

Rising Physical AI 시대, K-NPU 기업의 성공 조건

· 조재홍 | 딥엑스 상무 | jhcho@deepx.ai



들어가며: 0.1%의 현실

“한국 NPU(Neural processing unit)의 글로벌 시장점유율은 채 0.1%도 안 될 겁니다. 최근 국내 팹리스 기업 A사가 발표한 1,300만 불 양산품 수출 실적이 NPU 자체로서는 유일합니다.” 며칠 전 정부 관계자가 K-NPU 수출 애로사항 파악을 위해 연락을 해 왔다. 국산 NPU의 시장점유율이 0.1%도 안 된다는 필자의 답변에 관계자는 적잖이 당황한 눈치였다. 우리나라가 반도체 강국인데 ‘NPU 글로벌 시장점유율이 이렇게 낮을까?’ 생각하는 듯했다.

NPU는 인공지능(AI) 반도체의 한 종류이다. 복수의 행렬연산 계산기(Processing Element, Multiply-Accumulate)를 장착하고도 메모리, CPU 등과 효율적으로 통신할 수 있도록 설계된 NPU는 AI 추론을 위해 사용하는 전력이 적고, 열이 덜 발생하는 특징을 지니고 있다.

2025년 과학기술정보통신부의 ‘AI 고속도로’ 구축 사업 추진으로 우리에게 잘 알려진 GPU(Graphic processing unit)는 대규모 데이터를 학습하고 동시에 고성능 추론을 지원한다. 고성능 GPU는 AI 학습 및 연산을 위해 높은 전력 사용을 요구하며, 고전력 소비에 따라 뜨거워진 칩을 식히기 위해 대규모 냉각시설도 함께 요구한다. 비슷한 듯하지만 이처럼 GPU와 NPU는 소비전력 및 기능성 측면에서 확연한 차이점을 지니고 있다.

컴퓨팅 패러다임의 이동: Rising Physical AI

1990년대 개인용 컴퓨터가 본격 보급된 후, 우리는 2000년부터 지금까지 약 30년간 컴퓨팅 파워의 중심이 CPU → 모바일 AP → GPU 등 세 차례에 걸쳐 급속히 전환되는 것을 경험해 왔다. 매 순간 변화의 속도는 우리에게 때론 감탄을 때론 미래의 모습은 어떨지에 대한 궁금증을 일으켜 왔다.

[표 1] 세대별 컴퓨팅 환경 비교

시기	컴퓨팅 환경	핵심 프로세서	적용 영역
1990년대	개인용 PC	CPU	개인 정보 처리
2000년대	서버·인터넷 포털	서버 CPU	중앙 집중형 서비스
2010년대	스마트폰 기반 모바일 컴퓨팅	모바일 AP	저전력·소형화 컴퓨팅
2020년대	생성형 AI와 대규모 학습·추론	GPU	데이터센터·병렬연산
2020년대 후반	현장에서 구동되는 피지컬 AI	NPU(엣지 추론)	효율·분산·임베디드

2026년 지금 우리가 살고 있는 시대는 어떠한 컴퓨팅 패러다임의 혁신이 일어나고 있을까? 필자는 그 해답을 피지컬 AI NPU 전성시대로 표현하고 싶다.

왜 피지컬 AI인가: 전력 한계를 극복하는 유망 기술

피지컬 AI 시장이 주목을 받는 이유는 우리 삶의 현장과 직접적으로 맞닿아 있으며 생활 방식, 산업, 경제 및 사회를 직접적으로 변화시키고 있기 때문이다. 마치 2010년 스마트폰이 처음 등장한 이후 우리 모두가 스마트폰을 사용하게 된 것처럼 미래 우리의 산업, 경제 구조 속에서 피지컬 AI는 도처(Everywhere)에 적용될 전망이다.



피지컬 AI가 국내외 시장에서 주목받는 현실적인 이유를 몇 가지 관점에서 살펴보자. 첫째, 대규모 AI 서비스 운영을 위한 전력과 자원이 부족하다. 국제에너지기구(IEA, International Energy Agency) 발표에 따르면, 전 세계 데이터센터 전력 소비는 2024년 약 415TWh에서 2030년 약 945TWh로 두 배 이상 늘어나 전 세계 전력 수요의 3% 수준에 달할 것으로 분석됐다. 일본 연간 총 전력 소비량과 맞먹는 규모이다.

