전자통신동향분석, 41권 6호 2026, 102-115

피지컬 AI를 위한 온디바이스 AI 반도체 플랫폼 기술 동향

Technology Trends in On-Device AI Semiconductor Platforms for Physical AI

박유미 (Y.M. Park, parkym@etri.re.kr) 기업성장지원전략실 책임연구원
기문철 (M.C. Kee, mooncheol.kee@supergate.cc) 수퍼게이트(주) 연구소장

ABSTRACT

Physical artificial intelligence (AI), which autonomously perceives, reasons, and controls real-world environments, is emerging as a major AI paradigm. Because autonomous vehicles and robots operate in safety-critical environments where communication latency or network disconnection can have severe consequences, on-device AI capable of ultra-low-latency and highly reliable processing is essential. Accordingly, recent technological trends have evolved beyond standalone AI accelerators toward vertically integrated on-device AI semiconductor platforms that combine sensor-processing pipelines, functional safety mechanisms, hardware-based security, and comprehensive software toolchains. This paper presents a comparative analysis of recent technological trends in leading global platforms, including Tesla, NVIDIA, and Qualcomm, as well as AI semiconductor technologies developed domestically in South Korea, in terms of architecture, computational efficiency, functional safety, and software ecosystems. The analysis shows that global leaders have gained competitive advantages through tightly integrated hardware–software platform strategies, and that competitiveness depends not only on individual chip performance but also on the completeness of the integrated platform. Based on these findings, this paper assesses the current status of domestic technologies, identifies future development directions, and proposes strategic implications for strengthening the competitiveness and autonomy of the domestic AI semiconductor industry in the era of Physical AI.

KEYWORDS AI 반도체, 온디바이스 AI, 온디바이스 AI 반도체 플랫폼, 피지컬 AI

* DOI: https://doi.org/10.22648/ETRI.2026.J.410609
* 이 논문은 2026년도 정부(과학기술정보통신부)의 재원으로 정보통신기획평가원의 지원을 받아 수행된 연구(RS-2025-02304483, 온디바이스 AI 최적화 칩렛 기반 허브 SoC 개발, 90%)와 2026년도 한국전자통신연구원 내부연구과제(26ZV1100, ETRI 사업화 촉진, 10%)의 지원을 받아 수행된 연구임.

본 저작물은 공공누리 제4유형
출처표시+상업적이용금지+변경금지 조건에 따라 이용할 수 있습니다.
©2026 한국전자통신연구원

I. 서론

최근 생성형 AI(Generative AI)의 발전에 이어, 실세계 환경을 스스로 인식하고 판단하여 행동하는 피지컬 AI(Physical AI)가 차세대 AI 핵심 기술로 부상하고 있다. 피지컬 AI는 그동안 가상 공간에 머물렀던 AI가 자율주행차, 로봇, 드론 등 물리적 하드웨어와 결합하여 현실 세계와 직접 상호작용을 하도록 하는 기술이다. 피지컬 AI 시스템은 다양한 센서를 통해 주변 상황을 인식하고 AI가 이를 분석하고 판단한 뒤 액추에이터를 통해 물리적 제어 동작을 수행한다[1,2].

이러한 피지컬 AI의 '인식-판단-제어' 파이프라인에서는 수십 밀리초(ms) 수준의 통신 지연만으로도 안전사고를 초래할 수 있으므로 실시간 응답성이 핵심이다. 그러나 기존의 클라우드 기반 AI 처리 방식은 물리적 디바이스와 클라우드 서버 간 통신에 의존하기 때문에 실시간성을 만족하기 어려울 뿐만 아니라 운영 비용 증가, 보안 및 개인정보 보호 측면에서 한계를 갖는다. 이에 따라 네트워크의 최종 종단 디바이스에서 AI 추론을 직접 수행하는 온디바이스 AI(On-Device AI) 기술이 이러한 제약을 해결할 수 있는 핵심 기술로 부상하고 있다. 온디바이스 AI는 AI 모델을 디바이스 내부에서 실행함으로써 실시간 응답성을 보장할 뿐만 아니라, 통신이 불안정하거나 단절된 극한 상황에서도 서비스 연속성을 유지할 수 있다. 또한, 데이터의 외부 전송을 최소화하여 개인정보 보호를 강화하고, 네트워크 대역폭을 절감함으로써 통신 자원 활용 효율을 높이는 장점이 있다[2-4].

