 양산 시작 • ’23년 GPT-3와 같은 자연어 처리를 위한 거대 인공신경망용 반도체 출시 계획 • 워보이 가속카드는 카카오엔터프라이즈에 탑재돼 성능을 입증 중
리벨리온	• ’21년 금융에 특화된 NPU인 아이온 칩 발표 • 실시간 트레이딩과 같이 빠른 처리속도가 중요한 금융분야 AI 응용에서 빠른 연산 속도와 경쟁제품 대비 10% 낮은 전력 소모량을 자랑
칩스앤미디어	• 비디오를 목적으로 한 반도체 설계 자산(Silicon Intellectual Property, SIP) 전문업체 • 고화질 영상전용 NPU 개발했으며, 이는 저화질 영상을 고화질 영상으로 업스케일링(Upscaling) 해주는 슈퍼 레졸루션(Super Resolution) 등 다양한 AI 기반 영상처리 알고리즘이 적용됨 • 딥러닝 기반의 알고리즘을 통해 8K이하 영상을 8K TV에서 고화질로 구현하는 특화 기술로 2020년 첫 매출 이후 3개년 연속 라이선스한 검증된 AI 반도체 IP
가온칩스	• ’12년 설립된 시스템 반도체 전문 디자인 솔루션(디자인하우스) 기업 • 영국 반도체 기업 ARM의 공식 디자인 파트너사로 ARM과 NPU개발 • ’21년 6월~12월까지 ARM과 CPU, GPU, NPU, Hardening Platform을 개발 • 자체 개발한 AI 반도체는 자율주행 차량용 AI 가속기, AI CCTV 등 기존 응용처에 융합되어 사용되고 있음
텔레칩스	• 차량용반도체 전문 팹리스 업체로, 미드, 엔트리 시장을 목표로 차량 내 인포테인먼트 AP(애플리케이션프로세서)인 '돌핀' 시리즈를 국내외 고객사에 공급 중 • 자율주행의 핵심 요소인 ADAS(첨단운전자지원시스템)용 NPU를 개발하여 상용화 중
오픈엣지테크-놀로지	• AI 반도체 설계에 필요한 IP와 솔루션을 공급하는 업체 • 기존 제품 대비 최소 4배 개선된 성능을 자랑하는 고성능 신경망처리장치(NPU), 'ENLIGHT PRO (인라이트 프로)'를 출시
넥스트칩	• 차량용 반도체 전문 팹리스 업체로, NPU를 적용한 ‘아파치5’ 후속 제품인 ADAS용 칩인 ‘아파치6’ 제품 출시 • ‘아파치6’을 탑재한 차량은 특정 거리를 스스로 주행하는 레벨3 자율주행 기능을 갖춤
딥엑스	• ’22년 온디바이스, 자율주행차, 데이터센터 등 각 어플리케이션에 특화된 NPU인 ‘제네시스(GENESIS)’ 개발 • 테슬라 NPU 대비 5배 이상의 연산 성능으로 10 TOPS의 우수한 전력 효율 보유 • 1세대 제품(모델명 DX-M1) 양산에 돌입, 가온칩스와 양산 계약 체결
디퍼아이	• 팹리스 기업으로, NPU를 내장한 CCTV 및 로봇용의 AI 반도체 SoC 양산 • 다수의 AI 반도체 개발 특허를 보유하고 있으며 온디바이스 AI 반도체를 위한 NPU 기술을 독자 확보한 것으로 알려져 있음
한화뉴블라	• 한화임팩트의 시스템 반도체 계열사로, NPU반도체 설계 및 IP 개발에 집중 • ’23년 12월 가온칩스와 협력해 4nm 서버용 NPU 개발 계획을 발표했으며, 웨이퍼 한 장에 여러 팹리스의 반도체 시제품을 생산하는 서비스인 멀티프로젝트웨이퍼(MPW)와 웨이퍼 한 장당 하나의 반도체만을 생산하는 싱글런을 진행할 예정 • 한화그룹 내 에너지 계열사 타깃으로 저전력 서버용 NPU를 개발 중인 것으로 알려짐
에임퓨처	• IoT같은 작은 어플리케이션부터 오토모티브나 ADAS(첨단 운전자 보조 시스템)처럼 매우 높은 성능을 요구하는 시장까지 모두 적용할 수 있는 확장성, 여러 어플리케이션을 동시에 처리할 수 있고 미래 발생하는 것까지 대응가능한 유연성, 옵션의 다양성을 바탕으로 한 NPU 칩 개발 스타트업

