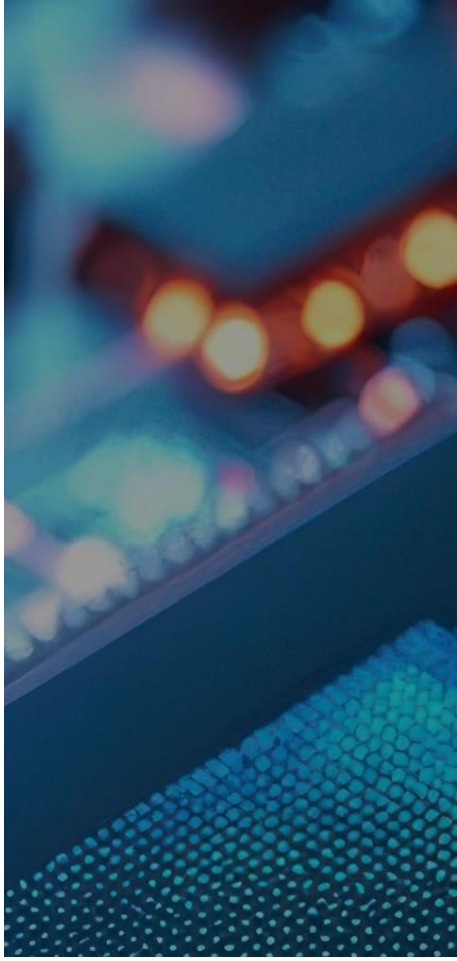
AI 시대를 견인하는 반도체 산업 전망



반도체 시장의 주요 동향과 성장 동력

December 2024

pwc

목차

Section 1

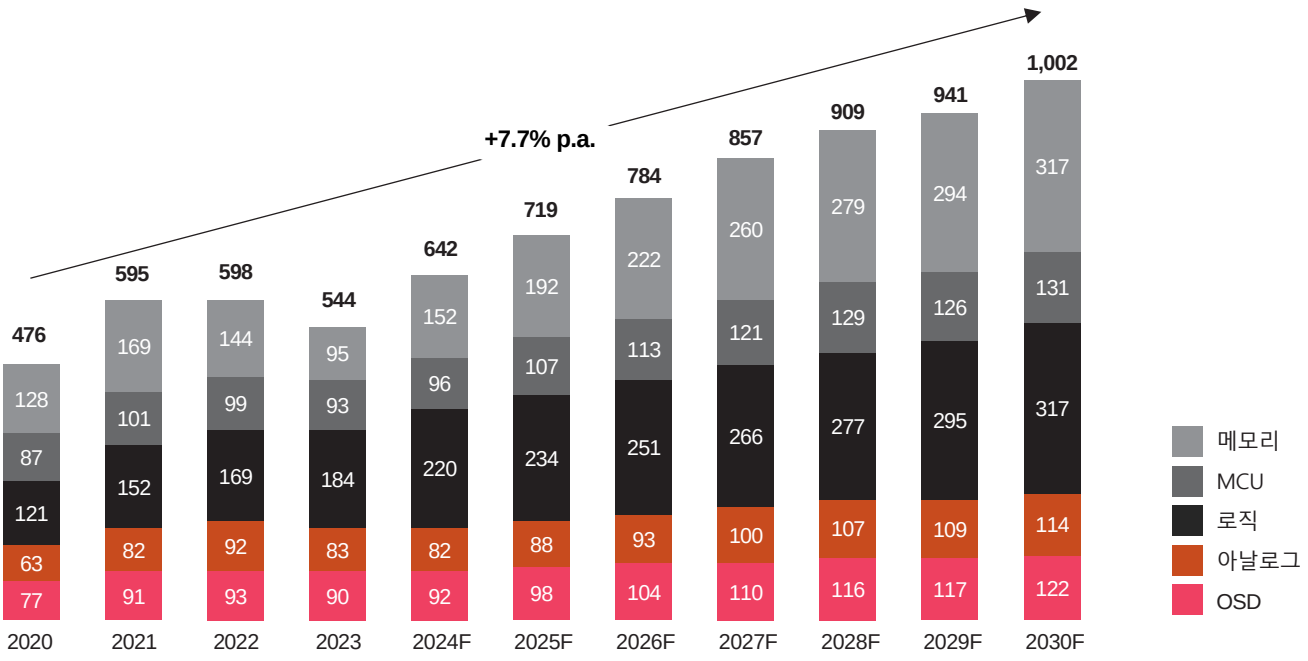
기술 필수재, 반도체 시장의 성장

반도체는 지난 70년 이상의 시간 동안 기술 혁신의 원동력으로써 다양한 산업 변화를 이끌어 왔습니다. 개인용 컴퓨터와 스마트폰에서 데이터 센터와 클라우드 컴퓨팅에 이르기까지 산업 발전은 반도체를 근간으로 하며, 앞으로도 반도체 산업은 전동화, 디지털화, 인공지능(AI) 및 사물 인터넷(IoT)과 같은 메가 트렌드를 선도하는 근간으로 지속 성장할 것입니다.

반도체 관련 시장 역시 2024년 6,420억 달러에서 2029년 1조 달러에 이를 것으로 전망됩니다. (도표 1참조)

도표 1

반도체 종류별 글로벌 시장 규모, 2020–2030 (단위: 십억 달러)



Source: Omdia Q3 2024; OSD – Optoelectronic, sensor and discrete

반도체 산업은 오랜 기간 글로벌 단위로 운영되어 왔습니다. 그러나 COVID-19 이후 지난 5년 동안 공급망 경직과 무역분쟁 증가로 인해 자국 내 반도체 생산시설 투자를 통한 반도체 공급망 주도권 확보 경쟁은 더욱 심해졌습니다.

COVID-19 기간 동안의 재택근무, 온라인 근무환경 구축 그리고 가전 제품 사용의 증가로 반도체 수요는 유례없이 급증하였습니다. 또한 전 산업에 걸친 AI와 IoT 기술의 도입으로 반도체 수요는 더욱 높아졌습니다. 이에 반해 일부 반도체 수요 급증에 대한 과소평가와 이로 인한 반도체 공급 업체들의 공급 능력 확장에 대한 보수적 태도로 2020년 하반기부터 2022년 말까지 전세계적으로 반도체 부족 현상이 발생하였습니다.

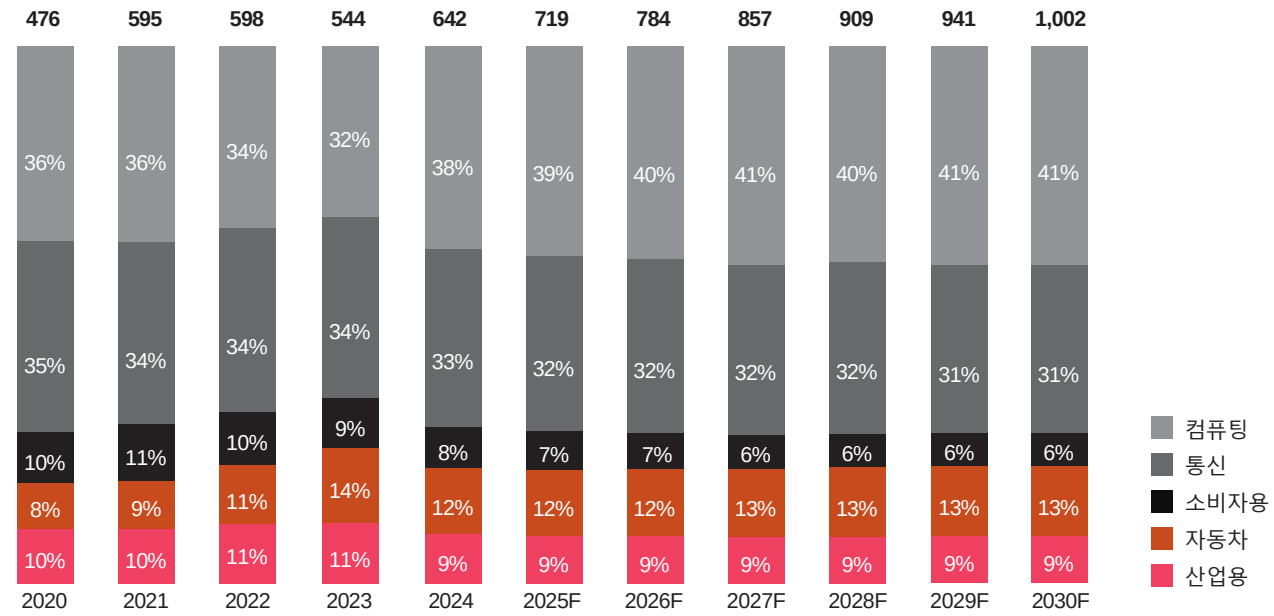
2023년에는 반도체 부족 현상을 해결하기 위해 생산량을 증가시켰으나, 수요 감소로 인해 재고가 증가하면서 반도체 산업은 침체기에 접어들었습니다. 하지만 현재 시장은 다시 안정화되고 있으며, 반도체 관련 매출 반등은 2022년의 최고치를 다시 한번 넘어설 것으로 예상됩니다.

AI, IoT, 자동차: 반도체 산업의 성장 동력

반도체의 7가지 제품 유형(메모리, 로직, 마이크로컴포넌트, 아날로그, 광전자, 센서, 개별소자) 중에서 가장 큰 비중을 차지하는 것은 메모리와 로직이며 이 제품들은 컴퓨팅 및 모바일 디바이스부터 산업용 및 차량용 제품에 이르기까지 전 산업에 필수 부품으로 자리매김하고 있습니다. 전 세계적으로 데이터를 더 빠르고 효율적으로 처리하기 위한 메모리 솔루션에 대한 수요가 증가하고 있습니다. AI, IoT, 클라우드 컴퓨팅 도입 증가로 이러한 수요는 더욱 증가할 것입니다. 또한 메모리, 로직 반도체는 전 세계 데이터 센터 확산에 따라 실시간으로 쌓이는 방대한 양의 데이터를 처리하는데 중요한 역할을 할 것으로 전망됩니다. (도표 2참조)

도표 2

사용처별 글로벌 반도체 시장 점유율, 2020–2030 (단위: 십억 달러)



Source: Omdia Q3 2024

반도체는 여러 분야에 사용되나, 그 중 컴퓨팅 시장은 2024년부터 통신분야를 넘어 가장 큰 비중을 차지할 것으로 예상되며, 2030년까지 연평균 성장률은 9%에 이를 것으로 전망됩니다. (도표 2참조). 이러한 높은 성장의 주요 요인 중 하나는 AI의 활용 확대입니다. 산업 전반에 걸쳐 AI 활용이 확대되어 복잡한 계산을 수행하기 위한 고성능 반도체 솔루션에 대한 요구가 증가하고 있습니다. 이와 더불어 머신 러닝, 뉴럴 네트워크, 데이터 분석 등의 기술은 반도체 산업의 지속적인 성장에 중요한 역할을 할 것입니다.

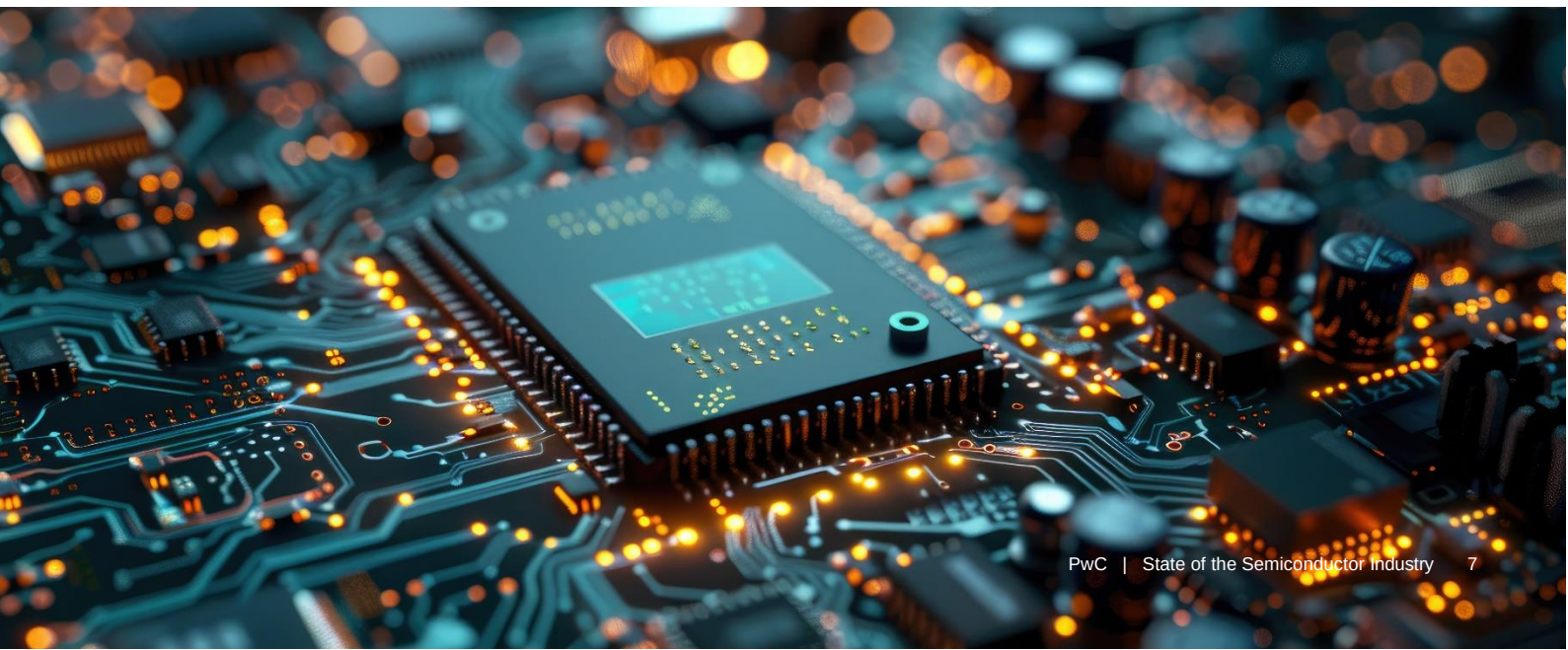
또한 범용 반도체를 사용하던 기업들이 각자의 목적과 사용 환경에 적합한 맞춤형 반도체로 전환하는 경우가 늘고 있습니다. 이는 산업용 데이터 센터부터 전자 제품에 이르기까지 다양한 분야에서 고성능, 에너지 효율성, 그리고 높은 보안을 필요로 하기 때문입니다. 일례로 애플은 일반 모바일 프로세서가 아닌 자사 제품에 특화된 사양이 반영된 자체 개발 반도체 칩인 M칩을 개발하여 더욱 뛰어난 성능을 확보하였습니다.