하지만 피지컬 AI에서는 온디바이스 상에서의 추론만으로는 충분하지 않다. 스마트폰 중심의 모바일 환경과 달리 피지컬 AI 시스템은 센서 처리를 통한 실시간 환경 인식 및 추론, 처리 시간 예측성을 보장하는 결정론적 실행(Deterministic Execution), 시스템 오류 발생 시 안전한 동작을 보장하는 기능 안전(Functional Safety), 외부 위협으로부터 시스템과 데이터를 보호하는 하드웨어 보안(Security), 하드웨어에 최적화된 SDK(Software Development Kit) 등의 개발 환경까지 통합적으로 제공하는 플랫폼 접근이 요구된다. 이러한 요구사항을 충족하기 위해 결정론적 실시간 처리를 위한 제어 구조, 기능 안전과 보안을 지원하는 전용 하드웨어 블록, 비전 및 센서 처리 파이프라인 등을 단일 다이에 통합하고 전용 SDK까지 함께 제공하는 '온디바이스 AI 반도체 플랫폼' 중심으로 기술이 발전하고 있다.

이에 본고에서는 자율주행차, 로봇, 드론 등 고신뢰성이 요구되는 피지컬 AI 도메인을 중심으로 글로벌 선도 기업들의 온디바이스 AI 반도체 플랫폼 기술을 비교 · 분석하고, 국내 플랫폼 기술의 현황을 진단하여 우리 산업 생태계의 경쟁력 확보 전략을 제시하고자 한다.

본고의 구성은 다음과 같다. Ⅱ장에서는 피지컬 AI용 온디바이스 AI 반도체 플랫폼의 하드웨어와 소프트웨어 핵심 요구사항을 분석하고, 이를 기준으로 Ⅲ장에서는 글로벌 주요 플랫폼의 최신 기술 동향을 정리한다. 이어서 Ⅳ장에서는 국내 플랫폼 기술의 현황을 파악하고, Ⅴ장에서는 국외 선도 기술 대비 국내 생태계의 차별화된 경쟁력 확보 방안을 논의한다. 마지막으로 Ⅵ장에서는 본고의 결론을 맺고 향후 피지컬 AI 반도체 플랫폼 기술의 발전 방향을 전망한다.