Source: 언론종합, KISTEP, 삼일PwC경영연구원

4-3. AI 반도체 소프트웨어(SW) 기술 현황 – AI 경량화 기술

- **AI 경량화 기술 (또는 경량 LLM (sLLM))은 매개변수를 줄여 학습·운용 비용을 낮추고, 미세조정을 통해 정확도를 높인 기술**
- **매개변수가 적은 경량화 모델은 기존 대규모언어모델 대비 낮은 연산 성능으로도 구동이 가능하여 기업의 고성능 하드웨어 기반 인프라 구축 부담을 줄이고 운용 비용을 낮출 수 있는 것이 장점임**
- **스마트폰과 같은 제한된 성능과 공간에서 AI를 제대로 구동하려면 모델 자체가 작거나, 큰 모델을 가볍게 만들 필요가 있기 때문에 'AI 경량화' 기술이 필수적**
- **해당 기술은 크게 2가지 방향으로 구분될 수 있음: ① 온디바이스 환경 전용 경량화된 AI 모델 및 추론 기술 개발과, ② 기존 AI 모델을 경량화하는 기술로 나뉨**
  - ① 경량 AI 모델(모델 자체를 작게 만드는 것): 기존 모델 대비 효율을 극대화하는 것으로, 딥러닝의 구조적 한계를 개선하고자 알고리즘 구조를 전면 수정하여 작게 만드는 것 → 특정 영역에서 성능이 좋고 비용 효율성이 높은 게 특징
  - ② AI 모델 경량화 기술(큰 모델을 가볍게 만드는 것): 만들어진 모델의 파라미터의 크기를 줄이는 것이 주 목적으로, 모델 압축(Model Compression), 지식 증류 등의 기법이 사용됨. 파라미터가 가지는 표현력을 최대한 유지하면서 불필요한 가중치를 없애기 위한 방법들이 있음

온디바이스 AI 소프트웨어 기술개발 동향

분류	접근방법	연구 방향
경량 AI 모델	모델 구조 변경	잔여 블록, 병목 구조, 밀집 블록 등 다양한 신규 계층 구조를 이용하여 파라 미터 축소 및 모델 성능을 개선
	합성곱 필터 변경	합성곱 신경망의 가장 큰 계산량을 요구하는 합성곱 필터의 연산을 효율적으로 감소
	자동 모델 탐색	특정 요소(지연시간, 에너지 소모 등)가 주어진 경우, 강화 학습을 통해 최적 모델을 자동 탐색
AI 모델 경량화	모델 압축	가중치 가지치기, 양자화/이진화, 가중치 공유 기법을 통해 파라미터의 불필요한 표현력을 감소
	지식 증류	학습된 기본 모델을 통해 새로운 모델의 생성 시 파라미터 값을 활용하여 학습시간을 줄이는 연구
	하드웨어 가속화	모바일 기기를 중심으로 NPU를 통해 추론 속도를 향상시키는 연구
	모델 압축 자동 탐색	알고리즘 경량화 연구 중 일반적인 모델 압축 기법을 적용한 강화 학습 기반의 최적 모델 자동 탐색 연구

Source: 언론종합, ETRI, 삼일PwC경영연구원

- **초거대 AI를 개발하고 운영하기 위해서는 막대한 양의 컴퓨팅 자원과 유지비가 필요하기에 기존에는 대기업만이 접근할 수 있는 영역으로 인식되었으나, AI 모델 경량화 기술을 통해 적은 비용으로도 보다 지속가능한 AI 모델 개발이 가능해져 관련 업계에서 상당한 수요를 얻을 것으로 예상**
- **다만 모델 크기가 압축됨에 따라 정확도가 다소 떨어질 수 있다는 장벽이 존재함을 고려해야 할 것**

■ **현재 국내외 주요 IT 기업 및 전문 스타트업에서 딥러닝 경량화에 대한 연구를 진행하고 있으며, 온디바이스 AI 실현을 위한 다양한 AI 경량화 기술을 내놓고 있음**

■ **각종 하드웨어에 얼마나 알맞게 경량화하고 최적화하느냐가 관건으로 작용할 것이며, 향후 더 다양한 분야에 AI가 제대로 활용이 되기 위해 경량화 기술에 대한 연구개발은 더 활발히 이루어질 것으로 예상됨**