차량용 반도체 시장은 2024년부터 2030년까지 반도체 세분 시장 중 가장 빠른 연평균 약 10%씩 성장할 것으로 전망됩니다. 전동화 트렌드는 차량용 반도체 시장 성장의 주요 요인으로 PwC 내 자동차 산업 전문 연구기관인 PwC Autofacts에 따르면 2024년 2분기 승용차 부문 글로벌 베터리 전기차(BEV) 침투율은 13.3% 정도이며, 2030년에는 42.5%로 증가할 것으로 예상됩니다.[1] BEV를 구동하는 모터와 배터리 관리를 위해서는 다양한 전자장비와 전력관리 반도체(PMIC)가 필요합니다. 결과적으로 BEV에는 내연 기관 차량 대비 두배 이상의 반도체 부품이 탑재됩니다. 즉, 전기차 시장의 성장에 따라 차량용 반도체 시장은 자연스레 성장할 것입니다.[2]

SDV(Software-defined Vehicle)로의 전환 역시 차량용 반도체 시장 성장을 뒷받침하고 있습니다. SDV는 소프트웨어를 통해 차량을 지속적으로 업데이트할 수 있습니다. 이로 인해 소유자의 취향에 따라 차량 맞춤화가 가능하며 기능 변화 주기가 짧아집니다. 더불어 자율주행과 고급 편의 기능 탑재 또한 차량용 반도체 수요를 더욱 증가시키고 있습니다. 차량 1대당 탑재되는 반도체는 금액 기준 2019년 420달러에서 2023년 800달러로 약 2배 가량 증가했으며, 2030년에는 1,350달러에 이를 것으로 예상됩니다.[2]

$1,350

차량 1대당 탑재되는 반도체의 금액



Section 2

AI 필수재, 메모리 반도체의 제2도약

메모리 반도체는 데이터 저장과 처리를 위한 필수 부품으로 수많은 산업에 사용되고 있습니다. 더 큰 저장 용량, 더 빠른 속도 및 대역폭 확대에 대한 수요가 증가함에 따라 메모리 반도체는 앞으로도 지속적인 기술 발전을 이룰 것으로 기대됩니다. AI, IoT, 클라우드 컴퓨팅, 고급 데이터 분석과 같은 데이터 집약적인 수요처 확산에 따라 지난 20년 동안 메모리 반도체는 주요 반도체 중 가장 빠르게 성장해왔습니다. 연평균 성장률은 8.6%에 달하며, 전체 반도체 매출 중 메모리 반도체는 2008년 18%에서 2024년 약 25%를 차지할 것으로 예상됩니다.

“더욱 효율적인 데이터 저장과 처리에 대한 수요 증가는 미래 반도체 산업 성장에 있어 메모리가 핵심 역할을 수행할 것임을 시사합니다.”

범용균 부대표
PwC Global Semiconductor Sector Leader

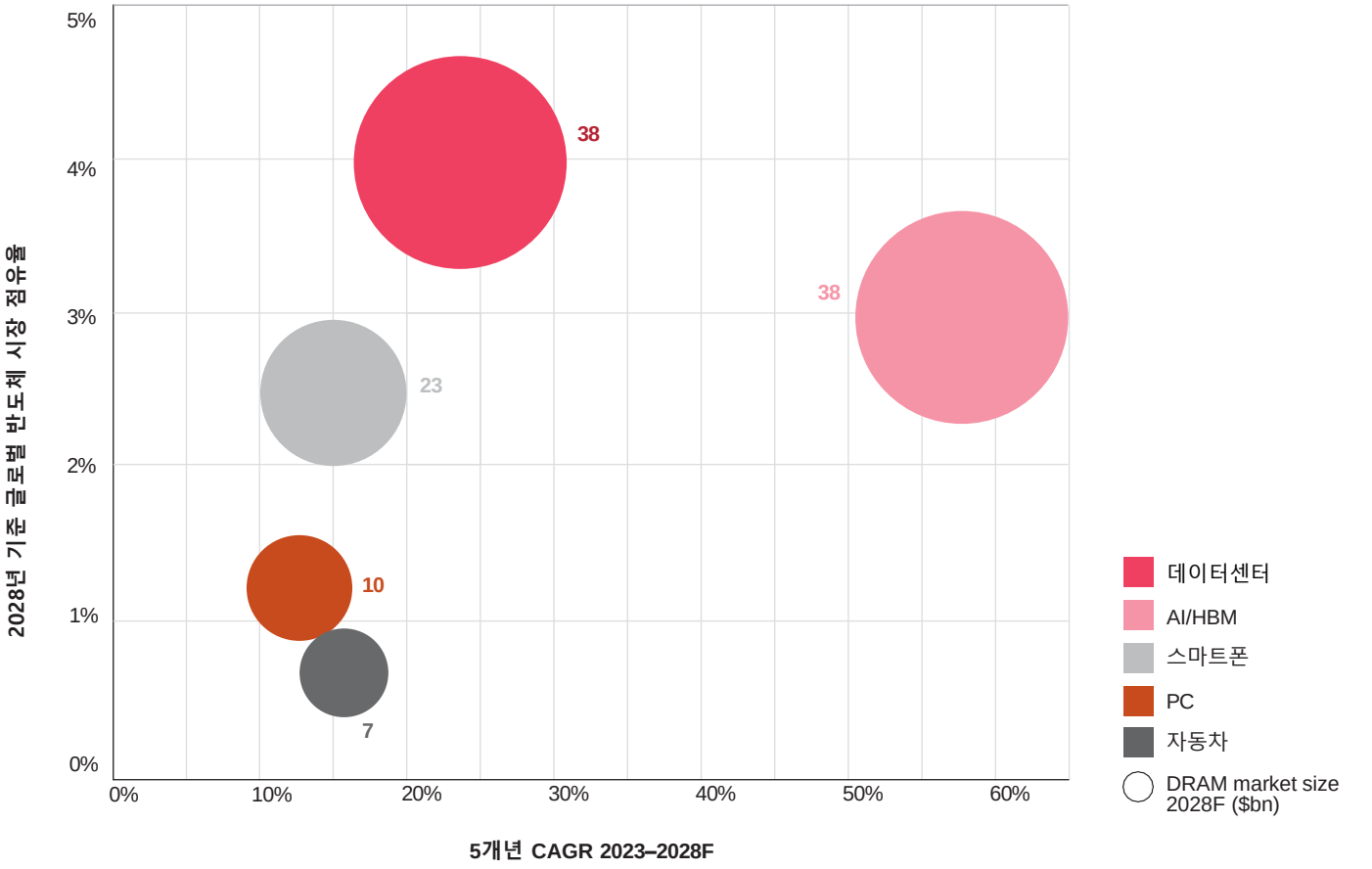
DRAM의 진화: AI와 데이터 중심 성장 촉진

DRAM은 2024년 전체 시장의 약 14%를 차지할 것으로 예상됩니다. 오랜 기간 개인용 컴퓨터와 데이터 센터에서 주로 사용되어온 DRAM은 앞으로 AI, 머신 러닝 및 클라우드 컴퓨팅과 같은 분야에 사용되며, 더 높은 대역폭, 더 빠른 속도 및 용량의 확대에 대한 요구가 증가할 것입니다.

HBM 시장은 2023년부터 2028년까지 연평균 57.5% 성장하여 2028년에 전세계 반도체 시장의 4.1%를 차지할 것으로 예상됩니다. 한편, 데이터센터용 DRAM은 전체 시장의 4.2%에 불과하지만, 연평균 성장률은 22.3%로 전망됩니다. 또한, 성숙 시장인 스마트폰 DRAM은 같은 기간 동안 2.6%의 시장 점유율을 유지하며, 연평균 성장률은 15.3%로 예상됩니다. (도표 3참조)

도표 3

글로벌 반도체 시장 내 DRAM 사용처별 시장 점유율 (%) 및 5개년 CAGR 2023–2028 (%)



Source: Omdia Q3 2023

HBM: 메모리 혁신의 시대를 앞당기다

HBM은 AI와 고성능 컴퓨팅의 성능 구현을 위해 개발된 제품입니다. HBM은 기존 DRAM 대비 매우 높은 데이터 전송 속도와 병렬 컴퓨팅 성능을 보유하고 있어 AI 트레이닝과 추론 작업에 필수적인 GPU에 대규모로 채택되고 있습니다.

또한, HBM의 신규 플랫폼 출시로 DRAM 탑재 용량이 연간 50~100% 가량 증가했습니다. 예를 들어, 2022년에 도입된 엔비디아의 H100은 80GB의 HBM3, 2년 후인 2024년에는 192GB의 HBM3E, 그리고 가장 최근에 발표한 HBM4 기반의 루빈(Rubin) 플랫폼은 764GB의 DRAM을 탑재하는 등 불과 5년 내 약 10배 수준으로 탑재 용량이 증가했습니다.[3]

HBM3E 이전까지는 메모리 전문 업체들이 베이스 다이(Base Die)를 포함한 모든 HBM 부품을 제조했습니다. HBM4 이후부터는 로직과 메모리칩의 일체화로 공정 난이도가 증가함에 따라 미세공정 기술과 대량 생산 인프라를 기 구축한 파운드리 업체가 HBM 제조 영역에 직접 뛰어들고 있습니다. 이에 향후 HBM 제조 영역에서의 파운드리와 메모리 업체 간 협력은 필수적으로 여겨집니다. 실제로 글로벌 선도사인 TSMC도 수년간 AI 전용 HBM 관련 다수의 메모리 업체와 협업 생태계를 구축해오고 있습니다.[4]

DRAM은 표준화 인터페이스를 가진 제품 특성상 비용 절감과 대량 생산이 시장에 중요한 영향을 미칩니다. 반면, HBM은 빠르게 발전하는 기술과 성능, 그리고 제품과 고객별로 상이한 요구에 대응하는 것이 핵심인 관계로 오히려 특정 소수 업체들이 주도하는 폐쇄적인 생태계에서 운영됩니다.

HBM 시장은 2028년까지 비트 그로스(Bit Growth) 기준 연평균 성장률 64%, 매출 기준 연평균 성장률은 58%로 약 380억 달러 규모의 시장으로 성장할 전망입니다. 이는 서버 DRAM 시장의 약 50%에 해당하는 규모이며, 전체 1,360억 달러 규모의 DRAM 시장의 27.6%에 해당합니다.

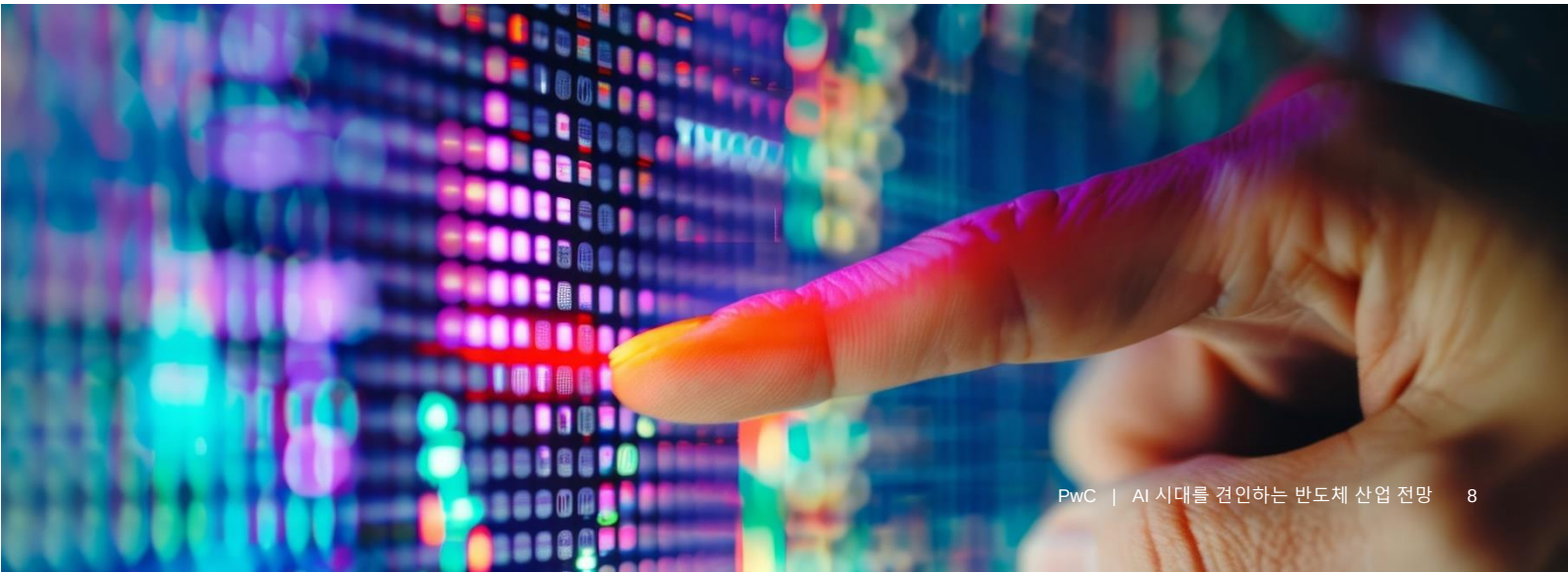
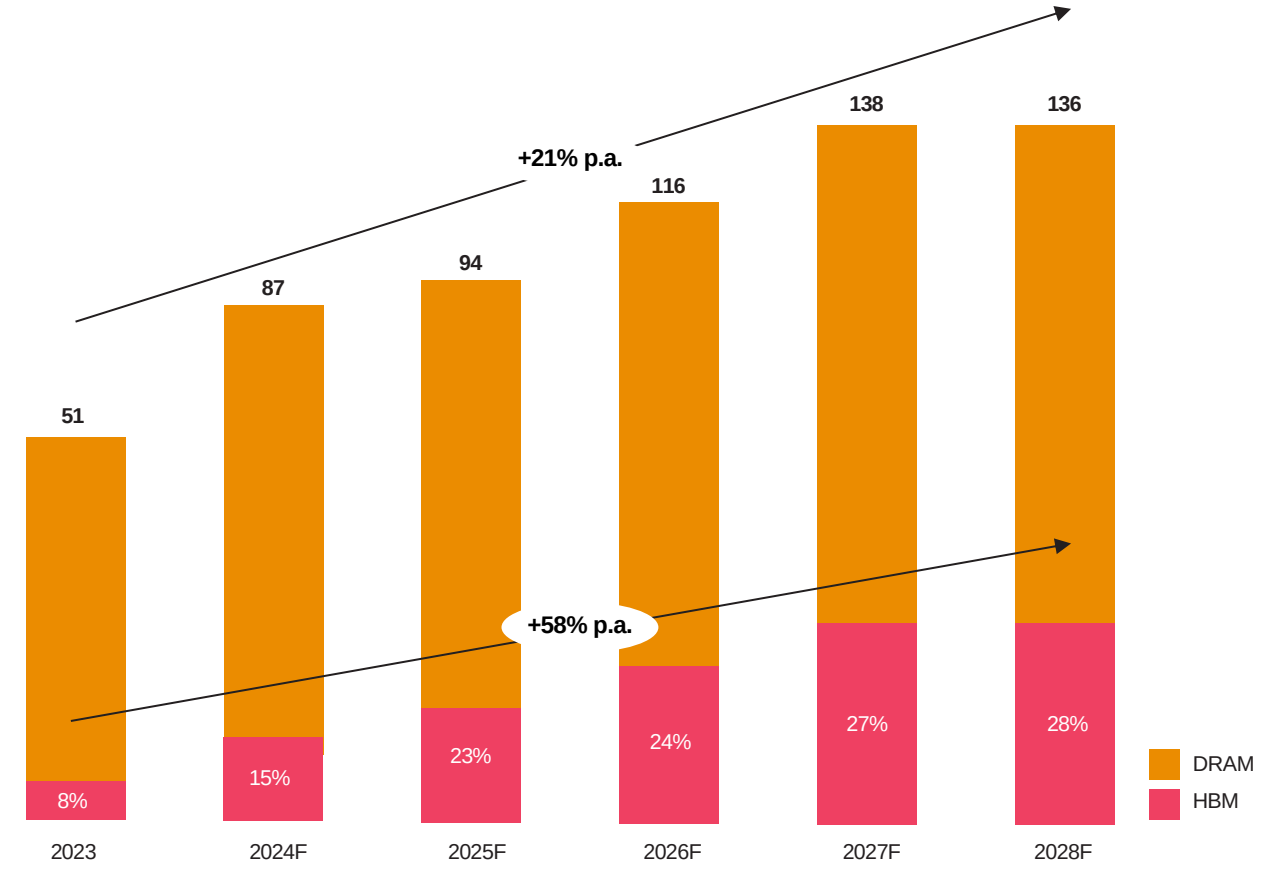


도표 4

글로벌 DRAM 및 HBM 시장 규모, 2023–2028 (단위: 십억 달러)



Source: Omdia Q3 2024

DRAM 수익성 개선에 따른 투자 확대

시장 내 DRAM 수요가 급증하여 DRAM의 단가와 이익이 모두 상승함에 따라 메모리 업체들은 자본적 지출(CAPEX)을 확대하고 있습니다. DRAM 산업의 영업이익률은 2024년 초에 20%를 돌파한 후, 연말까지 30~40%로 증가할 것으로 예상됩니다. 2025년에는 이러한 수익성 증가로 CAPEX가 2022년의 330억 달러를 크게 초과하며 사상 최고치를 기록할 가능성이 있습니다. 다만, 대부분의 투자가 HBM 후공정에 집중될 예정으로, 2025년 DRAM 생산 용량은 2022년과 비슷한 수준을 유지할 것으로 보입니다.