II. 온디바이스 AI 반도체 플랫폼 요구사항

피지컬 AI를 위한 온디바이스 AI 반도체 플랫폼은 물리적 세계의 역동적인 변화를 실시간으로 처

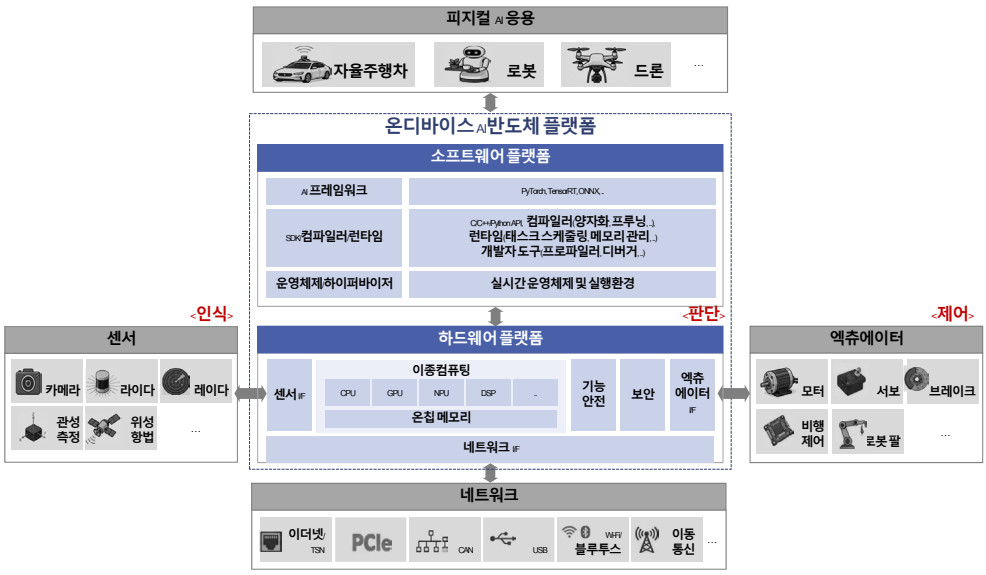

그림 1 피지컬 AI용 온디바이스 AI 반도체 플랫폼 구조

리해야 하므로, 단순한 연산 성능 향상을 넘어 시스템 전반의 고신뢰성을 보장할 수 있어야 한다.

그림 1과 같이 센서 입력부터 액추에이터 제어에 이르는 전 과정은 하드웨어 가속 아키텍처와 소프트웨어 실행 환경의 유기적인 결합을 통해 구현된다. 즉, 저지연 및 기능 안전을 보장하는 하드웨어 구조와 최적화된 소프트웨어 툴체인이 긴밀하게 연계될 때 실제 환경에서 요구되는 성능과 신뢰성을 동시에 확보할 수 있다. 이 장에서는 이러한 유기적 플랫폼을 구축하기 위해 필수적으로 요구되는 하드웨어 및 소프트웨어 플랫폼의 핵심 요구사항을 설명한다.

1. 하드웨어 플랫폼

피지컬 AI에서 요구하는 온디바이스 AI 반도체 하드웨어 플랫폼의 요구사항은 다음과 같다.

첫째, 고성능 AI 연산 기능이 요구된다. 이를 위해 CPU, GPU, NPU, DSP 등 이종 컴퓨팅 자원의 강점을 활용하는 아키텍처와 대용량 온칩 메모리 계층 구조를 통해 높은 연산 성능을 확보해야 한다. 또한, 온디바이스 환경에서 구동되는 AI 모델은 복잡도와 요구 정확도에 따라 최적 데이터 정밀도가 달라지므로, 단일 플랫폼에서 INT4, INT8, FP8, FP16 등 다양한 정밀도의 데이터 형식을 유연하게 지원해야 한다. 이는 연산 특성에 맞는 데이터 정밀도를 적용함으로써 추론 정확도를 유지하면서 메모리 대역폭과 에너지 효율을 동시에 개선할 수 있도록 한다. 나아가 피지컬 AI 시스템에서는 단순한 평균 지연시간 감소를 넘어 입력부터 출력까지의 처리 시간을 예측 가능하게 보장하는 결정론적 실행이 중요하며, 이를 위한 실시간 연산 자원 관리 및 데이터 처리 구조가 요구된다.

둘째, 피지컬 AI 단말의 지속 가능한 임무 수행을 위해 전력 소모를 최소화하는 고효율 설계가 필수다. 특히, 드론이나 이동형 로봇 등은 전력 자원이 제한된 환경에서 운용되므로 전력효율(TOPS/W)을 극대화하는 아키텍처가 요구된다. 이를 위해 차세

대 저전력 메모리를 도입하여 데이터 전송에 따른 전력 소모를 최소화하면서도 메모리 대역폭을 확보하고, 동적 전력 관리 기술(예: Dynamic Voltage and Frequency Scaling)을 도메인별로 적용하여 연산 부하에 따라 전력효율을 동적으로 최적화해야 한다.