미국 또한 2030년까지 데이터센터가 소비하는 전력량이 알루미늄·철강·시멘트·화학 등 에너지 다소비 제조업군 전체가 사용하는 소비량을 초과할 것이란 전망이다. 더 심각한 것은 전력 자체보다 생산된 전력을 이송하는 미국 내 전력 그리드의 노후화다. 60년 전에 깔린 미국의 송·배전망을 들어내지 않고는 증산된 전력을 보낼 방법이 없다. 그야말로 전력 생산과 전달 둘 다 진퇴양난의 상황에 빠져 있는 셈이다.

한편, 동일한 추론을 피지컬 AI 디바이스에서 처리하면 총소유비용(TCO)은 클라우드 대비 수십 분의 1 수준까지 내려간다. 칩값만의 비교가 아니다. 고속 통신망, 서버랙과 냉각, 부지비, 지역 민원 비용까지 포함한 결과다. 피지컬 AI가 필요한 AI 서비스를 제공하며 요구하는 자원의 총량은 대규모 AI 데이터센터 서비스 대비 월등히 낮은 셈이다.



둘째, 현장이 요구하는 AI 지능의 수준은 생각보다 낮다. 예컨대, 공장 벽면에 설치되어 24시간 생산 라인을 들여다보는 AI 카메라가 필요로 하는 것은 박사급 추론이 아니다. 지나가는 제품에 스크래치가 있는지 없는지를 실시간으로, 틀림없이 판독하고 이상이 감지되면 메시지 또는 소리로 경고를 내보내는 능력이면 충분하다. 이동 로봇의 장애물 회피, 에어컨의 재실자 인식, 밥솥의 취향 학습, 독거 어르신인 생체리듬 모니터링도 마찬가지다. 대형 AI 모델을 누구나 호출하여 사용하기에는 서비스 지연, 통신·인프라 비용 및 프라이버시 위험 등 총 비용이 기하급수적으로 늘어난다.

셋째, 피지컬 AI 적용 디바이스는 매년 기하급수적으로 증가하고 있다. 시장조사 전문기관인 IoT Analytics 발표에 따르면, 전 세계 연결된 사물인터넷(IoT) 디바이스는 2024년 185억대에서 2025년 211억대로 14% 늘었고, 2030년에는 390억대로 증가하여 2035년에는 약 500억 대를 넘어설 것이란 전망이다.

여기서 중요한 것은 이 수치가 PC와 스마트폰, 태블릿을 제외하고 센서를 집약하는 노드와 게이트웨이만 계산한 값이라는 점이다. 즉, 211억대의 IoT 디바이스가 모두 잠재적인 AI 추론 지점이며, 이 단말들은 사람이 하기 싫어하는 단순, 반복적인 일을 자동화하는 방향으로 끊임없이 AI 기능을 활용하고 있다. IoT Analytics 역시 2030년까지 IoT 시장의 성장동력으로 IoT와 AI가 결합된 피지컬 AI를 지목하고 있다.

반도체 시장의 전망도 같은 방향을 가리킨다. 글로벌 조사 기관에 따르면 피지컬 AI 반도체 시장은 2026년 362억 불에서 2033년 2,918억 불 규모로 연평균 34.7%씩 성장할 것이란 전망이다. 특히 주목할 것은 피지컬 AI 디바이스의 99.8%는 데이터 학습이 아닌 AI 추론용이란 사실이다. 또한 피지컬 AI 디바이스의 80.5%는 1~3W(와트) 전력을 소비하는 범주에 속한다. 시장이 요구하는 피지컬 AI 서비스 스펙이 어디에 있는지를 그대로 보여주는 숫자다.

넷째, AI의 희소성이 연산 중심에서 점차 메모리 영역으로 이동하고 있다. 지난 10년간 AI 학계와 산업의 화두는 ‘어떻게 더 많은 AI 연산 성능을 확보할 것인가’였다. 하지만 지금은 다르다. ‘만들어진 지능을 어디에 저장하고, 얼마나 빨리 메모리에서 AI 모델을 불러오며, 얼마나 적은 에너지로 메모리를 사용할 것인가’로 AI 발전을 위한 질문이 바뀌고 있다.