2022년부터 미국, 일본, 한국에서 도입한 보조금과 세제 혜택의 효과는 2025년부터 산업 전반적으로 영향을 미치기 시작할 것입니다. 삼성의 평택 P4 공장은 2025~2026년 중 대량 생산을 시작할 예정이며, SK하이닉스의 용인 공장은 2027년, 마이크론의 보이시, 히로시마, 뉴욕에 위치한 신규 공장은 2026~2029년 중 가동을 시작할 예정입니다.

3D DRAM으로의 전환: 2D 기술의 한계를 넘다

선단 노드 진입에 따른 기술적 난이도 증가로 DRAM 제조비용의 효율화 수준은 크게 둔화했으며, 추가 비용 절감은 더욱 어려워질 것으로 전망됩니다. 실제로 10nm급 노드 이전에는 DRAM 제조비용 효율화가 연간 20~30%에 달했으나, 2017년 10nm급 노드 이후에는 연간 6%로 감소했습니다.

3D DRAM은 2D DRAM의 뒤를 이어 업계의 장기적인 비용 절감 목표를 지속할 것으로 예상됩니다. 10nm급 이하 기준 3D DRAM은 2세대 제품부터 비용 효율성을 크게 제고할 것으로 기대됩니다. 현재 8단과 16단 DRAM 제품은 개발 중에 있으며, 고적층 3D DRAM의 대량생산 시기는 2030년 경으로 예상됩니다.[6]

NAND Flash의 회복: AI 주도의 수퍼사이클 진입

NAND는 높은 저장 밀도, 확장성, 비트당 저단가 경쟁력을 보유한 제품으로 소비자 가전과 데이터 센터 등 다양한 분야에서 활용되고 있습니다. 특히, NAND 수요를 견인하는 스마트폰, PC, 서버 분야의 성장으로 비트 그로스는 2013년 371억개(1Gb 환산 기준)에서 2023년 7,449억개로 증가하였습니다. 지난 10년간 약 20배 이상의 성장률을 보였습니다. 또한 AI 트레이닝 및 추론과 같은 대규모 작업을 처리하는 데 필수적인 대용량 SSD에 대한 수요 가속화로 2028년 기준 NAND 시장 규모는 1,150억 달러에 이를 것으로 예상됩니다.[2]

> **“산업 내 AI 채택이 확대됨에 따라 고용량 HBM에 대한 수요가 가속화되고 있으며, 특히 대용량 작업을 요하는 학습 및 추론 영역에서 이 같은 특징이 돋보입니다.”**
>
> **장유신 파트너**
> Partner, Strategy& Korea

NAND 시장은 과잉공급으로 인한 가격하락으로 2023년 2분기에 당해 역대 최저 실적을 기록한 후 주요 업체들의 공급 조절 전략으로 3분기 이후 회복세를 보였습니다. 다만, 2024년 업체들의 CAPEX 규모가 전년비 13% 감소함에 따라 NAND의 확대 생산에는 일부 제약이 있을 것으로 전망됩니다. AI 기반 수요 확대로 2025년 반도체 시황이 긍정적으로 전망되는 바 주요 업체들의 공장 가동률은 상승 추세에 있으며, NAND 시장 또한 중장기적으로 유의미한 슈퍼 사이클에 진입할 가능성이 있습니다.

스마트폰의 AI 기능 도입 확대는 NAND 수요 확대로 이어질 것입니다. 2028년 기준 스마트폰 메모리가 글로벌 반도체 시장의 5.4%를 차지하며, 연평균 성장률은 30%에 달할 전망입니다(다음 페이지 도표 5참조). 데이터 센터 부문은 대규모 AI 프로젝트와 저장 인프라 확대로 2028년 글로벌 반도체 시장의 3.5%를 차지하고, 연평균 성장률 33.4%로 스마트폰 분야의 성장률을 상회할 것입니다. PC 분야 역시 AI 수요 기반 전체 시장의 2.1%를 차지할 것으로 보입니다. 차량 부문은 현재 NAND 시장에서 소규모 시장에 불과하지만 고성능 메모리 솔루션 수요 증대로 2028년까지 연평균 성장률 23%로 빠르게 성장할 것으로 예상됩니다. (다음 페이지 도표 5참조)

NAND 시장의 CAGR 은

23%로

강한 성장세가 전망됨

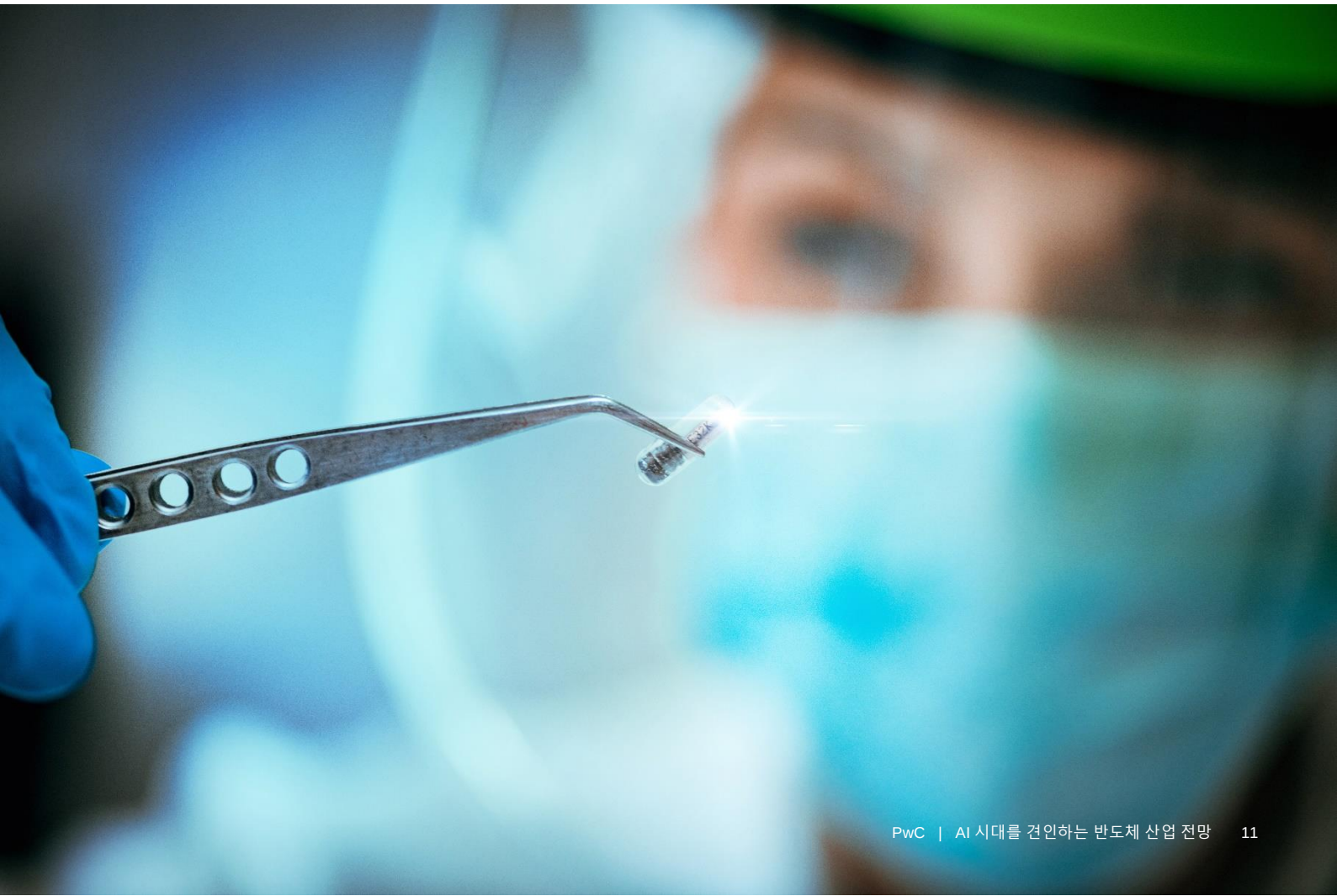
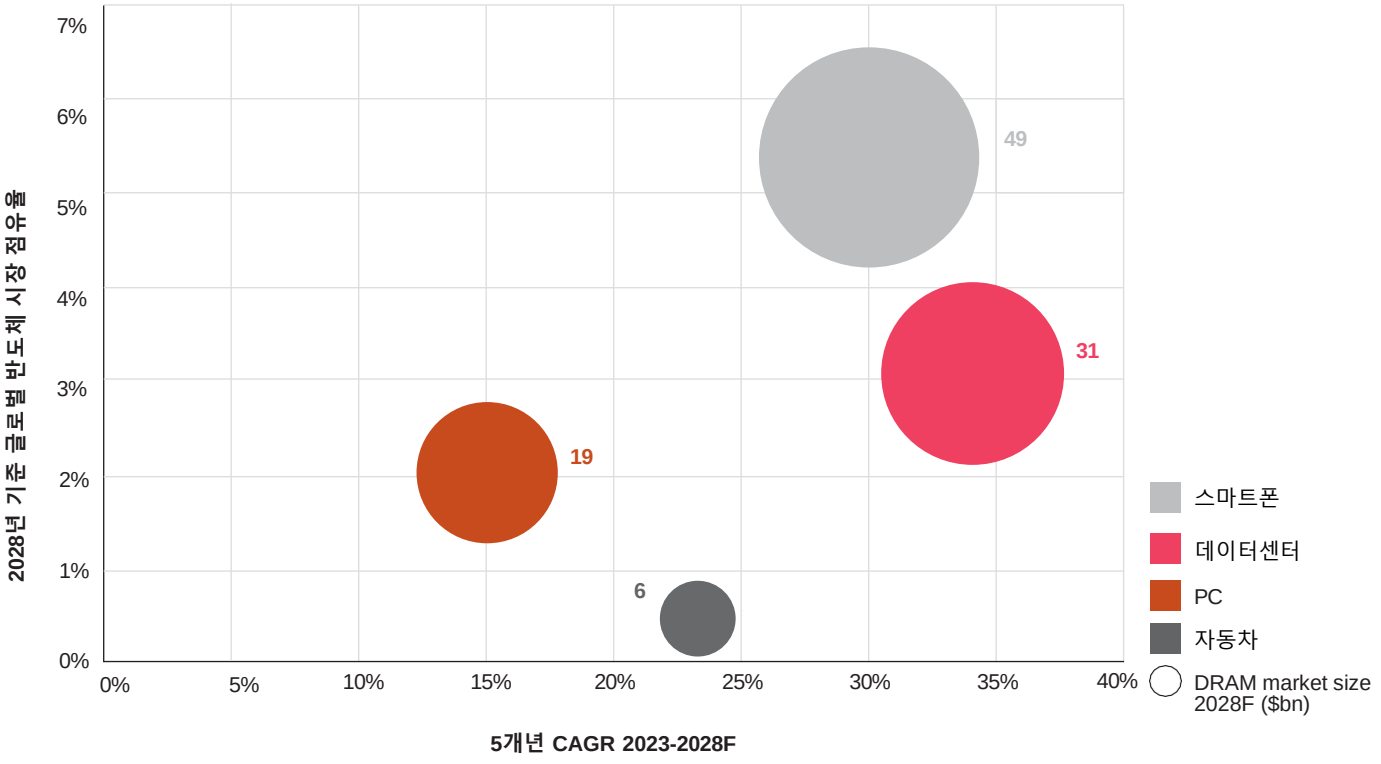


도표 5

글로벌 반도체 시장 내 NAND 사용처별 시장 점유율 (%) 및 5개년 CAGR 2023–2028 (%)



Source: Omdia Q3 2024

NAND의 확장: QLC와 1,000단 기술로의 변화

2024년 대부분의 메모리 업체들은 NAND 200단 이후 추가 적층 관련 기술적 한계에 봉착했습니다. 이러한 한계를 1,000단 NAND를 통해 극복하고자 활발한 연구개발을 진행 중에 있습니다.

AI 수요에 대응하기 위해서는 NAND 업체들은 TLC 기술에서 가격 경쟁력과 높은 저장 밀도의 이점을 지닌 QLC 기술로의 전환 중에 있습니다. 다만, QLC에는 성능과 내구성 간 일종의 트레이드 오프(Trade-off) 이슈가 발생할 수 있으나 기술적으로 해결 가능합니다. 2029년에 QLC는 NAND 시장의 절반 이상을 차지할 것으로 예상되며[2], 대표 업체로는 SK하이닉스, 솔리다임(Solidigm), 마이크론 등이 있습니다.[7]

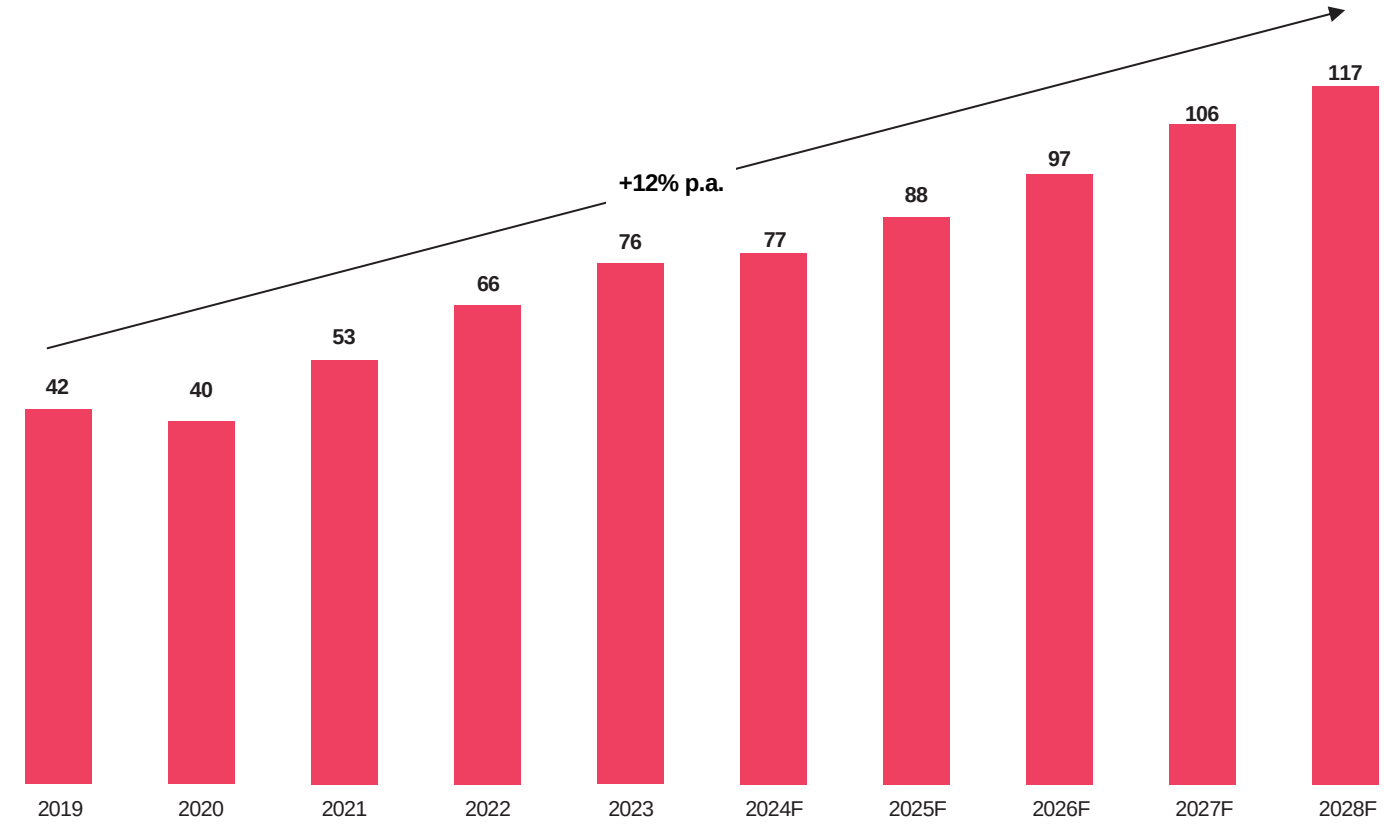
Section 3

자동차의 핵심, 엔진? 반도체!