셋째, 카메라, 라이다(LiDAR), 레이다(Radar), 관성측정장치(IMU: Inertial Measurement Unit) 등 이종 센서 데이터를 실시간으로 융합 · 처리하는 다중 센서 처리 및 통신 능력이 필수이다. 이를 위해 ISP(Image Signal Processor) 및 비전 가속기, 다채널 카메라 인터페이스를 탑재해야 하며, 차량 및 로봇 제어를 위한 고속 · 실시간 네트워크 인터페이스를 지원해야 한다.

넷째, AI 오동작이나 사이버 공격이 물리적 사고로 이어지지 않도록 하드웨어 수준의 안전 메커니즘과 보안 기술이 필요하다. 기능 안전 측면에서는 메모리 및 데이터 전송 과정의 오류를 탐지 · 수정하는 ECC(Error Correcting Code), 연산 오류 및 결함을 감지하는 하드웨어 락스텝(Lockstep), 그리고 메인 연산 영역과 독립적으로 동작하는 세이프티 아일랜드(Safety Island) 기반 안전 메커니즘 등이 필요하다. 보안 측면에서는 신뢰 가능한 코드 실행을 보장하는 시큐어 부트(Secure Boot), 암호 키와 장치 인증 기반을 제공하는 루트 오브 트러스트(RoT: Root of Trust), 중요 데이터 및 연산 보호를 위한 신뢰 실행 환경(Trusted Execution Environment) 및 암호화 엔진 등 하드웨어 기반 보안 구조가 요구된다.

2. 소프트웨어 플랫폼

소프트웨어 플랫폼은 AI 모델의 최적화, 배포 및 유지보수 전 과정을 지원하며 하드웨어 자원의 활용성을 극대화하는 핵심 인프라이다. 특히 글로벌 표준 프레임워크 및 산업별 미들웨어와의 호환성은 개발 편의성을 높이고 다양한 애플리케이션과 서드파티 생태계 확장을 촉진한다. 이를 위해 소프트웨어 플랫폼은 다음과 같은 핵심 요소를 갖추어야 한다.

첫째, AI 모델 최적화와 개발 생산성을 극대화하기 위한 통합 SDK가 필요하다. 컴퓨팅 자원이 제한된 온디바이스 환경에서 대규모 AI 모델을 효율적으로 배치 · 실행하기 위해 양자화(Quantization), 프루닝(Pruning) 등 경량화 도구를 포함하는 AI 컴파일러와 추론 런타임이 요구된다. 또한, 디바이스 드라이버와 부트로더를 포함하는 BSP(Board Support Package), 프로파일러 및 디버거를 제공하여 개발 및 최적화 과정을 지원해야 한다.

둘째, 미션 크리티컬한 제어 환경에서 실시간성과 운영 연속성을 보장하는 운영체제와 실행 환경이 필수적이다. 피지컬 AI의 가혹한 구동 환경 내에서 신뢰성 높은 제어 성능을 달성하기 위해서는, 범용 응용 실행을 위한 고성능 운영체제(Linux 등)와 실시간 제어를 위한 실시간 운영체제를 단일 플랫폼에서 동시에 운용할 수 있는 하이퍼바이저 기반 이기종 실행 환경 관리와 미들웨어 최적화가 요구된다.

셋째, 플랫폼의 지속 가능한 확장을 위해 글로벌 AI 및 산업 표준 생태계와의 호환성 확보가 필요하다. PyTorch, TensorRT, ONNX 등 글로벌 표준 AI 프레임워크와의 연동은 물론, 로봇 공학 표준인 ROS 2(Robot Operating System), 자동차 소프트웨어 표준인 AUTOSAR, 컴퓨터 비전 라이브러리인 OpenCV 등 산업별로 구축된 개발 생태계와 호환되어야 플랫폼의 범용성과 시장 경쟁력을 유지할 수 있다.