전문가 혼합 모델(MoE, Mixture of Experts)이 그 전환을 상징한다. 예컨대 DeepSeek-V3는 전체 6,710억 개의 파라미터를 보유하면서도 토큰(AI가 처리하는 텍스트의 최소 단위) 하나를 처리할 때는 약 370억 개 파라미터만 활성화한다. 전체의 18분의 1만 계산하는 셈이다. 그러나 계산하지 않는 파라미터가 사라지는 것은 아니다. 다음 토큰이 어느 전문가를 호출할지 알 수 없으므로, 전체 집합은 메모리에 상주하거나 즉시 불러올 수 있어야 한다. AI가 계산해야 하는 지능의 양은 줄었지만 보유해야 하는 지능의 양은 오히려 늘어난 셈이다.

이 역설은 피지컬 AI 영역에서 훨씬 중요하다. 데이터센터는 보유해야 하는 메모리 양이 늘어나면 메모리 반도체를 서버 랙에 더 꽂으면 되지만, 로봇과 카메라는 전력과 발열, 단가가 모두 고정된 상자 안에서 같은 문제를 해결해야 한다. 게다가 현장에 배치된 장치는 생애의 대부분을 추론으로 보낸다. 학습은 중앙에서 몇 차례 수행하면 되지만, 추론은 모든 현장 장치에서 반복된다.

데이터센터와 피지컬 AI 디바이스는 경쟁 관계가 아니라 역할이 분화되는 관계다. 데이터센터는 학습과 모델 갱신을 담당하는 두뇌가 되고, 현장은 즉시 감지하고 행동하는 에이전트가 된다. ‘데이터센터(학습) → 피지컬 AI 디바이스(추론) → 현장의 행동(에이전트)’으로 이어지는 이 구조가 피지컬 AI가 작동하는 방식이다.

[표 2] 피지컬 AI 도입 동기

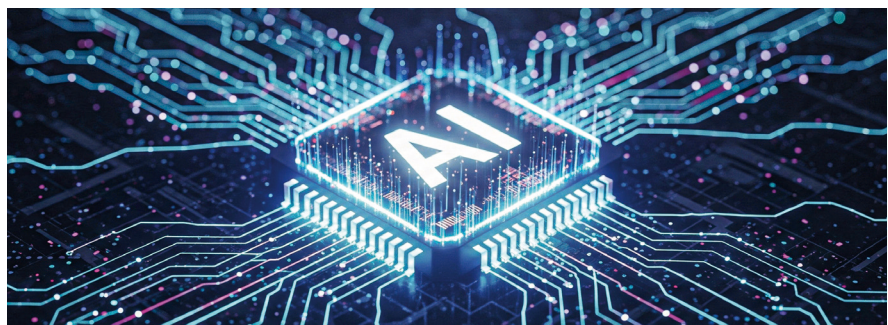
구분	필요성	적용 현장
실시간성	현장 판단 늦으면 사고로 직결	로봇 회피, 차량 보행자 인식
통신 단절 대응	네트워크 끊기면 전 공정 정지	공장 자동화, 선박·오지 설비
프라이버시·보안	공정 영상은 그 자체가 산업 기밀	제조 공정, 의료·교육 영상
비용(TCO)	비 클라우드 전송, 분석 연산 대비 낮은 비용	수만 대 규모의 CCTV·센서망
전력·발열	전원, 냉각 등 물리적 제한	드론, 배터리 기반 로봇, 방산

정리해 보면, 피지컬 AI는 대규모 학습 추론을 수행하는 데이터센터와 협력하여 현장의 일상적이고 반복적인 추론을 효율적으로 수행하는 구조로 가속 성장할 것이다. 예컨대 글로벌 클라우드 서비스 기업인 아마존 클라우드는 피지컬 AI 반도체·서비스 기업과 협력하여 IoT 배포 플랫폼(AWS IoT Greengrass)을 운영하고 있다. 엔비디아 역시 크로노스(Khronos) 그룹과 같은 개방형 표준화 단체에 참여하며 이종 컴퓨팅 프로세서 및 서비스 기업과의 협력을 넓히고 있는데, 이러한 노력은 데이터센터와 피지컬 AI 서비스가 결합한 형태의 AI 서비스 모델이 향후 확대될 것을 보여준다.