전기차 도입 확대와 SDV로의 전환 추세는 차량용 반도체 시장에 중대한 변화를 일으키고 있습니다. 2023년에 글로벌 자동차 생산량이 팬데믹 이전 수준을 기록하며 향후 2028년까지 연평균 성장률 8.9%로 지속 성장하여 1,170억 달러의 시장 규모를 형성할 것으로 예상됩니다. (도표 6참조)

도표 6

차량용 반도체 시장 규모, 2019–2028 (단위: 십억 달러)



Source: Omdia Q3 2024

전기차 시대를 주도하는 전력 반도체

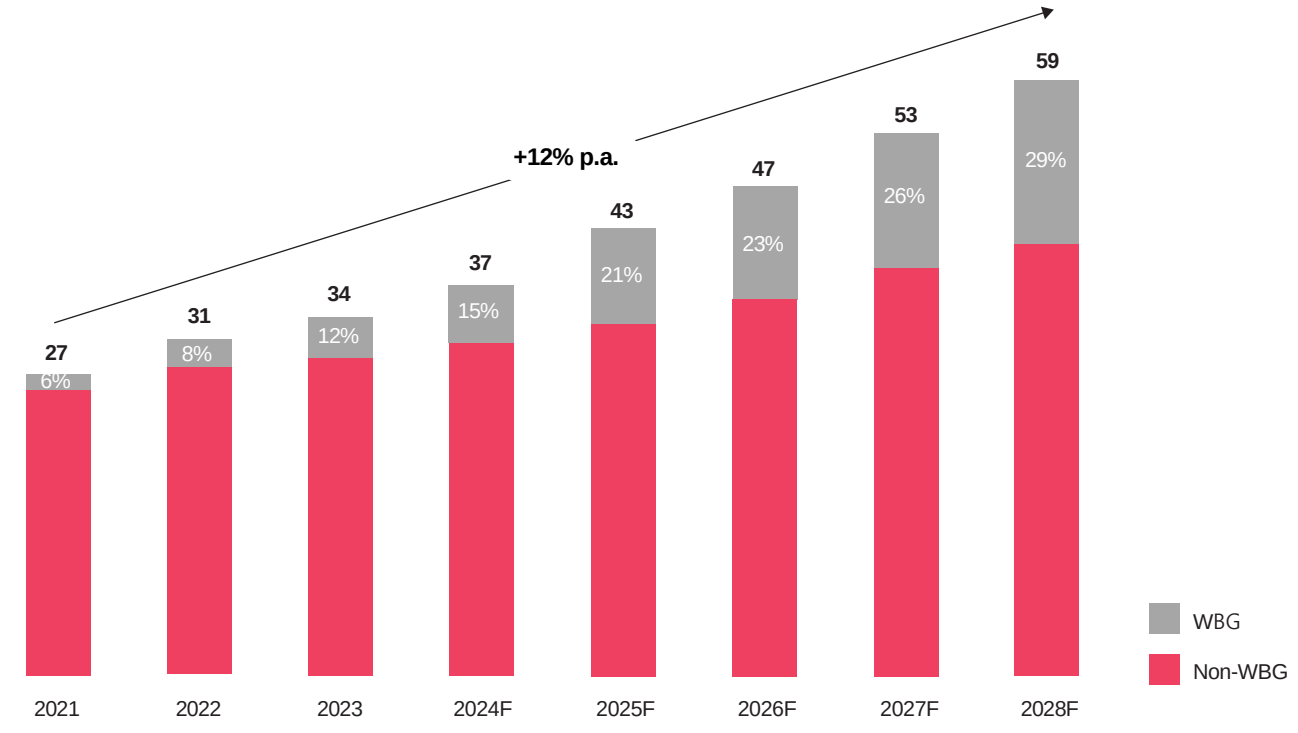
전력 반도체는 자동차 전동화에 따라 메인 인버터, DC-DC 컨버터, 온보드 충전기(OBC), 배터리 관리 시스템(BMS) 등에서 핵심 역할을 담당하는 동시에 매우 높은 시장 기회를 창출하고 있습니다. 테크 전문 연구 및 자문 기관인 옴디아에 따르면 전기차 한 대 당 탑재되는 전력 반도체는 금액 기준 내연 기관 대비 6배가 더 높습니다. 2023년 기준 자동차 시장 내 전력 반도체 매출은 210억 달러로 전년 대비 30.2% 성장했으며, 향후에도 강한 성장세가 지속될 전망입니다.

자동차의 전동화로 인해 실리콘 카바이드(SiC), 갈륨 나이트라이드(GaN), 갈륨 옥사이드(Ga2O3) 등 신소재를 활용한 차량용 와이드 밴드갭(WBG) 반도체 개발이 활발하게 이뤄지고 있습니다. 기존 실리콘(Si) 대비 신소재 활용 반도체는 고효율 에너지, 고밀도 전력, 고내열, 소형화, 경량화 등의 특성으로 차량의 주행거리와 시스템 성능을 향상시키는 데 최적화되어 있습니다.

인피니온, ST마이크로일렉트로닉스, 온세미, 로옴(Rohm), 넥스페리아(Nexperia) 등 주요 전력 반도체 업체들은 WBG 기술에 더욱 집중하는 한편 관련 시장 점유율을 크게 확대하고 있습니다. 옴디아는 2023년 글로벌 자동차용 전력 반도체 매출에서 WBG 비중은 12% 수준이었으나, 2028년까지 해당 비중은 29%까지 증가하여 약 170억 달러 수준의 시장을 형성할 것으로 전망하고 있습니다. (도표 7참조)

도표 7

전력 반도체 및 모듈 시장 규모, 2021–2028 (단위: 십억 달러)



Source: Omdia Q3 2024

실리콘과 WBG: 경쟁인가, 공생인가

실리콘 반도체는 여전히 소형 전기차와 하이브리드 차량 등 저전력 및 가격민감도가 높은 분야에서 수요가 지속되고 있습니다. 반면 고성능 전기차 분야에서는 SiC와 같은 신소재의 사용이 필수적으로 여겨집니다. 특히, 고효율 에너지 관리와 고성능 모델의 인버터에 적합합니다. 그러나 SiC 제조 공정에서 높은 에너지 소모와 취성(외부 충격이나 변형에 대해 쉽게 깨지거나 부서지는 성질)으로 인한 높은 가공 난이도로 수율이 이상적인 수준에 도달하지 못하고 있습니다.

주요 업체들의 SiC 제조 시설에 대한 대규모 CAPEX 투자로 SiC 가격은 점차 하락할 것으로 예상되며, 50개가 넘는 중국 내 SiC 신생 업체들의 시장 진출로 인해 하락 추세는 더욱 가속화될 전망입니다. 다만, 다수 업체의 SiC 시장 진출은 잠재적 공급 과잉으로 인한 가격 하락 리스크를 내재하고 있습니다. 현재 SiC 제품은 200mm 웨이퍼로의 전환하는 과도기에 있으나 200mm 웨이퍼 제조 공정의 복잡성과 장비 스펙트럼 한계 등으로 생산에 차질을 빚고 있어 단기적으로 공급 과잉으로 인한 가격 하락 리스크는 해소될 전망입니다. 향후 SiC 시장의 수요와 공급은 미국과 유럽의 전기차 수요 성장세에 크게 영향을 받을 것으로 전망됩니다.

반면, 업계에서 널리 사용되는 대구경 웨이퍼와 호환 가능한 GaN은 기존 제조 인프라 기반 비용경쟁력, 소형화, 경량화 등의 이점을 보유하고 있습니다. 이에 GaN은 소형 고속 충전기, 전력 컨버터, 그리고 온보드 충전 시스템 등에 활용도가 높습니다. 현재 GaN의 전압 커버리지 확장과 신뢰성 연구가 진행 중이며, 고전압 응용 분야에서 SiC 를 대체할 가능성도 존재합니다. 인피니온은 세계 최초로 300mm GaN 웨이퍼 기술 개발로 생산비용의 획기적 감소의 가능성을 보였습니다. 향후 GaN 시장에서는 인수합병이 활발히 진행될 것으로 예상됩니다. 실제로 2023년 인피니온은 GaN 시스템즈를 인수했으며 2024년 하반기에는 르네사스(Renesas)가 트랜스폼(Transphorm)를 인수를 완료할 계획입니다.

자동차 산업이 계속 발전함에 따라 SiC와 GaN 등 전력 반도체 기술은 다양한 차량의 성능, 전력, 비용 등 특정 니즈에 따라 공존할 것으로 보이며, 각 기술의 단점을 상호보완하여 향후 전기차 설계와 효율성 측면에서 시장의 기회를 확대할 것입니다.

전력 반도체 시장의 지각변동

전력 반도체 제조 프로세스는 Si 또는 SiC와 같은 원료를 사용한 고품질 웨이퍼를 생성한 후 웨이퍼 프로세스를 통해 반도체적 특성을 띠게 처리하고, 개별 다이로 절단하여 개별 소자(Discrete)로 사용하거나 전력 모듈로 조립하는 것입니다. 최종 단계인 패키징을 통해 열관리와 성능을 최적으로 관리하며, 다수의 다이를 결합해 전력 모듈을 제조합니다.

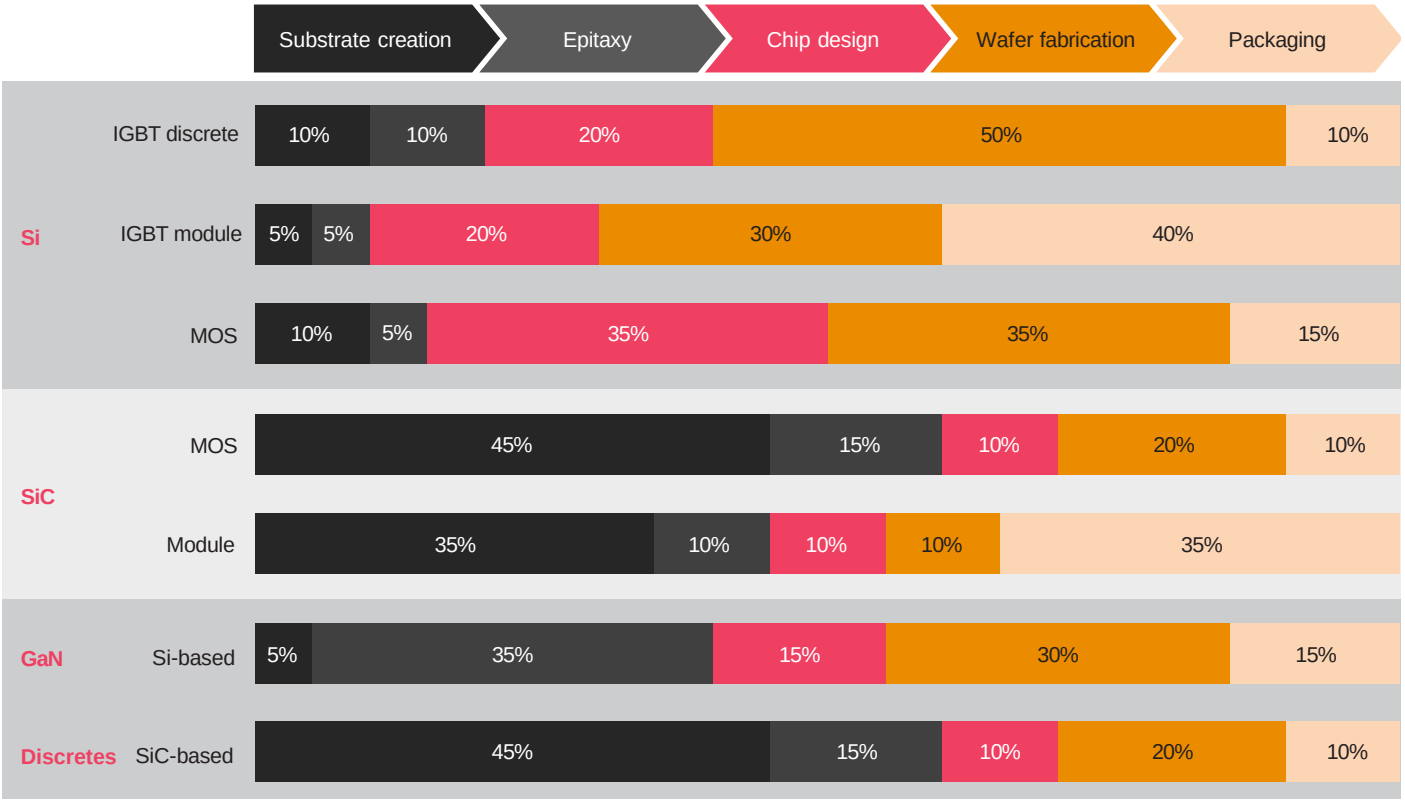
SiC 기반 반도체 제조 프로세스 중 고품질 웨이퍼 생성은 총 부가가치 중 35~45%를 차지하는 만큼 어려운 분야임을 보여줍니다. 다만 SiC의 시장 규모 확대와 기술 발전에 따라 제조 단계 등에서의 비용효율화 실현으로 SiC 기반 디바이스 접근성은 점차적으로 높아질 것으로 전망됩니다.

패키징 단계 또한 총 부가가치 중 35~40%를 차지하며 공정에 대한 이해 및 기술에 대한 요구가 높은 부분입니다. (다음 페이지의 도표 8 참조). 제품의 최종 신뢰도와 성능 향상을 위한 핵심 분야인만큼 반도체 업체들은 수직계열화 전략에 기반하여 관련 역량의 내재화에 집중하고 있습니다.

> **“SDV 트렌드는 소프트웨어를 통해 차량 기능을 지속적으로 업데이트하고 개선할 수 있게 되면서 더욱 확산되고 있으며, 하드웨어와 소프트웨어의 디커플링 트렌드는 향후 소비자 대상으로 더욱 많은 맞춤형, 서비스 유연성, 혁신의 기회를 제공할 것입니다.”**
>
> **Tanjeff Schadt**
> Partner, Strategy& Germany

도표 8

전력 반도체 유형별 반도체 제조 단계 내 부가가치 비중 (%)



Source: PwC analysis based on research from the beginning of 2023

소프트웨어 정의 차량 (SDV)과 반도체 수요

SDV로의 전환은 소프트웨어 기반 자동차 기능을 지속 업데이트하고 성능을 개선하려는 산업 트렌드를 반영합니다. 하드웨어 변경 없이 소프트웨어 업데이트 및 변경이 가능한 것은 소비자의 선택권을 확대하는 동시에 제조사의 혁신 속도를 제고할 수 있습니다. SDV는 현재 전기전자(E/E) 아키텍처에서 향후 영역(Zone) 및 중앙집중형(Central) 아키텍처로 전환될 것이며 이에 따른 고성능 프로세서 탑재 수요가 더욱 증가할 것입니다. 또한 이는 소수의 ECU(Electronic Control Unit, 전자제어장치)로 다수의 전자 장치들을 관리하는 등 간소화의 가치를 제공합니다.