III. 국외 기술 동향

이 장에서는 Ⅱ장에서 도출한 하드웨어 및 소프트웨어 요구사항을 기준으로, 미국의 Tesla, NVIDIA,

Qualcomm, 중국의 Rockchip, 이스라엘의 Hailo, 일본의 Renesas 등 주요 온디바이스 AI 반도체 플랫폼을 대상으로 AI 연산, 저전력 설계, 센서 인터페이스, 기능 안전 및 보안, 소프트웨어 생태계의 관점에서 비교 · 분석한다.

1. Tesla HW4(AI4)

자율주행 대표 플랫폼인 Tesla Hardware 4(HW4, 내부 코드명 AI4)는 Vision-Only 센서 구성과 End-to-End 신경망 아키텍처라는 Tesla의 자율주행 철학을 차량 내 컴퓨팅 하드웨어와 FSD(Full Self-Driving)라는 소프트웨어 스택과 긴밀하게 통합한 온디바이스 AI 이종 컴퓨팅 플랫폼이다. Tesla가 HW4의 공식 사양을 공개하지 않았으므로 이 절에서는 공개된 분석 자료를 바탕으로 정리하며 정량적 수치는 제3자 분석에 기반한 추정치임을 전제로 한다[5-7].

하드웨어 측면에서는 동일한 영상 입력을 두 개의 칩이 독립적으로 처리한 후 결과를 상호 검증하는 듀얼 SoC 구조를 채택하여 안전성을 높인 것으로 알려져 있다. 각 SoC는 Arm Cortex-A72 기반 CPU 코어와 다수의 커스텀 NPU, GDDR6 외부 메모리를 결합한 이종 연산 구조를 가지며, 전체 시스템 기준으로 수십~수백 TOPS(INT8) 수준의 추론 성능을 제공한다. 온칩 SRAM과 GDDR6 메모리 조합을 통해 데이터 이동 효율을 높이고, 12채널 고해상도(약 5 Mpixel급) 카메라 입력을 처리하는 비전 ISP를 통해 저지연 파이프라인을 구성하였다. 각 SoC는 전원 서브시스템과 메모리가 독립적으로 구성되어 특정 노드의 고장이나 과열 발생 시에도 전체 시스템의 안전 제어권을 유지할 수 있도록 물리적 이중화를 구현하였다. 또한, 암호학적으로 서명된 코드만 실행하도록 하는 보안 부트 체계와 두 FSD 컴퓨터 간 주행 계획 일치 여부를 판단하는 락스텝 기반 안전 판정 회로를 적용한 것으로 추정된다.

소프트웨어 측면에서는 PyTorch로 학습된 신경망을 전용 AI 컴파일러 및 추론 런타임을 통해 HW4의 NPU 구조에 맞게 최적화 · 매핑하는 소프트웨어 스택을 운용한다. 이 컴파일러는 LLVM 산하 MLIR 기반으로 진화하고 있는 것으로 알려져 있으며, 최신 FSD 버전에서 반응 시간과 모델 업데이트 효율을 개선하는 데 기여한 것으로 보고된다. 실행 환경에서는 차량 외부 카메라로부터 입력된 원시 비디오 스트림을 기반으로 조향 · 가감속 등 차량 제어 명령을 직접 산출하는 End-to-End 신경망 모델을 구동하며, 이는 FSD v12 이후 도입된 아키텍처이다. 아울러 섀도 모드(Shadow Mode)를 통해 실제 주행 중 수집된 데이터 및 예외 상황 클립을 자동 선별하여 클라우드 학습 인프라로 피드백하는 데이터 수집 · 학습 파이프라인을 구축하고 있는 것으로 알려져 있다.

요약하면, Tesla HW4는 듀얼 SoC 기반 이종 컴퓨팅 구조, 고해상도 비전 파이프라인, 보안 부트 및 락스텝 기반 안전 판정 회로, PyTorch-전용 컴파일러-End-to-End FSD 모델로 이어지는 하드웨어 · 소프트웨어 통합 스택을 통해 Ⅱ장에서 제시한 피지컬 AI용 온디바이스 AI 반도체 플랫폼 요구사항을 종합적으로 충족하는 대표적인 사례로 볼 수 있다.