국내외 AI 경량화 기술 개발 현황

기업명	국가	상세
삼성전자	한국	• '19년 '온 디바이스 AI(On-Device AI) 경량화 알고리즘' 공개 • 반도체가 특정 상황을 인식할 때 정확도는 유지하면서 기존 32비트로 표현되는 서버용 딥러닝 데이터를 4비트 이하로 낮출 수 있는 기술을 개발 • 학습을 거쳐 전체 정보 중 의미 있는 범위의 데이터만 양자화 해서 연산 속도를 높이고 동시에 전력 소모량을 낮춤
노타	한국	• AI 자동 경량화 플랫폼인 '넷츠프레소(NetsPresso)'를 개발 • AI 모델을 압축하여 연산량을 줄이면서 성능은 유지하는 방식으로 가지치기 기법, 양자화 기법 등 다양한 압축방법을 사용
앱스테이지	한국	• 자체 개발한 대규모언어모델(LLM) '솔라'를 구현할 때 작은 모델을 쪼개고 합치면서 최적의 성능을 내는 비율을 찾아냄 • 그 결과 107억 개 파라미터에 불과한 크기로 오픈소스 AI 모델의 글로벌 경연장 '허깅 페이스 리더보드'에서 1위를 차지한 바 있음
코난테크놀로지	한국	• 모델 크기를 줄이는 대신 학습량을 늘리거나 양질의 데이터만 학습시킴 • '23년 8월 자체 모델 '코난LLM'을 공개했으며, 해당 모델에 메타가 개발한 '라마2'보다 270배 많은 한국어를 투입했다고 알려짐
스퀴즈비츠	한국	• 양자화로 AI를 압축하는 기술을 개발했으며, 32자릿수 연산을 더 작은 단위의 연산으로 간단하게 표현해 빠르게 계산하면서도 똑같은 성능을 내는 원리
구글 (Google)	미국	• '23년 5월 매개변수가 수천억개인 '팜2(PaLM2)'를 공개했으며, 1년 전 버전인 'PaLM'보다 매개변수는 줄어든 반면, 학습 데이터 양은 5배 늘어남 • PaLM 2를 네가지 크기로 제공 - 활용 사례에 따라 작은 크기 순으로 '겍코(Gecko)', '오터(otter)', '바이슨(Bison)', '유니콘(Unicorn)'이 존재
메타 (Meta)	미국	• '23년 7월 자체 거대언어모델 '라마2(Llama2)'를 오픈소스로 공개한다고 발표 • '라마2'는 매개변수 규모에 따라 3가지 모델 – Llama 7B(70억 개), 13B(130억 개), 70B(700억 개)로 제공
엔비디아 (Nvidia)	미국	• 자사의 다양한 AI 하드웨어 기술을 기반으로 이들을 온디바이스 환경에서 활용할 수 있는 플랫폼인 'TensorRT'를 개발 • 이는 고성능 딥러닝 추론을 위한 소프트웨어 개발 키트(SDK)로, 임베디드 환경에서 학습된 모델을 기반으로 실시간 추론 및 모델 업데이트 기능을 제공
퀄컴 (Qualcomm)	미국	• 모바일에서 사용되는 비전 분석 기능을 최적화한 'FastCV' 라이브러리를 개발 • AI 모델 압축을 수행하여 온디바이스 AI를 지원하는 'AIMET(AI Model Efficiency Toolkit)'을 제공
노믹 AI (Nomic)	미국	• 일반 노트북에서 구동할 수 있는 경량화 모델 'GPT포올(GPT4ALL)'을 공개 • GPT-3.5를 이용해 질문과 답변을 수집하고, 'Llama 7B' 모델을 미세조정해 문서 작성이나 요약, 코드 작성 등 챗GPT의 주요 기능을 구현
알리바바 (Alibaba)	중국	• '19년 모바일 뉴럴 네트워크(MNN)라는 딥러닝 경량화 알고리즘을 선보임

Source: 언론종합, KETI, 삼일PwC경영연구원

4-4. 온디바이스 AI 적용 제품 현황

- **주요 글로벌 IT 업체들은 2023년부터 생성형 AI 경쟁을 본격화함에 이어, 자체 개발한 대규모언어모델(Large Language Model, 이하 LLM)을 탑재할 온디바이스 AI를 구현하고자 함**

- **국내에서 온디바이스 AI 기술을 선도하고자 빠른 행보에 나선 삼성전자의 경우, 2024년 1월 18일 공개한 갤럭시 S24시리즈에 자체 개발한 LLM '삼성 가우스(Samsung Gauss)'와 타기업의 AI 모델까지 온디바이스로 탑재한 것으로 알려짐**
  - 삼성전자는 스마트폰을 시작으로 스마트워치, 무선이어폰, 노트북 등 갤럭시 전 제품에 온디바이스 AI를 탑재할 계획이며, 이를 넘어 향후 확장현실(XR) 헤드셋으로도 확대 적용할 계획

- **구글의 경우 2023년 12월 6일 OpenAI의 생성형 AI 모델인 GPT-4에 대항할 '제미나이(Gemini)' 모델을 공개한 바 있으며, 가장 경량화 버전으로 온디바이스 AI용 '제미나이 나노(Gemini Nano)' 모델도 출시**

주요 IT 업체들의 대규모언어모델(LLM) 및 온디바이스 AI 개발 현황

기업명	국가	LLM명	LLM 및 온디바이스 AI 개발 현황
삼성전자	한국	가우스 (Gauss)	• 자체 LLM인 가우스는 언어모델, 코드모델, 이미지 모델 등 3가지 모델로 구성 • 이를 사내에서 활용 및 보완한 후 모바일에 탑재 • 가우스 외 OpenAI의 'GPT-4'나 구글의 'Gemini'가 갤럭시 S24 시리즈에 함께 탑재될 AI 후보로 유력하게 거론되는 중
구글 (Google)	미국	제미나이 나노 (Gemini Nano)	• 2가지 모델이 있으며, Gemini Nano-1은 18억 개, Nano-2는 32.5억 개의 파라미터를 쓴 가벼운 AI 모델 • 녹음기 요약, 스마트 채팅, 글과 그림의 활용에 집중될 것으로 예상 • '23년 10월에 Gemini Nano를 탑재하여 '픽셀 8 프로(Pixel 8 Pro)'라는 최초 AI 스마트폰을 출시한 바 있음(다만 픽셀의 세계 스마트폰 시장 점유율은 미비한 수준)
애플 (Apple)	미국	에이젝스 (Ajax)	• 공식 LLM 서비스 발표는 없음 • 다만 LLM R&D에 10억 달러를 투자할 예정이며, 애플이 자체 보유한 LLM 프레임워크 'Ajax'를 기반으로 하여 'AppleGPT(가칭)'을 올해 출시할 것으로 예상 → iOS 18에 탑재 예정
화웨이 (Huawei)	중국	팡구 (Pangu)	• 23년 8월, '화웨이 커넥트 2023'에서 자체 개발 차세대 OS인 'Harmony OS 4.0'에 'Pangu 3.0'(LLM) 탑재 예고 • 음성 비서 'Celia(Xiaoyi)'에 AI모델 탑재까지 완료, AI 성능을 실현시킬 하드웨어(반도체 칩 등) 개발이 핵심
샤오미 (Xiaomi)	중국	미LM (MiLM)	• '23년 4월 AI 모델 경량화를 위한 전담 LLM 연구팀을 설립, '23년 연간 200억 위안을 투자 • 모바일 비서 'Xiao AI'에 자체 LLM인 MiLM을 탑재, 초기 테스트 중 (월간 1.1억 명 사용자)
오포 (Oppo)	중국	안데스 (AndesGPT)	• 8월초, Oppo는 자체 LLM인 'AndesGPT'를 기반으로 AI 비서인 'Xiaobu'를 외부 테스트 • AndesGPT의 모델은 매개변수 1억, 3억, 10억 개로 구분되며, 향후 스마트폰 등에 탑재될 예정
아너 (Honor)	중국	외부협력	• Shanghai MWC 당시 LLM 대응을 위해 인터넷 서비스 업체와 협업하고 있다고 언급 • 바이두(Baidu)의 'Ernie bot', 알리바바(Alibaba)의 'Tongyi Qianwen' 등이 후보군이 될 것으로 예상