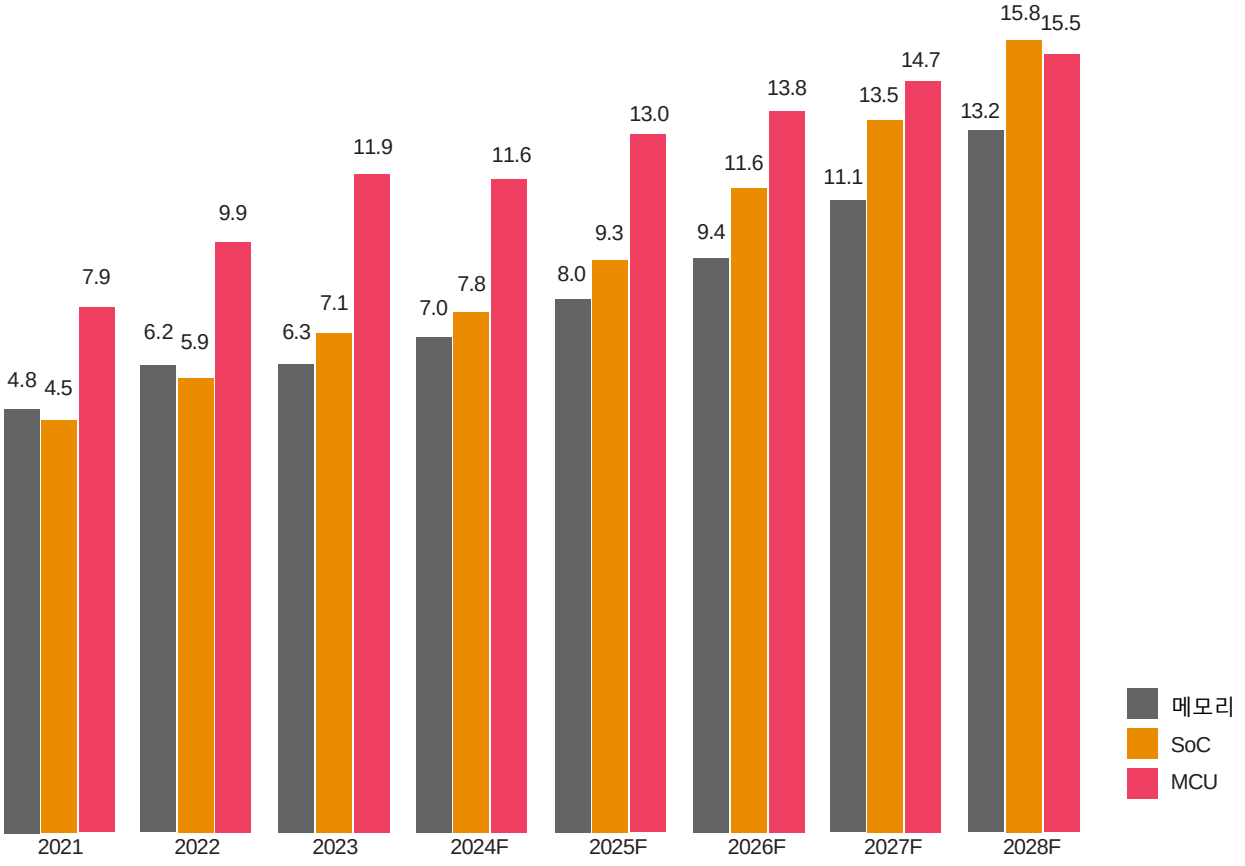
SDV 반도체에 있어 핵심은 실시간 데이터 처리, ADAS 제어, 보안 모듈, 인포테인먼트 시스템 등 기능을 하나의 칩에 내장 가능한 시스템 온 칩(System on Chip, SoC) 프로세서입니다. 자율 주행 등 머신러닝 알고리즘을 구동하는 GPU 및 NPU의 수요 증가 모두 SoC 시장 성장을 견인하고 있습니다. 이에 따라 2028년 기준 자동차 SoC 시장은 160억 달에 달할 것이며, 연평균 성장률은 17%로 전망됩니다. (다음 페이지의 도표 9참조)

SoC와 더불어 자동차의 실시간 제어 작업과 주변 장치 제어에 특화된 마이크로컨트롤러(MCU) 역시 중요한 요소입니다. MCU는 제어 타이밍, 전력 효율성, 신뢰성이 필수적으로 담보되어야 하는 엔진 제어, 임베디드 센서, BMS와 같은 시스템에 주로 사용됩니다. 또한, 차량과 외부 네트워크 간 실시간 데이터 교환을 요하는 이더넷, Wi-Fi, 블루투스, V2X 통신 등의 무선 연결 역시 지원합니다.

SDV의 특성상 방대한 양의 데이터를 저장하고 접근해야 하므로 고용량, 고속 메모리 솔루션 역량이 필수적으로 요구되고 있으며, 향후 소프트웨어가 고도화될수록 고사양 메모리 솔루션에 대한 수요가 증가할 것입니다. 2023년 기준 차량용 반도체 시장 내 메모리 칩은 8%를 차지했으나 2028년에는 해당 비중이 11%까지 증가하여 130억 달러 규모의 시장으로 성장할 것으로 전망됩니다 (도표 9참조).

도표 9

차량용 반도체 종류별 시장 규모 (단위: 십억 달러)



Source: PwC analysis based on research from the beginning of 2023

Section 4

고래싸움에서 살아남는 법: 지정학 리스크 대응방안

반도체 산업의 글로벌 공급망은 지정학적 리스크로 인한 각종 수출 규제 및 국내 수급 강제 등의 법적 및 정치적 제약에 노출되고 있습니다. 이에 안정적인 공급망 확보를 통한 리스크 헤징이 각 업체들의 반도체 산업 내 성공 방정식으로 자리매김하고 있습니다.

미-중 간의 기술적, 경제적 디커플링

최근 미·중 간의 기술적, 경제적 상호 독립 기조로 반도체 공급망 또한 변화의 압력을 받았습니다. 이에 따라 미국과 중국 양국을 중심으로 한 디지털, 커넥티비티 기술 권역(Technosphere)이 형성되고 있으며, 남반구에서는 중국 테크기업들이 대규모 디지털 인프라 프로젝트를 추진하고 있습니다.

중국 정부는 중국 자동차 제조사 대상으로 내년까지 중국산 반도체 사용 비중을 25%까지 확대하도록 권고

미국 정부는 수출통제개혁법(Export Control Reform Act, 2018), 외국기업책임법(Holding Foreign Companies Accountable Act, 2021), 안전장비법(Secure Equipment Act, 2021), 미국 반도체법(US CHIPS Act, 2022)와 같은 입법 조치 및 국방수권법(National NDAA) 제5949조 규칙 제정 논의 등으로 중국 반도체에 제재를 가하고 있습니다. 해당 법안 통과시 2027년 12월 23일 이후로 미국 정부기관은 특정 중국 반도체를 포함한 전자기기를 구매할 수 없게 됩니다. 현재까지는 정부기관만을 대상으로 한 법안이나, 과거 유사 사례를 고려할 때 민간 시장까지 규제가 확대될 가능성이 높습니다.

중국 또한 중국제조 2025(2015), IT 응용혁신 프로그램(ITAI, 2016), 수출통제법(2020), 데이터보안법(2021), 반외국제재법(2021), 외국국가면제법(2023) 등 유사한 규제를 수립하였으며, 자국 자동차 제조사에 내년까지 중국산 칩 사용 비중을 25%까지 높이도록 권고하였습니다.

특정 지역 및 국가가 글로벌 반도체 생산의 핵심적 역할을 담당하는 만큼 지정학적 갈등 고조 및 무역 압박은 반도체 공급망의 안정성을 위협하는 리스크로 작용할 것입니다.

지정학적 리스크 대응 전략

지정학적 리스크에 대한 대응 전략은 글로벌 기업들의 중요한 성공 요소가 되었습니다. 변화하는 경제, 사회 환경과 지정학적 리스크를 정확히 인지하고, 그에 따른 효과적 전략을 채택하는 것이 기업의 장기적 성공의 핵심이 될 것입니다.

멀티 소싱

지정학적 리스크에 대한 대응 전략은 글로벌 기업들의 중요한 성공 요소가 되었습니다. 변화하는 경제, 사회 환경과 지정학적 리스크를 정확히 인지하고, 그에 따른 효과적 전략을 채택하는 것이 기업의 장기적 성공의 핵심이 될 것입니다.

리스크 최소화

선도 기업들은 공급망 차질을 최소화하기 위해 전략적으로 다수의 공장과 공급처를 활용합니다. 생산 거점과 공급처를 다변화함으로써 기업은 지정학적 리스크가 운영에 미치는 영향을 최소화할 수 있습니다.

제품 현지화

정부 규제가 진화하면서 특정 지역, 시장에 대한 솔루션 사용이 의무화되기도 합니다. 테크 기업들은 기술 스택의 회복력을 높이고 규제에 빠르게 대응해야 합니다. 제품과 부품을 현지 요구사항에 따라 차별화해 제공하는 선제적 전략, 혹은 글로벌 제품을 유지한 채 필요 시에만 조정하는 수동적 전략을 선택할 수도 있습니다.

인재 확보

성공적인 현지화를 위해 우수한 인재 확보는 필수적입니다. 업체의 현지화 수요가 증가하면서, 인재 파이프라인 구축의 중요성 또한 증가하고 있습니다. 2023년 Strategy&의 연구 'Bridging the Talent Gap'[9]에 따르면 유럽에서도 EU가 설정한 2030년 글로벌 시장 점유율 20% 목표 달성을 위해서는 35만명의 전문가가 추가로 필요합니다.

Section 5

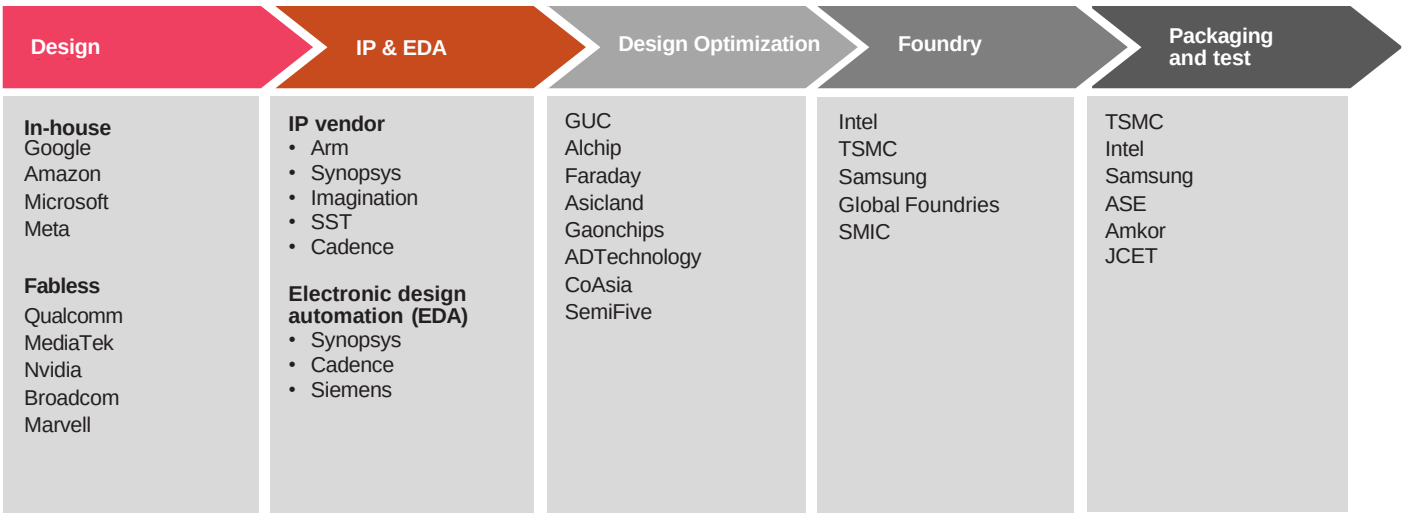
직접 만드는 반도체 DIY 시대

맞춤형 반도체 시장은 고성능, 고효율, 보안 수요의 증가에 따라 영광의 시기를 맞이할 것입니다. 맞춤형 반도체 시장의 확대로 디자인하우스, 파운드리, EDA 툴과 같은 밸류체인 또한 확장되었습니다. GUC, AIchip, ADTechnology 등은 전문 설계 서비스를 제공하며, 하드웨어 설계를 오픈소스로 제공하는 Open Compute Project는 Chiplet Marketplace 이니셔티브로 사전 설계된 프로세서 부품을 쉽게 활용할 수 있도록 했습니다. 프라운호퍼IC연구소(IIS)와 같은 소규모 고객 대상 파운드리는 수십만 개가 아닌 수만 개 규모로도 맞춤형 반도체를 생산할 수 있습니다. 향후에도 맞춤형 반도체 시장은 사용 가능한 IP의 증가, 소프트웨어 툴의 발달, 설계 비용의 감소 등으로 다양한 산업에 걸쳐 성장할 것입니다. (도표 10참조)

다만, 설계 비용이 높은 첨단 반도체의 확장성은 과제로 남아있습니다. 전자 산업 전문 컨설팅 업체 IBS는 10nm 칩 설계 시 1억 7천만 달러, 5nm 칩 설계 시 약 5억 달러 이상의 비용이 발생한다고 밝혔습니다. 한편 데이터 센터 부문에서는 아마존, 메타, 마이크로소프트, 구글 4사의 주도 (글로벌 데이터센터 관련 CAPEX의 약 30~40% 수준)로 맞춤형 반도체 개발이 활발히 이루어지고 있습니다.[2]

도표 10

맞춤형 반도체 개발 프로세스 (선별된 업체 기준)



Source: Public available information

영상 처리용 맞춤형 반도체

맞춤형 반도체의 첫 번째 성장 배경은 비디오 스트리밍 수요 폭증에 따른 영상 처리 수요 확대입니다. 구글에 따르면 비디오 스트리밍은 전세계 인터넷 트래픽의 약 60%를 차지합니다.[8] 이에 구글은 지난 5년간 비디오 코딩 유닛(VCU)으로 알려진 맞춤형 비디오 인코딩 IC를 개발하여 유튜브의 필요 서버 수를 줄였습니다. VP9 비디오 코딩 형식을 이용해 20개의 VCU를 장착한 서버 한 대로 기존 서버 여러 대를 대체한 것입니다. 이로 인한 유튜브의 3년간의 트래픽 관련 비용 절감액은 과거 대비 20~33배 수준에 달할 것으로 추정됩니다.[8]

메타와 텐센트도 자체 비디오 프로세서를 개발해 성능을 개선했으며, 네틴트(Netint)와 같은 비디오 프로세싱 반도체 스타트업 역시 서버, 전력, 비용대비 성능을 극대화하고자 ASIC(Application-Specific Integrated Circuit) 반도체를 활용하고 있습니다.(다음 페이지의 도표 11참조)

네트워크 및 보안 관련 맞춤형 반도체

다량의 데이터 처리를 동반하는 네트워크 및 보안 부문 역시 맞춤형 반도체의 주 적용 분야입니다. 옴디아에 따르면, 아마존 인프라의 약 20%가 네트워크 및 보안 처리에 할당됩니다. 이러한 중요성 아래 아마존은 DPU와 이더넷 컨트롤러를 통합된 맞춤형 반도체, 니트로 카드(Nitro Card)를 자체 개발했습니다. 이 DPU는 가상 PC 데이터의 평면 처리(캡슐화, 라우팅 등), 암호화 및 기타 네트워크 관련 기능을 수행하며, 아마존은 각 버전마다 스토리지 제어, 보안 모니터링, 시스템 제어 및 분석 기능을 추가했습니다. 그 결과 확보된 여유 CPU 코어는 IaaS(Infrastructure-as-a-Service) 형태로 기업에 판매되었고 해당 성공사례는 맞춤형 보안 칩 개발의 배경이 되었습니다.