2. NVIDIA Jetson AGX Orin

Jetson AGX Orin은 자율주행, 로봇 및 다양한 엣지 AI 응용을 위한 대표적인 범용 온디바이스 AI 컴퓨팅 플랫폼이다[8].

하드웨어 플랫폼 관점에서는 차량용 및 고성능 임베디드 프로세서인 Arm Cortex-A78AE CPU와 Ampere 아키텍처 기반 GPU, 그리고 심층학습 전용 가속기인 DLA(Deep Learning Accelerator)와 컴퓨터

비전 가속기인 PVA(Programmable Vision Accelerator)가 유기적으로 결합한 하이엔드 이종 컴퓨팅 구조를 갖추고 있다. 이를 통해 최대 275 TOPS(INT8)의 AI 추론 성능을 제공하며, 메인 연산 외의 심층학습 및 비전 처리를 저전력 도메인인 DLA와 PVA로 분산 배치함으로써 전력효율을 극대화한다. 구체적으로는 Ampere GPU 및 DLA를 통해 EO/IR 영상 기반의 실시간 AI 추론 및 3D 맵핑을 수행하는 한편, PVA를 활용한 필터링 및 특장점 처리 및 VIC를 통한 렌즈 왜곡 보정 등의 컴퓨터 비전 알고리즘을 결정론적(Deterministic)으로 처리한다. 또한, 대용량 온칩 캐시와 고대역폭 LPDDR5 메모리 구조를 탑재하여 데이터 이동 시 발생하는 연산 병목을 최소화하며, 루트 오브 트러스트(RoT), AES-256 기반 암호화 가속기, 메모리 및 저장장치 차원의 보안 아키텍처를 통합함으로써 가혹한 자율 구동 환경에서 요구되는 높은 기능 안전성과 데이터 무결성을 보장한다. 이와 함께 다채널 MIPI CSI-2 인터페이스를 통한 고해상도 비전 센서 입력은 물론 라이다, 레이다, 관성측정장치 등 이종 센서 데이터를 융합 및 처리하는 멀티센서 파이프라인을 구축하였다. BSP 기반 시큐어 부트, 신뢰 실행 환경, 디스크 및 메모리 암호화 등 보안 기능을 지원하며, 특히 미션 크리티컬 환경의 신뢰성을 보장하기 위해 산업용 라인업의 경우 메모리 전반에 데이터 비트 오류를 보정하는 ECC 기술을 적용하였다.

소프트웨어 플랫폼 측면에서 Jetson AGX Orin은 Jetson Linux 기반 BSP와 NVIDIA 고유 병렬 컴퓨팅 아키텍처인 CUDA를 축으로 하는 JetPack SDK를 통해 하드웨어의 연산 잠재력을 극대화하고 있다. 특히 핵심 텐서 컴파일러인 TensorRT를 통해 모델 경량화 및 신경망 레이어 융합을 수행함으로써, CPU, GPU, DLA, PVA 등 이기종 하드웨어 자원을 최적으로 활용하게 하여 실질적인 연산 효율성을 높인다. 또한, PyTorch 및 ONNX 등 글로벌 표준 AI 프레임워크와의 유기적인 연동을 지원하고, 최적화된 하드웨어 가속 미들웨어 환경을 제공함으로써 Jetson 기반으로 로봇 및 엣지 AI 애플리케이션 생태계를 지속적으로 확대하고 있다.

요약하면, Jetson AGX Orin은 CPU, GPU, 딥러닝 · 비전 가속기를 통합한 이종 컴퓨팅 아키텍처, 멀티센서 입력 · 통신 파이프라인, ECC 및 시큐어 부트 · TEE 기반 기능 안전 · 보안 메커니즘, CUDA-TensorRT-JetPack SDK-표준 프레임워크 연계로 하드웨어 · 소프트웨어 요구사항을 균형 있게 충족하는 대표적인 범용 엣지 AI 플랫폼으로 평가할 수 있다.