한국과 같이 K-AI 소버린 전략을 추진하는 상황에서, 국내 AI 기업은 거인과 같은 글로벌 빅테크와 정면 경쟁이 아니라, 거인이 들어가기 어려운 틈새시장(낮은 소비전력, 피지컬 AI 추론)에서 협력할 수 있는 기회가 생겨난다.

피지컬 AI 구현: AI 반도체(NPU) 기술

그렇다면 피지컬 AI 현장에서 구현하는 NPU는 GPU, 서버형 NPU 등과 무엇이 달라야 할까? 피지컬 AI NPU의 경쟁력은 단일 성능 지표가 아니라 현장의 AI 적용 수요를 잘 이해하고 특성에 맞게끔 최적화된 AI 연산 환경을 제공하는 데서 출발한다. 현장의 수요를 최대한 분석하고 현재 가능한 피지컬 AI 기술로 구현하는 것이 피지컬 AI 반도체 기술의 시작이다.



[표 3] 피지컬 AI 반도체 기술 특성

구분	기술 특성	적용 현장
연산 구조 (MAC 배열·메모리 대역폭)	초당 연산 횟수(TOPS)로 표현되는 추론 처리량 반영	카메라 8채널 동시 판독
전력 및 발열	AI 반도체 작동 환경 조건 (소비전력, 최고/최저 온도 등) 반영	드론·웨어러블·CCTV 탑재 가능 여부
메모리 용량·대역폭	현장에 필요한 AI 모델 크기 및 정확도 반영	CV, VLM, VLA 등

먼저, AI 모델이 처리해야 하는 연산 환경을 부동소수점(FP, Floating Point)에서 정수 8비트, 나아가 4비트로 낮추면(양자화 기술) AI 반도체 도입 비용은 크게 감소한다. AI 반도체 구현을 위해 필요한 메모리 규모와 전력 소비가 줄어들기 때문이다.

하지만 단순히 비트 수를 줄이게 되면 AI 성능(판독·분류·생성 정확도 등)은 열화된다. 양자화는 단순히 AI 모델을 압축하는 기술이 아니라 모델 내 특정 계층의 가중치를 어느 정도 추상화해도 현장 성능이 유지되는가를 판단하는 설계·컴파일러 기술의 집합이다. AI 반도체 기업 간 기술력이 극명히 구별되는 지점이기도 하다.

흔히 AI 반도체에 관한 오해가 하나 더 있다. 고성능 GPU를 적당히 줄이면 피지컬 AI 프로세서가 될 것이라는 생각이다. 하지만 현장은 다르게 말한다. 로봇용 고성능 GPU가 100W를 효율적으로 쓴다고 하더라도, 팬을 달 수 없는 밀폐형 산업용 PC나 배터리로 구동되는 로봇에는 고전력 GPU가 들어갈 수 없기 때문이다.

반도체 공정 미세화(7나노, 5나노 등)만으로는 전력 효율이 급속히 개선되지 않는다. 저전력 NPU 설계 구조를 처음부터 만들어야 하는데 이러한 과정은 기술 병목도 심하고 개발 비용이 높아 조기 시장 접근이 필요한 현대 AI 반도체 시장 특성과는 거리가 멀다.

국내는 이미 일부 팹리스 기업을 중심으로 삼성전자 5나노 공정에서 5W 전력으로 최대 25 TOPS를 구현하는 NPU가 양산되고 있다. 특히 1장의 웨이퍼에서 불량 없이 생산되는 AI 반도체 양산 수율이 90% 수준을 넘어 완성도 높은 기술력을 확보한 것으로 평가받고 있다.

해외 고객사가 직접 수행한 비교 평가를 보면 의미가 더 분명해진다. 동일한 영상·문자인식 알고리즘을 수백 와트급 서버 GPU와 엄지손톱만 한 피지컬 AI 전용 NPU에서 돌렸을 때, 전력은 수십 분의 1이면서도 알고리즘에 따라 성능은 오히려 두세 배가 나오기도 했다. 예컨대 공장에 설치되는 산업용 PC는 분진 유입을 막기 위해 팬을 없앤 밀폐 구조가 많고, 내부 온도가 이미 40~50℃에 이르기도 한다. 그 안에서 칩 온도 85℃를 넘기지 않고 몇 년을 버티는 것, 이것이 현장이 요구하는 AI 반도체 품질이다.