AI용 맞춤형 반도체

AI는 가장 성능 집약적이자 상업적으로 중요한 부문 중 하나로, 모든 주요 클라우드 서비스 제공사(Cloud Service Providers, CSP)들은 더 높은 효율, 더 빠른 속도를 구현하는 AI용 맞춤형 반도체를 개발해 경쟁우위를 확보하려 하고 있습니다. 구글은 AI 추론에 특화된 맞춤형 반도체, 텐서 처리 유닛(Tensor Processing Unit, TPU)을 CSP 중 최초로 개발했으며, 2024년까지 100만 개 이상의 TPU를 적용할 예정입니다.[2]

다가오는 미래: 중국과의 맞춤형 AI 반도체 경쟁

중국은 미국 주도의 대중국 GPU 수출 제재로 AI용 맞춤형 반도체를 자체 개발해야만 합니다. 중국의 CSP는 AI 추론과 학습에 최적화된 맞춤형 반도체를 구축하고 있습니다.

2023년 텐센트는 엔비디아의 추론용 A10 GPU를 대체하기 위해 ASIC인 Zixiao v1을 확대했고, 뒤이어 AI 학습을 위한 Zixiao v2Pro를 출시하며 미국 제재에 대응했습니다.

화웨이 또한 2019년 어센드(Ascend) 910을 시작으로 AI 전용 맞춤형 반도체를 개발, 미국 제재 이후에는 SMIC과 협력해 신제품을 개발하였습니다. 2023년 8월 화웨이와 iFLYTEK은 SMIC의 N+2 7nm 기반 어센드910B를 탑재한 'StarDesk AI 워크스테이션'을 출시했으며, 출시 이후로 수 백 개에서 2만 개에 걸친 클러스터 규모로 어센드910B을 10만 개 이상 배포했습니다.[2]

도표 11

맞춤형 반도체를 활용하는 대표적 업체 사례

	업체명	목적	반도체명	주요 기능
영상처리용	Google	Video processing unit	Argos	Video processing and encoding
	Meta	Video processing unit	Meta scalable video processor	Video processing and encoding
	Tencent	Video processing unit	Canghai	Video processing and encoding
	Netint	Video processing unit	G4/5; T400	Video processing and encoding
네트워크 및 보안	aws	Data processing unit	Nitro	Data plane processing Encryption and other network functions
	Cisco	Networking	UADP and Silicon One	Networking, switching and security
AI 맞춤형	Google	AI accelerator	TPU v4/5	AI processing (training and interference)
	aws	AI accelerator	Tranium; Inferentia	AI processing
	Microsoft	AI accelerator	Maia	AI processing
	Meta	AI accelerator	MTIA	AI processing
	Tencent	AI accelerator	Zixiao V1/V2	Image and speech recognition
	Huawei	AI accelerator	Ascend 910B	AI workloads
	Baidu	AI accelerator	Kunlun	AI computing
	Alibaba	AI accelerator	Hanguang 800	AI inference

Source: Public available information

맞춤형 반도체의 새로운 물결: 보안, 웹 서비스, 데이터베이스, 데이터 분석

맞춤형 반도체의 차세대 흐름은 보안, 웹서비스, 데이터베이스 및 분석 부문의 성능 및 계산효율 극대화 니즈가 이끌 것입니다. 데이터베이스 처리용 반도체 칩을 활용하면 쿼리 응답 시간은 줄이고 서버 당 처리 가능 사용자 수는 늘릴 수 있습니다. 마이크로소프트는 데이터 분석에 FPGA(Field Programmable Gate Array)를 적용한 반도체를 활용하는 초기 연구를 진행중입니다. 한편 보안 전용 IC 개발은 디지털 경제 확산으로 촉진되고 있습니다. 이미 소버린 클라우드 운동(Sovereign Cloud Movement) 등으로 보안용 IC 개발의 요건들이 정립되고 있습니다.

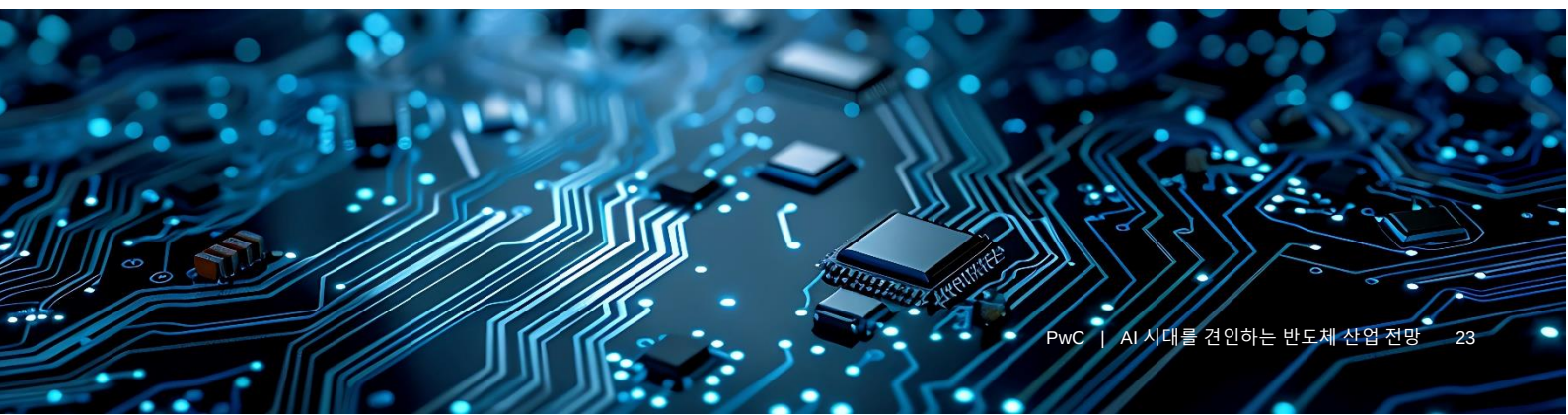
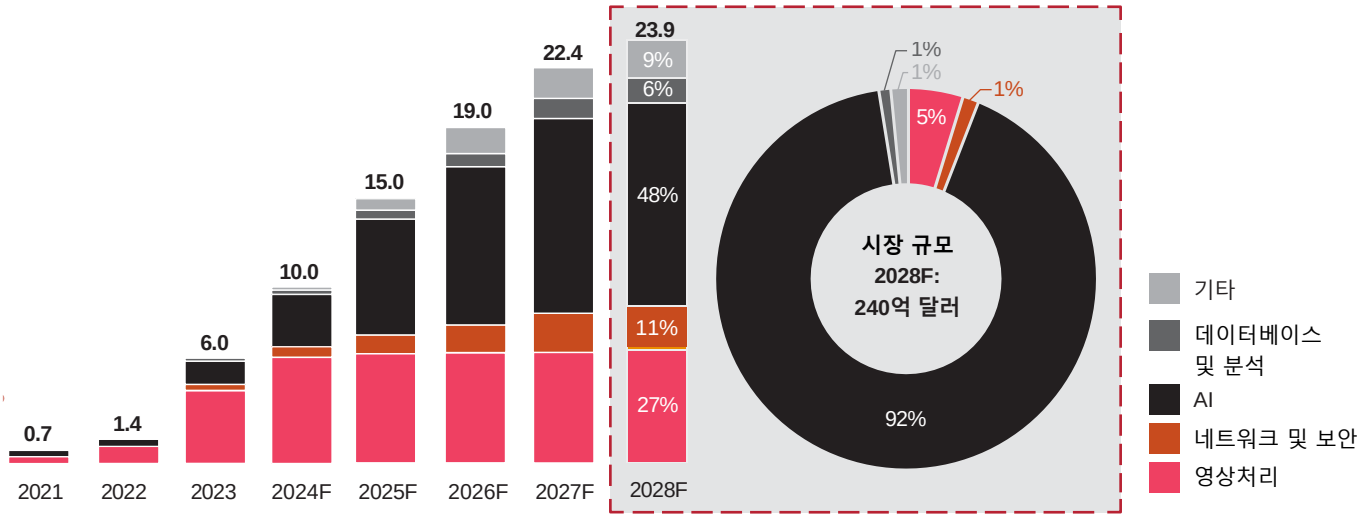


도표 12

데이터센터용 맞춤형 반도체 수요 및 시장 규모 (단위: 백만 개 / 십억 달러)



Source: Omdia Q3 2024

컴퓨팅 시장 구도에 미치는 영향

옴디아에 따르면 맞춤형 반도체 수요는 2028년 기준 연간 약 2,500만 개에 달할 것이며, 이러한 칩의 처리 능력을 활용하는 서버는 시스템 당 10개 이상의 칩을 장착할 것입니다. 그에 따라 연간 서버 배포량은 2028년 기준 200만 대를 초과할 것으로 보입니다. 2023년과 2024년에는 영상 처리용 칩이 가장 많이 장착될 것이며, AI 칩은 과거 투자에 대한 성과를 드러내기 시작할 것입니다. 통상적으로 AI 칩은 고비용의 HBM 메모리와 수많은 컴퓨팅 코어를 포함하고 있어 영상 처리 칩이 더 저렴합니다. (도표 12참조)

데이터 센터용 맞춤형 반도체 시장규모는 2028년 말 기준 약 240억 달러에 이를 것으로 예상되며(도표 12참조), 이러한 성장은 엔비디아와 같은 선도사와 경쟁할 수 없었던 반도체 설계, 제조사에게 새로운 기회가 될 것입니다. 브로드컴, 마벨(Marvell), 인텔 등은 데이터 센터 분야에서 증가하는 맞춤형 반도체 수요를 잡고자 전략을 수립하고 있습니다.

2028년 기준
데이터센터용 맞춤형
반도체 시장 규모는 약

$240억

맞춤형 반도체의 자체 제작 시도는 데이터 센터 분야에서 시작해 자동차, 헬스케어 등 타 산업까지 확산되고 있습니다. 자동차 산업에서는 다수 업체들이 엔비디아, 인텔, 화웨이 등과 파트너십을 통해 기성 프로세서의 최적화를 추진중입니다. 테슬라는 안전 관련 기능을 포함한 자율주행용 신경망 처리 프로세서를 설계하였고, 덴소는 자동차용 맞춤형 프로세서를 제작해 전세계 자동차 제조사에 공급하고 있습니다.[10] BYD는 배터리 관리 시스템(BMS), 파워트레인 제어, 실시간 센서 데이터 처리를 위한 MCU 생산 역량을 내재화하고 있습니다.

업체들이 기존 프로세서의 성능 최적화에서 자체적인 칩 설계 및 개발로 전략을 수정함에 따라, 2028년 기준 158억 달러에 달하는 자동차 SoC 매출의 상당 부분이 맞춤형 반도체로 구성될 것으로 예상됩니다.

선택의 기로: 파트너십과 자체 역량 구축

각 사는 전략적 필요성, 운영 규모, 원하는 기술 통제 정도를 고려하여 맞춤형 반도체 개발 여부를 선택해야 합니다. 자사가 보유한 전문 지식과 리소스 등 내부 역량을 고려해 자체 칩 개발을 수행할 수 있는가, 혹은 외부 파트너와 협력이 적합한가를 결정해야 하는 것입니다. 맞춤형 반도체의 이점은 얻으면서 리스크를 줄이고자 하는 기업은 팹리스 또는 IDM과의 파트너십을 고려할 수 있습니다. 완전 맞춤화 서비스부터 공동 설계까지 파트너십의 형태는 다양합니다.

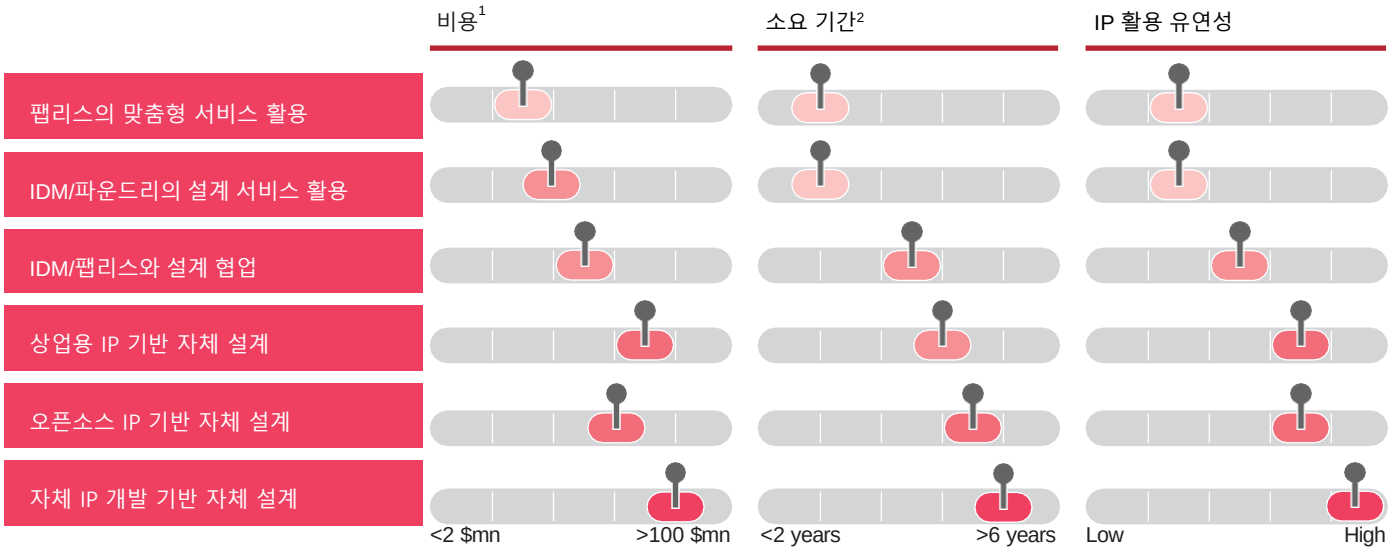
내부 역량 수준이 높은 경우 맞춤형 반도체를 자체 개발하는 것이 이익이 될 수 있습니다. 전력 소비 최소화, 성능 고도화 등 목표와 용도에 맞는 칩을 개발할 수 있으며, IP에 대한 통제력을 강화해 기술을 보호하기 용이해집니다. 나아가 IP 라이센싱, 니치 시장 대상의 신제품 개발 등 신사업 기회도 얻을 수 있습니다.

반도체의 맞춤화로 소프트웨어와 하드웨어는 더욱 긴밀하게 통합되며, 각 사용처에 최적화된 형태로 발전하고 있습니다. 기존 하드웨어 전문 기업들도 소프트웨어 중심으로 변모하고 있으며, 이는 제품 자체의 경쟁력을 높일 뿐 아니라 새로운 수익 모델을 창출하고 경쟁우위를 구축하는 기반이 될 것입니다.