3. Qualcomm Dragonwing IQ10

로봇 대표 플랫폼인 Dragonwing IQ10은 휴머노이드 로봇, 자율 이동 로봇(AMR: Autonomous Mobile Robot), 협동 로봇(Cobot) 및 첨단 공장 자동화 시스템을 목표로 하는 퀄컴의 프리미엄 로보틱스 플랫폼이다[10].

하드웨어 플랫폼 관점에서 Dragonwing IQ10은 퀄컴 오라이온(Oryon) CPU, 아드레노(Adreno) GPU, 벡터 및 텐서 연산에 특화된 헥사곤(Hexagon) NPU가 통합된 이종 컴퓨팅 구조를 바탕으로 최대 약 700 TOPS(INT8, sparse 기준)/350 TOPS(INT8, dense 기준) 수준의 AI 연산 성능을 제공한다. NPU 도메인은 객체 인식뿐만 아니라 비전-언어 모델을 포함한 생성형 AI 워크로드를 저전력 · 저지연으로 처리하도록 설계되어 있으며, 배터리 기반 엣지 환경에 최적화된 고효율 오라이온 CPU와 하드웨어 수준의 스파시티(Sparsity) 연산을 지원하는 헥사곤 NPU를 주축으로 워크로드를 분산 배치함으로써 전력효율을 극대화한다. 센서 측면에서는 최대 12채널의 고

해상도 카메라 입력과 라이다, 레이다, 관성측정장치, Time-of-Flight(ToF) 기반 3차원 공간 인식 등 이종 센서 데이터를 실시간으로 융합 처리하는 고대역폭 센서 데이터 처리 구조를 갖추고 있다. 또한, 메모리 전체에 대한 ECC 적용, 독립적인 세이프티 아일랜드 기반 안전 메커니즘, 고속 데이터 전송용 PCIe Gen5, Ethernet/TSN 및 EtherCAT 인터페이스 통합 등을 통해 산업 및 자동차 기능 안전 요구를 지원하면서 로봇 제어용 실시간 고속 네트워크를 제공한다.

소프트웨어 플랫폼 측면에서, Dragonwing IQ10을 위해 AI 모델 경량화, 양자화 및 컴파일을 통합적으로 지원하는 Qualcomm AI Stack을 제공한다. AI Hub 개발 플랫폼에서 PyTorch, ONNX 등 글로벌 표준 프레임워크로 개발된 모델을 AI Engine Direct 런타임이 넘겨받아 CPU, GPU, NPU 등 이종 컴퓨팅 자원에 최적으로 배치 · 실행함으로써 개발 기간을 단축하고 칩의 연산 효율을 극대화한다. 생태계 측면에서, 본 플랫폼은 개방형 협력 아키텍처를 지향하며 ROS 2 및 Linux BSP 기반 Qualcomm Robotics SDK를 서드파티 개발자 환경에 전면 공개하여 글로벌 로봇 소프트웨어와의 호환성을 강화하고 있다. 나아가 다양한 센서 인터페이스용 BSP · 드라이버와 사전 검증된 로보틱스 레퍼런스 디자인을 포함한 통합 풀스택을 공개하였다. 이러한 기술 공개는 파트너사들과의 상생적 생태계 구축을 유도하며, 개별 부품 통합의 기술적 장벽을 제거함으로써 서드파티 개발사들이 엣지 AI 로봇 시스템의 초기 프로토타이핑부터 최종 상용화 및 양산에 이르는 전 주기를 획기적으로 단축할 수 있게 한다.