이외에도 메모리와 반도체를 단일 구조로 설계하는 것, 메모리와 MAC 간 데이터 이동을 최소화하며 AI 연산을 최적화하는 설계 기술 등이 피지컬 AI 반도체가 요구하는 핵심 기술로 분류될 수 있다.

[표 4] 피지컬 AI 반도체 적용 분야(예시)

영역	적용 분야(서비스)
로보틱스	배송 로봇의 보행자 회피와 수령인 얼굴 인식, 휴머노이드 구동
스마트팩토리·제조	공정 불량 검사, 물류 분류와 적재 균형 제어
영상감시·보안	매장 재고·도난 탐지, 교실과 공공시설 안전 모니터링
스마트모빌리티·ITS	고속도로 번호판 인식, 자율주행 보조 인지
의료	내시경 영상에서 수 픽셀 크기의 초기 병변 검출
AIoT·엣지컴퓨팅	5G 중계기 내 영상분석, 회의 시스템의 화자 식별, 문서 스캔
농축산·환경	가축 체중·질병 모니터링, 농작물 생육 관측

AI 반도체 생태계 성장 및 애로사항: 정책적 지원과 기업의 도전

AI 반도체 산업은 높은 기술개발 난이도로 진입장벽이 높을 뿐만 아니라, 장기간의 개발 기간 및 대규모 자본 투자가 요구되는 첨단 산업이다. 이 장에서는 AI 반도체 생태계 가치사슬 관점에서 어떤 과정을 거쳐 AI 반도체가 설계를 거쳐 제품화에 이르는지 살펴보고자 한다.

[표 5] AI 반도체 기술 생태계 및 기능

IP	EDA	설계	칩 제작	패키징/테스팅	검증 및 사업화
- NPU 코어 아키텍처·명령어 세트 자체 개발 - CPU, 메모리 컨트롤러, 인터페이스(PCle·LPDDR·MIPI) IP 라이선스 확보	- 합성·배치배선(P&R)·타이밍/전력 분석 툴 라이선스 확보 - 시뮬레이션 및 에뮬레이션 검증 환경 구축 - 공동활용 플랫폼·설계지원센터 활용으로 초기 비용 절감	- 목표 전력·성능·면적(PPA) 정의 - 연산기 배열 및 메모리 계층 구조 설계 - 양자화(INT8/INT4) 및 컴파일러·SDK 동시 설계 - 기능 검증과 다이 면적 최적화(수율 확보)	- 공정 선정 및 PDK 적용, 마스크 제작(테이프아웃) - MPW로 시제품 웨이퍼 제작 - 폴마스크 전환 및 웨이퍼 양산 - 수율 분석과 공정 피드백	- 패키지 형태 결정 및 열·전력 방출 설계 - 웨이퍼 테스트(EDS)와 최종 테스트(FT) - PCB·M.2 등 모듈 보드 실장 - 신뢰성 인증(-40~85℃, 진동·EMC)	- 브링업, 드라이버·펌웨어 안정화 - 고객 모델 포팅 및 PoC·현장 실증 - 고객사 QA 통과와 양산 승인 - 유통·모듈 파트너 확보, 해외 레퍼런스와 반복 주문

AI 반도체 설계는 CPU 등 다양한 이종 컴퓨팅 프로세서 간 호환성을 확보하는 것으로부터 시작한다. 첫째, IP(Intellectual Property, 설계자산) 부문은 NPU 코어와 함께 칩에 들어가는 CPU, 메모리 컨트롤러, 인터페이스 회로를 공급하는 역할을 맡는다. 고도의 기술 안정성과 상호운용성 검증이 요구되는 영역이 IP 분야이다. 국내에도 IP를 공급하는 기업이 있지만 대부분 Arm을 비롯한 소수의 글로벌 기업이 IP를 과점적으로 공급하고 있다. 시제품 제작 단계 시 초기 IP 라이선스 비용 부과는 물론 양산품 판매 시에는 매출액 기준 로열티까지 붙는데, 양산 물량을 아직 증명하지 못한 팹리스 스타트업은 IP 기업과 협상 테이블에서 불리한 조건을 받아들이게 된다. 국산 IP 풀이 얇아 대체 선택지가 없다는 것도 구조적 애로다.