Source: 하이투자증권, 언론종합, 삼일PwC경영연구원

5. 향후 전망 – 시장 전반

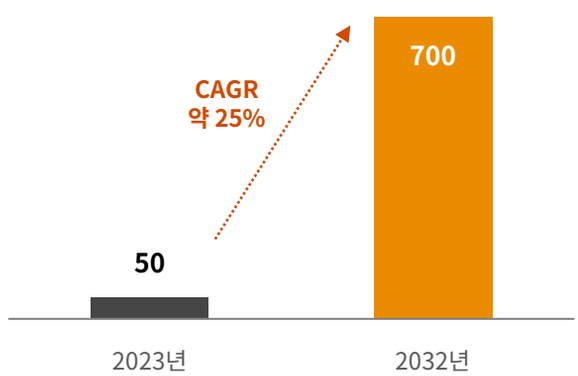
- **온디바이스 AI 시장은 AI 기술 향상 및 기기 적용의 본격화와 5G 네트워크 기술의 발전으로 향후 가파른 성장세를 보일 것으로 전망됨**
  - 글로벌 시장조사기관 GMI(Global Market Insight)에 따르면, 글로벌 온디바이스 AI 시장규모는 2023년 50억 달러(약 7조 원)에서 2032년 700억 달러(약 87조원)로 연평균 25% 가량 성장할 전망

- **또한, 올해부터 온디바이스 AI가 급격히 확대되며 2028년에는 전체 PC 시장의 80%, 스마트 폰 시장의 60%가 AI를 탑재할(AI PC, AI 스마트폰) 것으로 전망됨 (KB증권)**
  - 이는 IT 업체 인텔(Intel)이 2025년까지 AI PC 1억대 보급 계획과, 2027년 기준 AI 스마트폰 출하량이 5억대를 넘어서며 향후 4년간 누적 출하량이 11억대 달할 것으로 예상되기 때문

- **특히 온디바이스 AI의 성장에 따라 해당 기술에 필수적으로 탑재되어야 하는 추론용 AI 반도체인 신경망처리장치(Neural Processing Unit, 이하 NPU)와 온디바이스 기기의 생성형 AI 구현을 위한 메모리 반도체인 DRAM(Dynamic Random Access Memory)에 대한 수요도 급증할 것으로 전망됨**

글로벌 온디바이스 AI 시장 규모 전망

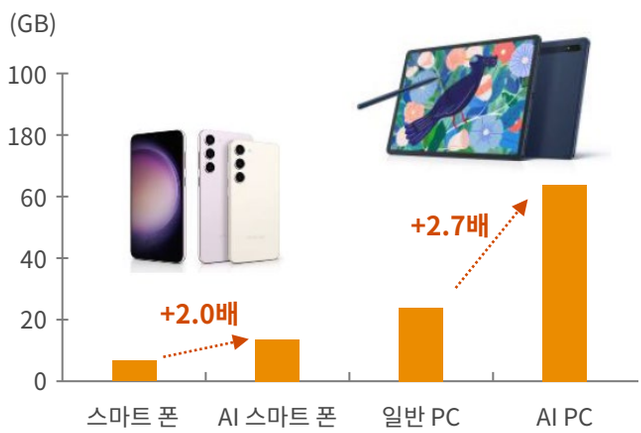
(단위: 억 달러)



Source: Global Market Insight, 삼일PwC경영연구원

온디바이스 AI 메모리 탑재량 2배 이상 증가(2028년 추정)

(GB)



('24년 1월 기준)