> **“이러한 하드웨어와 소프트웨어 간의 시너지는 전통 하드웨어 제조업체들이 점차 소프트웨어 기반 업체로 탈바꿈하게 되는 등 산업 내 근본적인 변화의 바람을 불러 일으키고 있습니다.”**
>
> **Tom Archer**
> Global Technology Consulting and Alliances Leader, PwC US

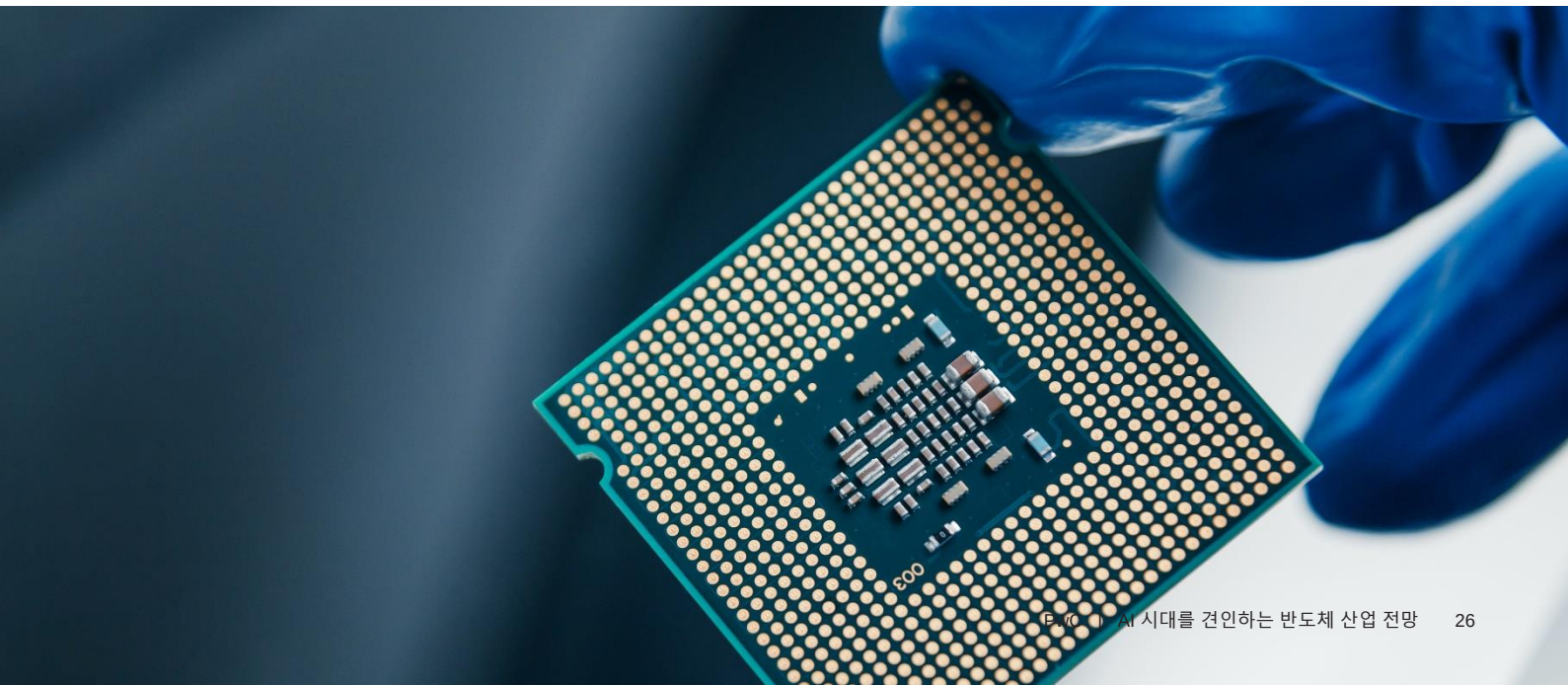
도표 13

맞춤형 반도체 개발 방식의 평가



1 세부 비용은 반도체 기술 (노드), 제품 유형 및 공급업체에 따라 상이함
2 필요 역량 확보 위한 빌드업 단계 포함
Source: PwC analysis

결국 기업의 선택은 기업의 목표, 어플리케이션의 복잡성과 난이도, 맞춤화와 빠른 시장 진출 사이의 균형에 달려 있습니다. 기성 혹은 오픈 소스 IP를 사용하면 개발 일정은 단축하고 일정 수준의 유연성은 유지할 수 있으며, 소규모, 중견 기업은 기존 업체와의 파트너십을 통해 효율적으로 시장에 진입할 수 있습니다. 대기업은 원하는 IP를 자체 소유함으로써 용도에 맞는 최적화된 솔루션을 개발할 수 있습니다. (도표 13참조)



Section 6

AI의 일상화와 AI 반도체 공급의 다양화

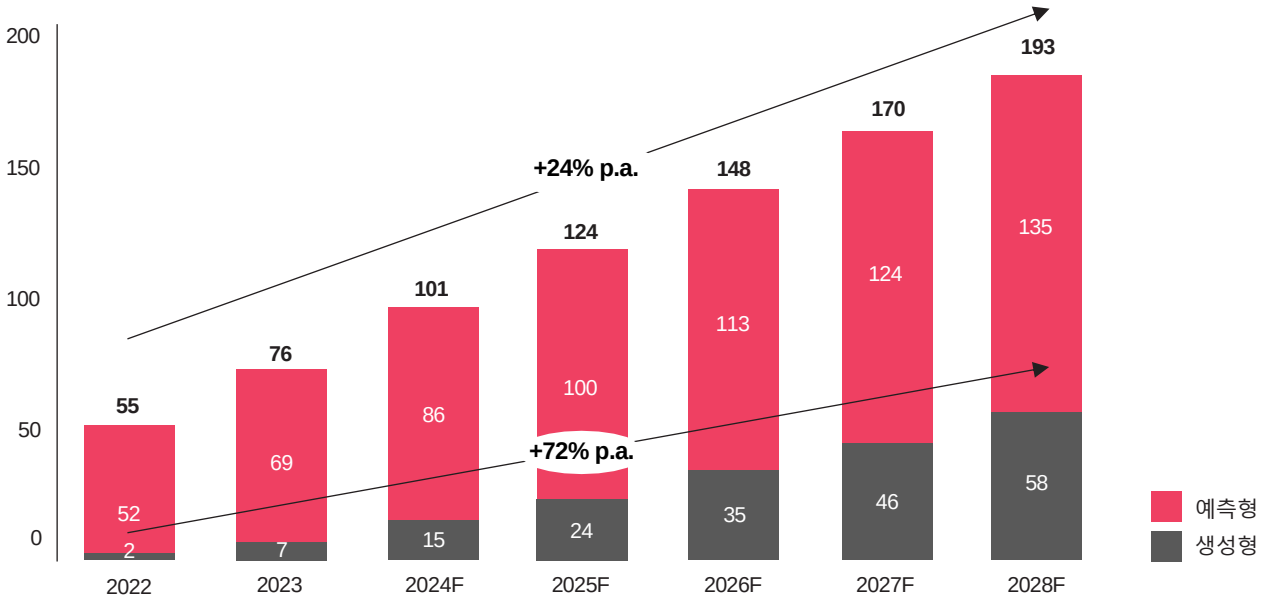
AI는 지역, 산업을 불문하고 빠르게 적용되고 있으며, 반도체 산업에도 두 가지 거대한 기회를 제공합니다. 첫 번째, 반도체 칩의 설계와 공정 최적화 과정에 AI를 적용해 효율을 높이고 오류는 줄이며 빠르게 제품을 출시할 수 있습니다. 두 번째 예측형 및 생성형 AI으로 대표되는 AI 애플리케이션의 성장에 따라 고성능 반도체 시장 수요의 폭발적 확대를 기대할 수 있습니다.

예측형 AI는 자동 품질검사 및 공급망 최적화 등 과거 데이터 기반의 예측 및 비즈니스 프로세스 개선에 활용되며, 그 시장규모는 2028년에는 약 1,350억 달러에 달할 것으로 예상됩니다. (도표 14참조)

생성형 AI는 학습된 데이터 패턴을 기반으로 새로운 콘텐츠를 생성하는 AI로, 태동기를 지나 이미 대중 시장으로 확산되고 있습니다. 2028년 기준 시장규모는 580억 달러로 예상되며, 2023년부터 연평균 성장률 54%로 빠르게 성장할 것으로 전망됩니다. (도표 14참조)

도표 14

예측형 및 생성형 AI 소프트웨어 시장 규모, 2022–2028 (단위: 십억 달러)



Source: Omdia Q3 2024

AI 모델의 규모 경쟁은 2019~2021년 사이 가장 치열했고, 2021년 구글이 1.6조 파라미터의 Switch-C를 발표한 후로 누그러지는 추세입니다.[11] GPT-4나 기타 유사 모델이 Switch-C의 규모를 능가하지는 않고 있지만 전반적인 모델의 사이즈 자체는 증가하고 있습니다. 2023년 2월 메타의 라마(LLaMa) 모델 유출을 기점으로[12], 오픈 소스 커뮤니티에서는 50억에서 700억 파라미터 규모의 '작은 AI 모델'이 확산됐습니다. 이는 일정 수준 이상의 하드웨어만 있다면 연구자나 개인이 충분히 의미 있는 연구를 할 수 있는 규모로, 최근에는 '중간 규모' 모델을 중심으로 혁신이 이루어지고 있습니다.

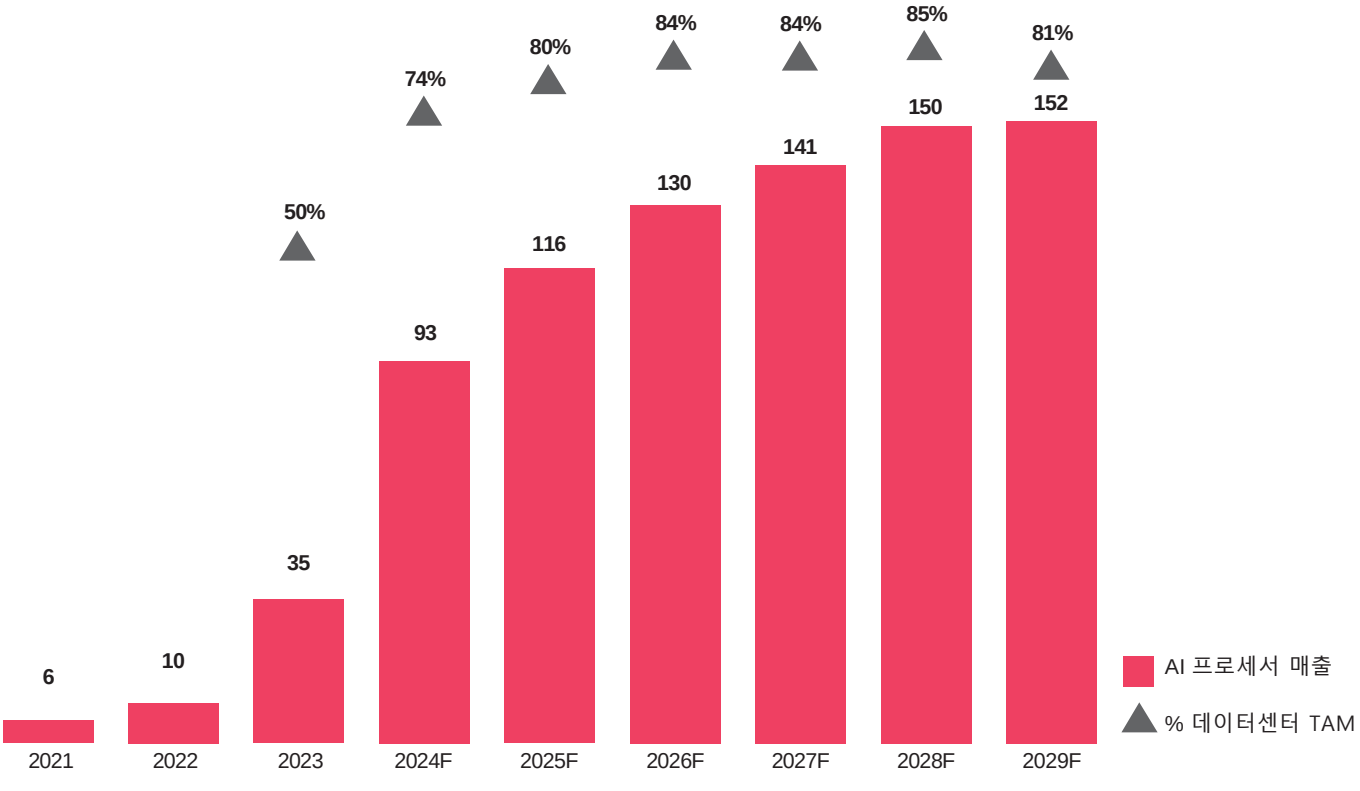
AI 가속기가 이끄는 데이터 센터의 변화

예측형 AI에서 생성형 AI로의 전환, 특히 고성능 반도체를 사용하는 AI 가속기와 HBM의 수요 증가는 반도체의 수요를 견인하는 주 요인입니다. YOLO-v5와 같은 5억 파라미터 이하의 전통적인 머신 러닝 모델 및 소형 신경망은 50~700억 파라미터의 Phi, Gemma, Mistral-7B 등의 대형 모델로 대체되고 있습니다. AI 컴퓨팅 수요 증가에 따라 엔비디아의 데이터 센터 사업 부문 매출은 올해 800억 달러에 달할 것으로 보이며, 해당 매출이 전부 컴퓨팅 혹은 AI 부문 매출은 아니나, 2023년 2분기 이전의 엔비디아 총 매출은 그 절반 이하였음을 고려할 때 놀라운 성과라고 할 수 있습니다.

데이터 센터 분야의 AI 반도체 수요 증가는 메인 프로세서인 CPU와 GPU, ASIC 간의 관계 변화를 불러일으켰습니다. 점차 CPU는 GPU의 로직 보조 프로세서 역할을 수행하거나, 여러 가속기 간 조율 기능을 수행하고 있습니다. (도표 15참조)

도표 15

클라우드 및 데이터센터용 AI 프로세서 시장 규모, 2021–2029 (단위: 십억 달러)



Source: Omdia Q3 2024

2028~2029년 데이터 센터용 프로세서 시장 중 AI 가속기 및 AI 기능 포함 제품의 비중은 80% 이상, 시장 규모는 1,500억 달러 이상이 될 것으로 예상됩니다. 이러한 성장의 배경은 트랜스포머 모델 아키텍쳐의 채택에 있습니다. 본래 기계 번역(Machine Translation, MT)을 위해 설계된 트랜스포머는 현재 거의 모든 영역에서 사용됩니다. 그러나 추론 처리량이 메모리 대역폭에 크게 좌우되며, 필요 메모리 용량 또한 컨텍스트 윈도우 크기의 제곱에 비례해, 때에 따라 매우 커질 수 있다는 한계가 있습니다. 따라서 대역폭과 레이턴시, 특히 예측 가능한 레이턴시(Predictable Latency)는 매우 중요해집니다. 현재 사용되는 AI의 경우 학습과정에서 최소 한 번의 올투올(All-to-all) 커뮤니케이션이 필요한데, 이때 관련 기기 혹은 네트워크 링크 중 가장 느린 속도를 지닌 곳에서 병목이 발생해 작업이 지연될 수 있기 때문입니다.

시장 매출 기준
데이터센터용 프로세서의
80% **이상이**
AI 가속기 및 AI 기능 포함
제품일 것으로 예상

충분한 메모리와 I/O 대역폭을 확보하는 데에는 단순히 컴퓨터 성능(FLOPs)을 확대하는 것보다 각각 약 10배 더 많은 에너지가 소모됩니다. 또한 메모리 용량을 확보하더라도 빠른 속도 역시 필요하기 때문에 단순히 DRAM을 추가하는 것이 아닌, AI 가속기와 HBM의 직접 통합이 필요합니다. 이러한 배경 아래 엔비디아의 B200, AMD의 MI300과 같은 플래그십 GPU가 탄생했습니다.[2] B200은 1킬로와트 이상의 많은 전력을 소모합니다. 따라서 향후 AI 시스템의 운영 비용을 줄이기 위해서는 반도체 단위의 효율 향상이 필수가 될 것입니다.