둘째, EDA(Electronic Design Automation, 전자설계자동화툴) 부문은 설계와 검증을 가능하게 하는 소프트웨어 도구를 제공한다. 시놉시스, 케이던스, 지멘스 등 3사가 과점하는 EDA 시장은 필수 라이선스 하나가 수십만 불에 이르기도 한다. AI 반도체 설계부터 검증, 테스트까지 반도체 설계를 위한 핵심 작업이 글로벌 벤더의 통제권 위에서 이루어지며 국내 팹리스 기업은 대개 글로벌 벤더에게 충성 고객이 된다. 연간 라이선스 비용 또한 협상을 하지만 공급자 우위의 시장이다 보니 제시되는 가격을 국내 팹리스 기업이 받아들이는 형국이다. 정부는 스타트업을 대상으로 EDA 공동활용 플랫폼을 제공하여 초기 시장 진입장벽을 낮춰 주지만 지원 규모 및 대상 기업 수는 여전히 제한적이다.

셋째, 설계 부문은 목표 전력·성능·면적을 정의하고 연산기 배열과 메모리 계층, 양자화와 컴파일러를 함께 설계하는 생태계의 핵심으로 팹리스 기업이 속해 있는 영역이다. 국내에는 AI 반도체 설계 경험이 있는 전문인력이 절대 부족하다. 또한, 글로벌 기업은 물론 국내 대기업과의 임금 보상 격차 탓에 확보한 인력을 지키는 것도 쉽지 않다. 또한, 설계에서 양산까지 통상 5년 이상 매출 없이 버텨야 하므로, 인력과 자본을 동시에 붙들어 두는 것 자체가 애로가 된다. AI 반도체 생태계 전체에서 가장 큰 개발 위험을 감당하는 분야이다.

넷째, 칩 제작 부문은 파운드리와 마스크 업체가 담당한다. 여기서 팹리스는 생산물량 기준 후순위에 놓인다. 선단 공정의 생산능력은 대량 고객에게 먼저 배정되고, 마스크 비용은 회차당 수십억 원에 이른다. 다품종 소량이라는 AI 반도체의 초기 특성이 파운드리의 수익 구조와 어긋나기 때문에, MPW(Multi project wafer) 일정을 확보하는 일조차 경쟁이 된다.

다섯째, 패키징·테스팅 부문은 칩을 보드에 실장 가능한 형태로 만들고 동작을 확인한다. 국내 후공정 역량은 범용 패키지에 강점이 있지만, 고집적·고방열이 요구되는 구조는 해외 의존도가 높다. 영하 40도에서 영상 85도 구간의 신뢰성 인증과 EMC(Electromagnetic Compatibility, 전자파 적합성) 시험 비용도 대부분 기업 부담으로 남는다.

끝으로, 검증 및 사업화 부문은 수요기업과 유통·모듈 파트너가 함께 참여하는 마지막 관문이다. 애로는 여기서 가장 크다. 국내 수요기업은 팹리스를 전략적 파트너보다 기술 공급 하청으로 인식하는 경향이 있고, 칩을 전달한 뒤 고객사 QA와 양산 승인을 거쳐 첫 상용 주문이 나오기까지 빨라야 9개월, 길면 18개월이 걸린다. 경험칙으로 100개 기업에 사전 평가(PoC: 개념검증)를 수행해야 3~4곳이 최종 구매를 결정한다. 결국 PoC 대상을 글로벌 기업으로 확대해야 하는데, 글로벌 비즈니스 비용은 개발비와 별개로 발생하며 정책적 지원이 가장 덜한 영역이다.

이처럼 AI 반도체는 IP 도입, 라이선스 구매, 제품 제작, 패키징·테스팅 및 PoC 등을 통해 제품 제작이 완성된다. 통상 5나노급 선단 공정을 활용하는 AI 반도체 제작은 5~7년이 소요되며 개발 비용은 1천억 원 이상이 소요된다.