Source: KB증권, 삼일PwC경영연구원

5. 향후 전망 – 메모리 반도체 D램

- **온디바이스 AI 시장이 열리며 기존보다 더 작은 크기의 고성능과 저전력의 반도체가 주목받게 될 것**
- **AI 작업에 최적화된 NPU와 같은 시스템 반도체가 필수적으로 탑재될 것은 물론, 'LLW(Low Latency Wide, 저지연 광대역폭)' D램, 'LPDDR5T(Low Power Double Data Rate 5 Turbo)', 'LPCAMM (Low Power Compression Attached Memory Module)' 등의 고성능·저전력 메모리 반도체 D램 시장도 온디바이스 AI 활성화에 힘입어 향후 급성장할 것으로 전망됨**
- D램은 CPU로부터 전송된 데이터를 임시 보관하는 주기억장치 역할로, CPU의 성능 향상에 따라 D램 역시 데이터를 빠르게 처리해야 AI가 적용된 제품이 제대로 구현됨
  → 더 빠른 데이터 전송을 위한 대안으로 DDR(Double Data Rate), 휴대 전자기기가 등장하면서 LPDDR(Low-Power Double Data Rate) 등의 저전력 DDR로 발전하게 된 것
- **온디바이스 AI에 특화된 D램을 개발한 국내 주요 기업 중에는 삼성전자와 SK하이닉스를 꼽을 수 있음**

삼성전자, SK하이닉스 온디바이스 AI용 D램 개발 현황

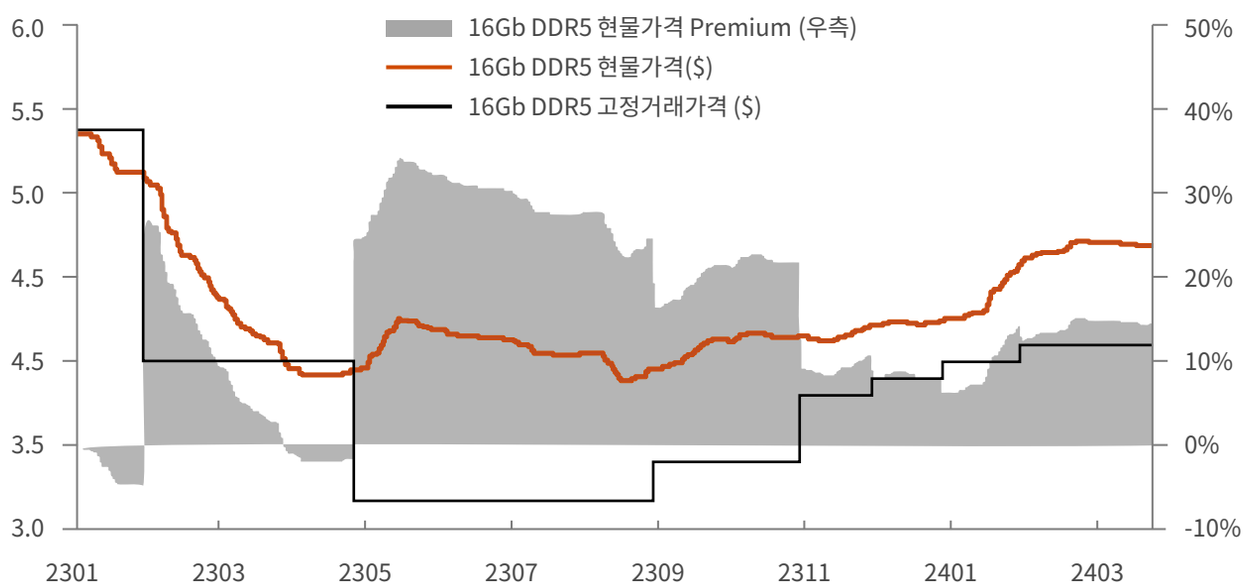
기업명	D램 개발 현황
삼성전자	• 올해 말 양산 목표로 'LLW' D램 개발 중 (LLW는 정보가 들어오고 나가는 통로 입출구(I/O)를 늘려 기존 모바일 제품인 'LPDDR' 대비 대역폭을 높인 특수 D램) • '23년 9월 개발한 PC·노트북용 D램 신제품 'LPCAMM'도 온디바이스 AI 맞춤형 제품으로 평가되고 있으며, 종전 제품(DDR) 대비 성능은 50%, 전력효율은 70% 높은 제품으로 알려짐 • 올해 삼성전자의 자체 개발 생성형 AI '삼성 가우스'의 모바일 제품 탑재를 시작으로 본격적인 시장 선점에 나설 것으로 전망
SK하이닉스	• '23년 11월 초고속 모바일용 D램인 'LPDDR5T'를 개발, '23년 11월 고객사에 공급까지 진행 • 해당 제품은 풀HD급 영화 15편 분량의 데이터를 1초 만에 처리하는 등, 모바일용 D램 가운데 데이터 처리 속도가 가장 빠름 • 애플(Apple)의 차세대 증강현실(AR) 디바이스 '비전프로'에 특수 D램을 공급하게 되었으며, 애플이 비전프로용으로 새롭게 개발한 'R1'이라는 칩과 연동해 실시간 고화질 영상을 처리를 지원

- **올해부터 온디바이스 AI 제품들의 서비스 확대가 예상되면서 서버에서 차지하는 D램 비중은 2023년 17%에서 2027년 38%로 4배가량 증가할 것으로 전망**

- **이에 2023년 글로벌 스마트폰 출하량이 10년 만에 최저치에 머물며 급락세를 보이던 모바일 D램의 가격이 2023년말부터 회복세를 보이고 있으며, 당분간 10% 내외의 추가 상승이 전망됨. 이는 온디바이스 AI가 적용된 스마트폰 신제품 출시 등이 맞물린 것도 한 몫 한 것으로 판단**