맞춤형 가속기, GPU의 지위를 흔들다

데이터센터의 전력 수요 증가와 GPU의 높은 자본비용은 맞춤형 반도체와 AI 가속기 스타트업의 성장을 촉진했습니다. 이는 1981년 소니의 CTO 츠기오 마키모토가 설명한 전자 산업의 주기적인 변동, 즉 기술의 최전선에서 일어나는 '맞춤화' 단계와, 수요가 일정해지며 오는 '표준화' 단계의 사이클과도 일맥상통합니다. AI는 이러한 '마키모토의 파도'의 다음 단계를 이끌고 있습니다. IBM, 테슬라, 화웨이, 애플과 하이퍼스케일 클라우드 서비스 업체들은 자체 맞춤형 가속기 ASIC을 도입 중이며, 마벨의 익명의 고객 "C"도 2026년까지 맞춤형 AI ASIC 생산을 확대할 것으로 예상됩니다.[13]

현재 GPU 시장을 지배하는 엔비디아에 견줄만한 기록을 세우고 있는 유일한 분야는 구글의 클라우드 TPU로 대표되는 ASIC입니다. 구글의 TPU와 메타의 MTIA(Training and Inference Accelerator)의 ASIC 아웃소싱 파트너인 브로드컴의 2023년 AI 부문 매출은 전년 대비 3배 이상 증가했습니다. 이는 엔비디아의 데이터센터 사업부문보다도 빠른 속도입니다.

> **"반도체 제품의 효율성을 대폭 개선하는 부분이 앞으로 AI 애플리케이션의 운영 비용을 절감하는데 핵심 과제가 될 것입니다."**
>
> **Kimihiko Uchimura**
> Partner, PwC Japan

반면 AI칩 스타트업은 SDK(Software Development Kit), 각종 도구 및 제반 솔루션 구축에 필요한 인재 유치의 어려움으로 영향력을 발휘하지 못하고 있습니다. 이는 파운드리와 ASIC 협력사에게는 통합 서비스를 제공해 고객의 진입 장벽을 낮추는 비즈니스 기회가 됩니다. 이러한 통합 서비스는 패키징과 리소그래피 등의 공정뿐만 아니라 인접 IP, 특히 소프트웨어 영역까지 확장되어야 합니다.

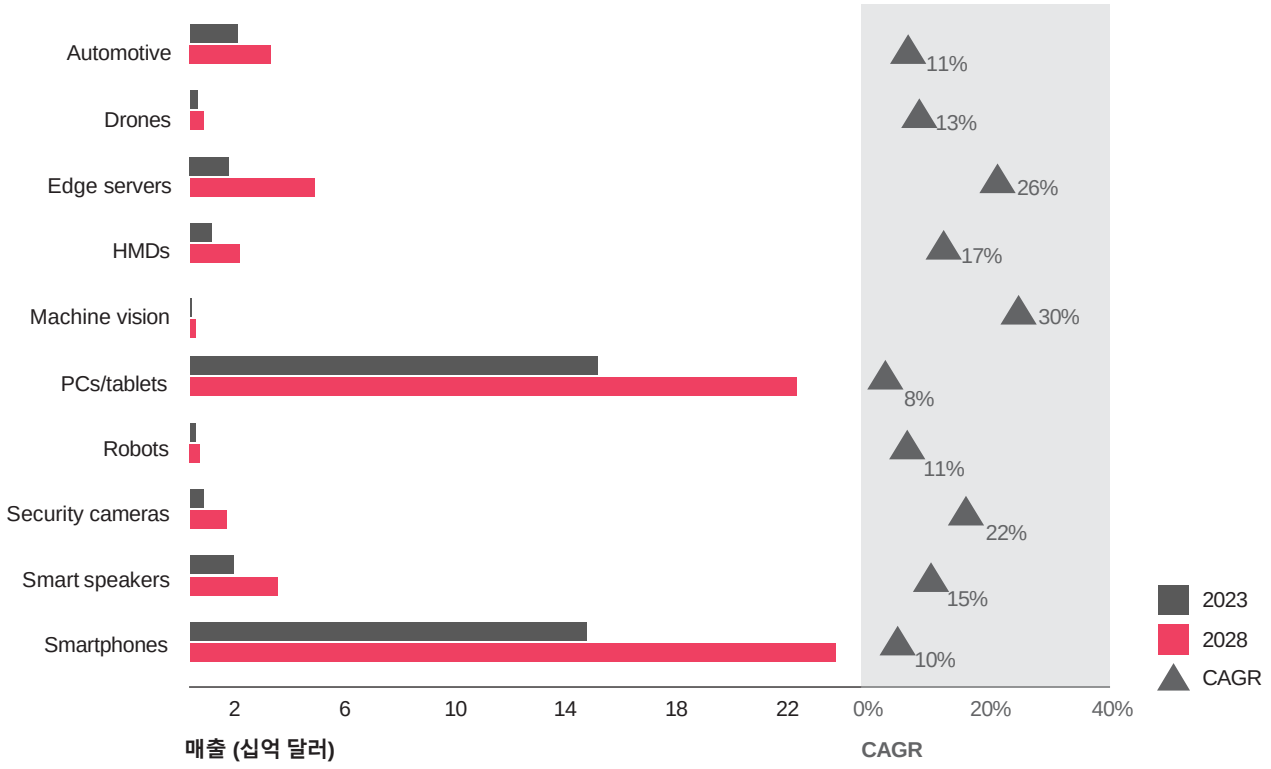
엣지컴퓨팅과 AI: 가속기 반도체의 진화

AI 반도체는 엣지 컴퓨팅, 클라이언트 컴퓨팅 영역까지 확산되고 있습니다. 스마트폰 부문에 가장 먼저 적용되기 시작하였으며, 2017년 애플과 퀄컴이 AI 가속기 코어를 SoC에 통합한 것을 시작으로 2023년 기준 약 66%의 스마트폰이 일종의 AI 가속 기능을 갖추고 있습니다. 이제는 120달러 이하의 보급형 기기도 AI 가속 기능을 탑재할 정도로 범용적 기술이 되었습니다.

퀄컴과 같은 선도사는 70~100억 파라미터 모델을 구동할 수 있을 정도로 모바일 기술이 발전하는 동안 PC 부문은 다소 뒤쳐져 있었습니다.[14] 인텔의 메테오 레이크(Meteor Lake) CPU 출시 전까지 PC 부문의 AI 가속기 탑재는 GPU 기반 게이밍 디바이스 혹은 애플 실리콘(Apple Silicon) 프로세서를 탑재한 맥제품에 그쳤습니다. 그러나 이후 인텔, AMD, 퀄컴은 연이어 AI 가속 기능을 갖춘 PC 전용 CPU를 발표했고, 이러한 프로세서들은 점점 스마트폰의 SoC 설계와 유사해지고 있습니다. (도표 16참조)

도표 16

디바이스별 엣지 데이터센터용 AI 프로세서 시장 규모, 2023–2028 (단위: 십억 달러)

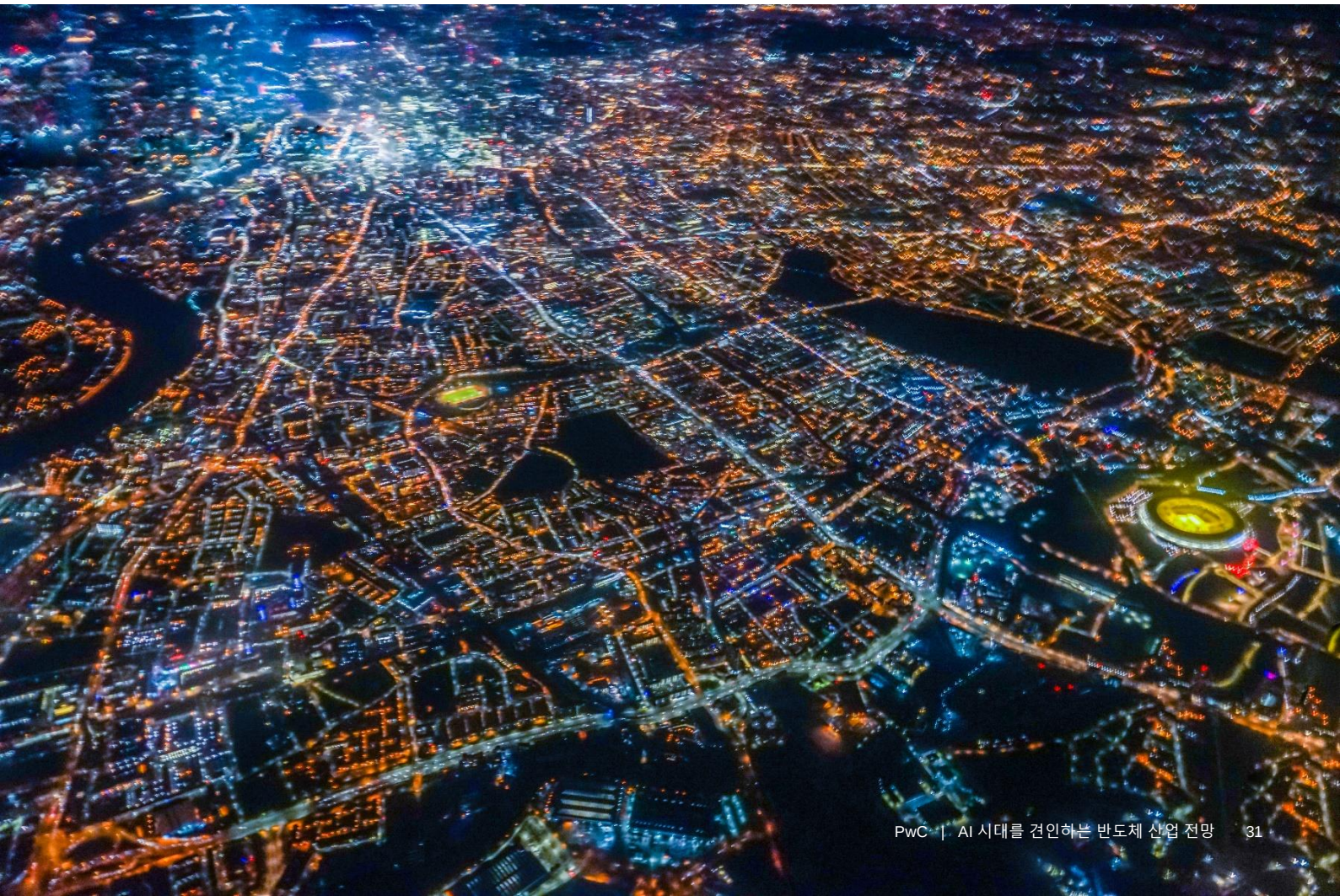


Source: Omdia Q3 2024

떠오르는 과제: 장벽을 극복하고 트랜스포머 모델 이후의 AI를 그리기

향후 몇 년 간 핵심적인 질문은 바로 '얼마나 많은 업체들이 맞춤형 반도체를 채택할 것인가'와 '필요한 최소 투자 금액은 얼마인가'일 것입니다. 흔히 맞춤형 반도체 칩 1종의 테이프 아웃까지 약 12개월의 시간과 5천만 달러의 일회성 비용이 든다고 알려져 있으나, 그보다 진정한 과제는 반도체 칩을 지원하는 소프트웨어 도구를 개발하는 것입니다. 이는 특히 AI 칩 스타트업들이 직면한 주요 과제입니다.

또 다른 질문은 '트랜스포머 모델을 대체할 차세대 모델은 무엇인가' 입니다. 현재 시장에서는 스트립트 하이에나(Striped Hyena)와 맘바(Mamba) 등이 대안으로 거론되고 있습니다. 이들은 트랜스포머의 메모리 집약적인 어텐션 매커니즘을 순환 신경망(Recurrent Neural Network)의 개념에서 차용한 상태 기계(State Machine)로 대체합니다. 또한 1.5 비트 숫자 표현의 3진법 체제로 전환해 트랜스포머 모델의 사용 범위를 확장하려는 시도도 존재합니다. 이러한 시도로 AI가 다시금 CPU 중심으로 전환된다면, 현재의 전세계적인 고성능 GPU에 대한 투자의 영향력도 비교적 축소될 수 있습니다. 결국, 기업의 경쟁우위 유지를 위해서는 지속적이고 깊이 있는 AI 연구를 통해 다가올 기술 변화에 대비해야 할 것입니다.



Endnotes

1. PwC 2024: Electric Vehicle Sales Review Q2-2024
2. Omdia analysis and research Q3 2024
3. AnandTech March 2024: NVIDIA Blackwell Architecture and B200/B100 Accelerators Announced: Going Bigger With Smaller Data
4. Taipeitimes September 2024: Samsung, TSMC collaborating in HBM solutions
5. Trendforce August 2024
6. TrendForce June 2024: SK Hynix's 5-layer 3D DRAM Yield Reportedly Hits 56.1%
7. SK Hynix press release May 2019; SOLIDIGM press release July 2023
8. PwC 2023: Bridging the talent gap
9. European Commission (2022). European Chips Act, retrieved 16th August 2023.
10. Tesla FSD chip: company website
11. William Fedus, Barret Zoph, Noam Shazeer: Switch Transformers: Scaling to Trillion Parameter Models with Simple and Efficient Sparsity
12. Center for Security and Emerging Technology 2023: Meta's Language Models Leak
13. TrendForce April 2024: Marvell's AI Business Reportedly Accelerates, Potentially Benefiting TSMC
14. Forbes: Qualcomm On-Device AI Powers Future Products From Phones To PCs March 2024

About the authors

범용균 부대표는 글로벌 PwC 반도체 부문 리더이자 한국 PwC Strategy& 파트너입니다. Technology, Media, Entertainment 및 Platform 산업 리더이며 전 세계 PwC Strategy& 내 삼성그룹 Global Relationship 담당 파트너로서 활동하고 있습니다. 성장 전략, 투자 전략, 가치 창출 및 디지털 전략 등 다양한 분야에서 고객을 지원하고 있습니다.

PwC Semiconductor Center of Excellence (CoE) 는 한국, 일본, 미국, 유럽 등 PwC 반도체 전문가들로 구성되어 있으며, 글로벌 협력을 통해 다양한 컨설팅 서비스를 제공합니다. 또한, 반도체 생태계 전반에 걸쳐 고객이 직면한 문제에 대한 혁신적 솔루션을 제공하며, 고객의 성장을 지원하고 있습니다.

본 보고서 작성에 필요한 데이터와 인사이트를 제공해주신 Omdia Semiconductor Research에 깊이 감사드립니다.

Contacts

Korea

장유신
Partner,
Strategy& Korea
yoo-shin.chang@pwc.com

이주형
Partner,
Strategy& Korea
tommy.lee@pwc.com

김태영
Partner,
Strategy& Korea
ty.kim@pwc.com

EMEA

Tanjeff Schadt
Partner,
Strategy& Germany
t.schadt@pwc.com

US

Tom Archer
Partner,
PwC US
thomas.archer@pwc.com

Amit Dhir
Partner,
PwC US
amit.dhir@pwc.com

Arup Chatterji
Partner,
PwC US
arup.chatterji@pwc.com

Japan

Kimihiko Uchimura
Partner,
PwC Japan
kimihiko.uchimura@pwc.com

Toshihiro Murata
Partner,
PwC Japan
toshihiro.murata@pwc.com