과학기술정보통신부를 비롯한 정부는 국내 AI 반도체 기업의 건실한 성장을 위해 2019년부터 꾸준한 R&D, 실증 지원, 사업화 정책사업을 추진하고 있다. EDA 툴 관련 시놉시스·케이던스·지멘스가 과점하는 이 시장은 라이선스 하나가 수십만 불에 이르기도 하는데, 정부는 2019년 EDA 공동활용 플랫폼을 AI 반도체 클러스터 내에 구축하여 초기 팹리스의 고비용 설계툴 활용을 지원하여 진입장벽을 크게 낮췄다.

검증 단계에서도 에뮬레이터 등 고가 장비를 공동활용하는 시스템반도체 개발지원센터가 문을 열었고, 여러 기업의 설계를 한 장의 웨이퍼에 함께 올려 시제품 비용을 나누는 MPW 사업도 지속 지원하고 있다. 최근에는 기업 수요를 반영하여 10나노 이하 선단공정에 대한 지원도 확대하고 있다.

이러한 정책적 노력의 결실로 국내 한 팹리스 기업은 칩 단가 50불, 메모리를 결합한 M.2 카드 100~150불 수준으로 양산하며 1,300만 불 이상을 10개국 이상에 수출하는 데 성공하였다. 정부의 긴 호흡에 걸친 지원과 불굴의 스타트업이 함께 만들어낸 성과이다.

마지막 관문: 수출 레퍼런스 구축

양산에 성공한 AI 반도체가 글로벌 수출까지 하며 기술력과 시장성을 인정받고 있는 반면 국내 대기업 등 대규모 수요처의 K-NPU에 대한 수요는 여전히 미온적이다. 초기 위험 투자를 회피하려는 기업 문화가 국내에는 여전히 존재하고 있으며 사업 내 수직계열 구조에 익숙한 국내 대규모 수요처 등은 스타트업인 국내 NPU 기업과 협력에 적극적으로 나서지는 않는 모양새다.

반면 해외 수요처는 급속히 성장하는 AI 시장을 선점하기 위해 첨단 기술력을 지닌 국내 팹리스 기업과 협상을 가속화하고 있다. 앞서 언급한 국내 팹리스 기업은 양산 1년 만에 10여 개 국가에서 77건, 1,300만

불 이상의 상업용 구매 주문(PO)을 확보했고, 누적 매출의 절반 이상이 해외에서 발생했다. 특히 전체 77건 PO 중 62%에 해당하는 48건이 최근 3달에 집중됐다는 점이 흥미롭다. AI 반도체 수출은 선형이 아니라 계단식으로 오른다는 특징을 보여주고 있기 때문이다. 초기 레퍼런스를 넘는 순간 동반 주문과 반복 주문이 비교적 쉽게 이어지기 때문이다.

피지컬 AI 반도체 성공 조건

본고를 통해 우리는 피지컬 AI 시장에 대한 성장 추세 및 기술 원동력에 대해 살펴보았다. 약 211억 대가 연결된 현장 디바이스는 AI 기술의 발전에 따라 로봇틱스, 스마트 제조, 자율형 모빌리티, 첨단 국방 등 다양한 영역에 걸쳐 적용 분야를 확대하고 있다.

피지컬 AI는 양산 국면으로 전환
피지컬 AI 산업을 커버할 수 있는 하드웨어 플랫폼이 필요

"초저전력 AI 칩 없는 피지컬 AI는 존재할 수 없습니다"

초저전력 AI 칩 솔루션

DX-M1
25TOPS / 5W

DX-M1M
25TOPS / 3W

DX-AIBOX
25TOPS / 5W

엣지 AI 서버 솔루션

DX-H1 V-NPU
50TOPS / 30W

DX-H1 Quattro
100TOPS / 20W

로봇틱스

스마트 모빌리티

방산

스마트 팩토리

반도체

조선

스마트 가전

AI 헬스케어

로봇, 스마트팩토리, 스마트모빌리티, 리테일, 의료, 국방 등 여러 산업에서 피지컬 AI 수요가 동시에 발생

세계 1위의 메모리 반도체 강국인 우리나라는 비메모리 반도체(AI 반도체 등) 영역에서도 글로벌 경쟁력을 확보하기 위해 다양한 정책적 노력을 확대하고 있다. 소중한 정부의 노력 덕분에 국내에도 딥엑스(피지컬 AI), 퓨리오사AI(서버형) 등 양산급 AI 반도체를 양산하는 기업이 등장하였다. 참으로 괄목할 만한 성과이다.