- **고부가 D램이 온디바이스 AI의 주요 솔루션으로 떠오르면서 '공급자 우위' 추세가 당분간 지속될 전망이며, 업계에서는 AI 메모리 반도체는 다양한 영역의 맞춤형 주문이 가능해 향후 수주형 비즈니스로 변할 것으로 전망**
  - 메모리 반도체 제조사들은 단가와 판매량을 동시에 끌어올리며 이익 극대화에 나설 것으로 예상

16GB DDR5 현물가격과 고정거래가격 차이



Source: DRAMeXchange, 삼일PwC경영연구원

5. 향후 전망– 하드웨어 기기

- **앞서 온디바이스 AI 기술이 발전함에 따라 PC, 스마트폰, 가전, 자동차 등 전 산업 분야에서 신규 수요를 창출할 것이라고 전망**

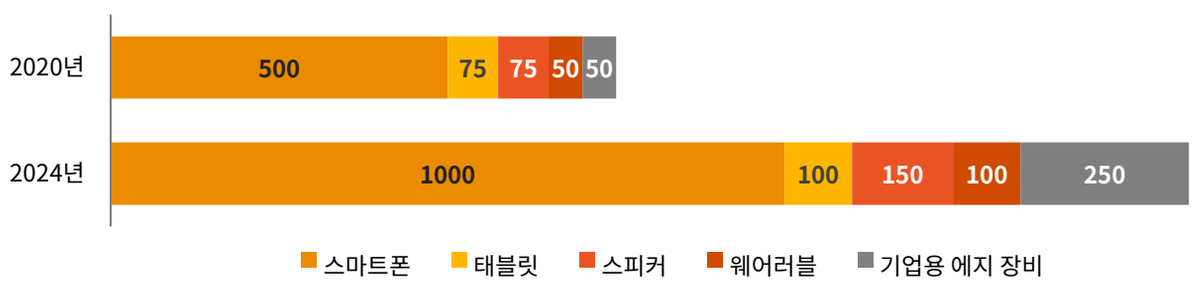
- **단기적으로는 '개인용 스마트폰'과 '개인용 PC' 시장이 더욱 진화되어 활성화될 전망(2028년 전체 PC 시장의 80%, 스마트 폰 시장의 60%가 온디바이스 AI 탑재 예상)**
  - 2025년까지 인텔(Intel)의 AI PC 1억대 보급 지원 계획 발표와 더불어 24년 상반기를 기점으로 전 세계 40개 PC 업체들이 인텔의 'Meteor Lake'를 탑재한 PC 250종을 출시할 것으로 예상되고 있음 ('27년 1억 7천만 대 육박할 전망)→ **전 세계 스마트폰 및 PC 출하량은 정체 상태에서 회복세로 전환 예상**

- **스마트폰과 PC외에도 온디바이스 AI는 XR 헤드셋, 로봇, 자율주행차, 드론 등 다양한 하드웨어 기기에 빠르게 적용될 것으로 예상됨에 따라, 하드웨어 경쟁력 강화 시도가 이어질 것**

  → 기기 특성에 맞는 최적화된 AI를 수행할 수 있는 역량을 키우기 위한 시도가 더욱 늘어날 전망이며, AI 기반 고사양 하드웨어 개발을 위한 경쟁이 더욱 치열해질 것으로 예상

주요 하드웨어 기기별 탑재된 온디바이스 AI 반도체 규모

(단위: 백만 개)



Source: MarketsandMarkets, DBR, 삼일PwC경영연구원

- 이에 온디바이스 AI 시대에 상당한 경쟁우위를 갖출 수 있는 player는 바로 이 다양한 IT 디바이스 시장을 이미 확보하여 상당한 수요 기반을 갖춘 업체일 것으로 판단

- **이에 업계에서는 현재 애플(Apple)이 온디바이스 AI 시대를 주도하는 주요 기업일 것으로 기대하고 있으며 삼성전자 역시 두각을 보이고 있음.**
  - 애플 및 삼성전자는 스마트폰, 태블릿, 노트북, 웨어러블까지 이미 다 갖추었기에 모든 영역에 AI 알고리즘을 접목시켜 AI 시장에서 유리한 입지를 갖출 수 있을 것으로 판단

    * 애플은 이미 다양한 엣지 AI(Edge AI) 애플리케이션을 아이폰과 그 외 주요 제품에 접목하기 위해 기술 개발을 추진 중인 것으로 알려짐

    * 삼성전자, '갤럭시 S24' 시리즈 공개 "세상과 소통하는 AI폰의 시작"
  - 이외 LG이노텍은 온디바이스 AI폰에 들어가는 고성능 카메라 모듈 공급이 확대됨에 따라 주목받고 있음

- **하드웨어 기기에 이어 소프트웨어, 칩, 서비스까지 수직 통합 기반의 시스템을 구축할 수 있는 player가 AI 시대의 진정한 강자가 될 수 있을 것**