하지만 높은 투자비용과 장기간 R&D가 동반되는 AI 반도체 시장에서 국내 팹리스 기업이 글로벌 시장으로 진출하기 위해서는 글로벌 가격 경쟁력을 갖추어야 한다. 또한 압도적 기술 경쟁력을 확보해야 한다.

2019년부터 정부는 AI 반도체 산업 육성을 위한 정책 수요조사, 핵심 기술 분야에 대한 단계별 R&D 및 사업화 적시 지원 정책 등을 속도감 있게 추진해 왔다. 이러한 정책적 노력 덕분에 국내에도 글로벌 시장에 도전하는 AI 반도체 팹리스 기업이 탄생하였다. 어렵게 싹틔운 첨단 산업 내 기술 기업이 거침없이 글로벌 시장에 도전하고 글로벌 빅테크와 협력하며 새로운 시장을 개척해 나가야 할 골든 타임이 바로 지금이다.

그렇다면 K-NPU 기업의 마지막 성공 조건은 무엇인가. 첫째, 수출용 AI 반도체에 대한 감세 또는 면세다. 경쟁국이 국가 단위 보조금으로 초기 단가를 낮추는 상황에서, 수출 과정에서 발생하는 해외 매출에 대한 세수를 한시적으로 유예해 가격 경쟁력을 만들어 주는 것은 재정 지출 없이 가능한 가장 직접적인 수단이다. 둘째, 글로벌 실증(PoC)을 확대하는 지원이다. 100개 기업에 평가를 수행해야 3~4곳이 구매를 결정하는 산업에서, 데모 보드와 현지 기술 지원, 해외 출장과 전시 비용은 개발비가 아니라 매출을 만들기 위한 선행 투자로 인식되어야 한다. 셋째, 공공 조달 규격의 전환이다. 공공 제안요청서가 외산 고성능 GPU 부품 사양이 아니라 AI 인식 정확도와 응답시간, 소비전력이라는 결과 기준으로 쓰이기만 해도 국산 AI 반도체의 초기 레퍼런스가 열린다. 넷째, 국내 실증을 수출의 출구로 설계하는 일이다. 실증이 국내 스펙에 머물면 제품도 국내용으로 굳어진다. 설계 단계부터 해외 수요기업을 참여시키고, 실증을 종착지가 아니라 구매와 양산으로 이어지는 조건부 단계로 두어야 한다. 지금까지의 정책이 AI 반도체를 만드는 일을 도왔다면, 이제 필요한 정책은 칩을 세계 시장에 파는 일을 돕는 것이다. 산업 육성의 목적은 시제품 제작이 아니라 매출이기 때문이다. AI 핵심 기술로 주목받고 있는 피지컬 AI 반도체 시장에서 K-NPU 기업이 거센 돌풍을 일으키길 기대한다.

참고문헌

- 김낙원, <AI의 희소성은 어디로 이동하는가?>, 딥엑스(2026)
- DeepSeek-AI, DeepSeek-V3 Technical Report
- IoT Analytics, State of IoT 2025
- International Energy Agency, Energy and AI — Executive Summary, 2025
- Grand View Research, Edge Artificial Intelligence Chips Market Report, 2026-2033
- MarketsandMarkets, Edge AI Hardware Market
- Fortune Business Insights, Edge AI Market
- 디지털데일리, “딥엑스, ‘DX-M1’ 양산 1년 만에 1300만불 수주…온디바이스 AI 수출 본격화”, 2026. 8. 14.
- ZDNet Korea, “딥엑스, NPU 양산 1년 만에 글로벌 77건 구매 수주”, 2026. 8. 18.
- 대한민국 정책브리핑, “판교 전역 팹리스에 시스템반도체 설계SW 무료 제공”, 2021. 1. 8.