6. 시사점 및 제언

시사점

✓ **2022년부터 LLM 중심으로 AI 대중화 시작**

✓ **2024년부터는 모든 기기에 AI가 탑재되는 온디바이스 AI 시대가 시작될 것임**

✓ **이에 따라 관련 산업의 폭발적 성장이 예상되는 바, 이에 대한 전방위적 준비가 필요해 보임**

1) 기기의 폭발적 대체 수요:
흑백 TV가 컬러 TV로 교체되듯이, 또는 휴대폰이 스마트폰으로 바뀌듯이, 모든 전자기기 및 가전 등에 AI가 탑재된 기기로 바뀌는 과정에서 모든 전자기기의 대체 수요가 빠른 속도로 증가하게 될 것으로 전망됨(스마트폰, PC, TV, 냉장고, 세탁기 등 가전제품, 자동차 등 모빌리티 수단 등)

2) 온디바이스 AI 전용 반도체: NPU, 뉴로모픽

3) AI 경량화 기술 또는 경량 AI 모델

4) 반도체 전반: DRAM 등 메모리 반도체, 팹리스, 파운드리 등
* 온디바이스 AI를 탑재한 디바이스의 형태 (Form factor) 변화는 메모리 반도체의 성장을 견인할 것으로 전망됨
- AI 반도체로 사용되는 '그래픽처리장치(GPU) + 고대역폭 메모리(HBM)'모델로 봤을 때, 메모리 반도체는 HBM이 성장을 견인할 것임. 하지만 궁극적으로는 인간의 뇌를 닮은 차세대 반도체, NPU로 귀결될 것으로 예상됨.

5) 배터리: 온디바이스 AI 적용을 위해서는 배터리의 경박 단소화가 중요한 이슈이기 때문에 배터리 에너지 밀도를 높이는 관련 기업들의 경쟁 치열 예상됨.

제언

1 온디바이스 AI 시장 본격 개화에 따라 관련 기술(생성형 AI, 엣지 기술, 고성능·저전력 AI 반도체 등) 보유업체에 대한 적극적인 투자 및 M&A 경쟁 불붙을 예정 → 기업들의 신속한 대응 방안 요구됨

- **몇몇의 주요 글로벌 업체들은 이미 온디바이스 AI 관련 업체들을 인수**
  - 인텔: '19년 12월 이스라엘 AI 반도체 스타트업 '하바나 랩스(Habana Labs)' 인수 (20억 달러)
  - 애플: '23년 12월, 저전력·고효율 딥러닝 알고리즘과 온디바이스 AI 처리 기술 개발한 프랑스의 AI 스타트업체 '데이터칼랩' 인수 → '24년 3월, 더 작고 빠르게 만드는 기술을 개발해 온 캐나다의 '다윈 AI' 인수
  - AMD: '22년 2월, AI 반도체(FPGA 전문) '자일링스(Xilinx) 인수' (490억 달러) → '23년 8월, 프랑스의 AI 추론 스타트업 '밈솔로지'를, 10월에는 소규모 오픈소스 컴파일러 업체인 '노드.ai' 인수 → '24.7월 핀란드 소재 유럽 최대의 민간 AI 연구소인 '사일로 AI' 인수(6억 6,500만 달러)
- **한편 국내 주요 업체들 역시 관련 업체들에 대한 투자 및 파트너십 진행**
  - **삼성전자: 국내 AI 반도체 스타트업 '크래프트테크놀로지스'와 '리벨리온'과 협업해 NPU 생산, '24.7월 지식 그래프 기술을 보유한 영국 스타트업 옥스퍼드 '시맨틱 테크놀로지스' 인수**
  - **KT: '22년 7월 AI 반도체 스타트업 '리벨리온'에 300억 원 투자**
  - **LG전자: '23년 5월 AI 반도체 스타트업 '텐스토렌트'와 AI 및 칩렛 기반 반도체 공동개발 발표**

2 기업의 주어진 데이터를 기반으로 질 높은 데이터(의미 있는 인사이트)를 확보하는 역량이 중요해질 것 → 기업의 데이터 경쟁력이 곧 비즈니스 성과와 직결될 것

- 온디바이스 AI는 기존의 대규모언어모델보다 작은 규모의 데이터를 기반으로 학습 및 연산하기 때문에 그만큼 가치 있는 데이터를 뽑아낼 수 있는 역량이 요구됨

3 온디바이스 AI의 한계점도 고려해 예상하지 못한 문제점들을 대응할 수 있는 역량 길러야 할 것 → 잠재적 리스크 사전에 파악하여 예방할 수 있는 역량 중요

- **제한된 학습 데이터 규모 및 컴퓨팅 능력으로 왜곡된 아웃풋을 내놓거나 오작동 발생 가능**
- **개발 기기에서 구동되기에 AI 기술 악용 위험을 제때 효과적으로 대응하지 못할 가능성 존재**

4 조만간 '앰비언트 컴퓨팅(Ambient Computing)*' 시대가 올 것 → 온디바이스 AI 기기들이 조화롭게 운용될 수 있는 연계 방안 요구됨

- 온디바이스 AI는 사용자가 놓여진 환경 속에 존재하는 다양한 기기들로 부터 수집된 정보를 바탕으로 사용자의 라이프스타일을 이해하고 사용자의 행동 흐름에 맞는 맞춤형 서비스를 선제적으로 제공할 것
- 온디바이스 AI의 역할이 보다 더 중요해질 것으로 예상됨에 따라 기기들 간의 컴퓨팅 충돌, 데이터 통일성 혼란 등의 이슈들이 생기지 않도록 대응 방안 필요

*주) 앰비언트 컴퓨팅은 인간의 직접적인 명령이나 개입 없이도 사용자 주변에 있는 장치들이 사용자가 필요로 하는 서비스를 제공하는 것. IoT와 기술면에서 유사하지만 앰비언트 컴퓨팅은 사용자 중심의 경험 제공에 초점이 맞춰진 반면, 사물인터넷은 다양한 기기들 간의 연결과 상호작용에 집중하고 있음.

Appendix

국내 온디바이스 AI 관련 주요 기술 및 지원 정책

■ **온디바이스 AI 기술은 아직 초기 단계에 있으나, AI 기술이 어디서나 만연하게 자리잡은 요즘 사회에 필요한 장점들을 지녀 향후 다양한 분야에서 활용될 것임**

- 안전한 자율주행 서비스와 같은 특수 목적의 기능이 요구되는 분야나, 네트워크 환경이 열악한 분야, 개인 정보 보안이 중요한 분야 등에서 특히 주목받을 것으로 판단됨

■ **온디바이스 AI 기술을 성공적으로 구현하기 위한 필수적인 하드웨어(HW)와 소프트웨어(SW) 기술은 각각 '지능형 반도체 기술'과 '경량형 AI 기술'이라고 볼 수 있음**

- 이들은 온디바이스 AI 기술 활성화에 따라 향후 성장세가 돋보이는 기술들로, 정부 차원에서 관련 기술 개발을 통한 경쟁력 제고를 위한 정책적 대응 방안을 제시 중

온디바이스 AI 구현을 위한 주요 HW·SW 기술

구분	기술	상세
하드웨어	신경망처리장치(NPU)	딥러닝에 적합한 연산들을 가속해주는 추론용 AI 반도체
소프트웨어	AI 모델 경량화	높은 연산량 성능은 유지한채 소형기기에 탑재 가능하도록 함

NPU 및 AI 모델 경량화 기술 경쟁력 강화를 위한 주요 정책

구분	발표 시기	정책 상세
NPU	'23. 06	• 'K-클라우드 프로젝트' 사업 본격화 * 국산 AI반도체를 개발해 이를 데이터센터에 공급하기 위한 정책으로, '23년부터 '30년까지 총 8,262억 원을 걸쳐 투자할 계획 * AI반도체를 신경망처리장치(NPU), 저전력 PIM(Processing-In-Memory), 극저전력 PIM 3단계에 걸쳐 고도화 할 예정이며 단계별(3단계)로 실증 사업 추진
	'22.12	• 'K-클라우드 추진방안' 발표 - 2030년까지 3단계에 걸쳐 국산 AI 반도체의 국내 데이터센터 시장 점유율을 80%까지 끌어올릴 계획 • 1단계: 2025년까지 국산 NPU의 국내 점유율을 23%까지 올림 → 2단계:'28년까지 D램 기반 PIM과 국산 NPU를 접합해 기술 우위를 가진 해외 GPU급 성능을 구현 → PIM 기술을 개발하여 '28년엔 중국을, '30년까진 미국을 따라잡고자 함
	'22. 06	• '제1차' 열어 'AI 반도체 산업 성장 지원대책'을 발표 • '27년까지 R&D에 총 1조200억원을 투입할 예정 • 주요 과제로는 AI 알고리즘 연산에 최적화한 차세대 NPU 개발, 연산과 저장기능을 통합하는 PIM 반도체 개발, NPU와 PIM 반도체 성능을 융합하는 초거대 AI 시스템 구축 등이 있음
AI 모델 경량화	'23.04	• '디지털플랫폼정부 실현계획 보고회' - '26년까지 2,655억 원 투입해 초거대 AI 기술의 논리적 리즈닝(인과관계 이해), 편향성 필터링, 모델 경량화·최적화 등을 구현하기 위한 기술 등에 집중할 계획 • 과학기술정보통신부는 서울대학교 '초거대 AI 모델 및 플랫폼 최적화(CHAMP) 센터'를 선도연구센터(ERC) 지원사업으로 선정하고 140억 5,000만 원을 지원 → 2026년까지 오픈AI의 GPT-3.5와 동일한 성능을 내면서도 규모는 100분의 1로 줄인 모델을 개발할 계획

Source: 언론종합, 삼일PwC경영연구원

Author contact

삼일PwC경영연구원

이 은 영 상무
eunyoung.lee@pwc.com

최 형 원 책임연구원
hyungwon.choi@pwc.com

신 서 윤 연구원
seoyoon.shin@pwc.com

삼일 PwC Business Research Center

최재영 원장
jaeyoung.j.choi@pwc.com

Business contact

Semiconductor/Display

Assurance

정재국 Partner
jae-kook.jung@pwc.com

남상우 Partner
sang-woo.nam@pwc.com

김경환 Partner
kyung-hwan.kim@pwc.com

Tax

이윤석 Partner
yoon-sok.lee@pwc.com

Deals

최창대 Partner
chang-dae.choi@pwc.com

장성욱 Partner
sung-wook.jang@pwc.com

문상철 Partner
sang-chul_1.moon@pwc.com

www.samil.com

S/N: 2408W-RP-049