요약하면, Dragonwing IQ10은 CPU, GPU, NPU 기반 이종 컴퓨팅과 고신뢰성 및 결정론적 실시간성을 보장하기 위한 TSN/EtherCAT 네트워크, 세이프티 아일랜드 안전 구조 및 개방형 SDK 생태계를 유기적으로 결합함으로써 피지컬 로봇 도메인의 필수적 온디바이스 AI 요구사항을 충족하는 로보틱스 특화 플랫폼이다.

4. Rockchip RK3588

저가형 범용 플랫폼인 Rockchip RK3588은 서비스 로봇, AIoT, 스마트 비전, 산업용 엣지 컴퓨팅 등을 목표로 하는 AI SoC이다[11]. 오픈소스 기반의 방대한 보드 생태계와 저비용 대비 준수한 컴퓨팅 성능을 바탕으로 상용 엣지 AI 시장에서 활용 범위를 넓혀가고 있으며, 저가형 임베디드 AI 시스템 및 프로토타입 개발에 널리 사용되고 있다.

하드웨어 플랫폼 관점에서 RK3588은 옥타코어 CPU(Cortex-A76/A55)와 Mali-G610 GPU, 최대 6 TOPS(INT8) 성능을 제공하는 3코어 NPU를 단일 다이에 통합한 범용 이종 아키텍처를 채택한다. 각 NPU 코어는 독립적으로 작동하거나 공동 가속이 가능하도록 설계되어, YOLO 등 경량 컴퓨터 비전 모델의 추론 가속에 널리 활용된다. 이와 함께 고성능 ISP와 대용량 영상 처리를 하드웨어로 가속하는 8K VPU(Video Processing Unit)를 통합하여 다채널 카메라 영상의 실시간 전처리와 AI 추론을 지원하며, PCIe 3.0, 기가비트 이더넷, CAN/CAN-FD 인터페이스 등 다양한 주변장치 및 네트워크 인터페이스를 탑재하여 센서 연동성과 시스템 확장성을 확보하고 있다.

소프트웨어 플랫폼 측면에서는 임베디드 리눅스 및 안드로이드 BSP와 함께 전용 NPU 가속 프레임워크인 RKNN-Toolkit2를 지원한다. 본 프레임워크는 PyTorch, TensorFlow, ONNX 기반 모델을 효과적으로 양자화하고 RKNN 실행 포맷으로 변환함으로써, 저가형 엣지 환경에서도 AI 모델이 최적의 효율로 손쉽게 배치 · 구동되도록 돕는다. 아울러 생태

계 호환성 관점에서는 Ubuntu 기반 ROS 2 개발 환경 및 다양한 싱글보드 컴퓨터 하드웨어를 수용하여, ROS 2 기반 로봇 소프트웨어와 연계된 임베디드 AI 시스템의 프로토타이핑과 검증 속도를 획기적으로 향상시킨다.

요약하면, Rockchip RK3588은 CPU, GPU, 경량 NPU 기반 이종 컴퓨팅 구조와 ISP, 8K VPU 기반 비전 처리 능력, RKNN-Toolkit2 및 ROS 2 생태계를 결합하여 저비용 · 범용 엣지 AI 영역의 필수 요구사항을 충족하는 실용적인 플랫폼이다.

5. Hailo-15H/15L

비전 특화 플랫폼인 Hailo-15H/15L은 스마트 카메라, 산업용 비전, 로보틱스 등 비전 응용을 목표로 하는 고효율 온디바이스 AI SoC 제품군이다[12,13].

하드웨어 플랫폼 관점에서 Hailo-15는 쿼드코어 Arm Cortex-A53 CPU와 독자적인 데이터플로우(Dataflow) 아키텍처 기반의 NPU(최상위 Hailo-15H 모델 기준 최대 20 TOPS), 벡터 비전 DSP 및 하드웨어 ISP를 단일 칩에 통합한 이종 구조를 채택하고 있다. 본 플랫폼은 소형 드론이나 임베디드 카메라 등 열 · 전력 제약이 엄격한 엣지 환경에 최적화되어 저전력 한도 내에서 높은 전력효율을 달성하는 장점이 있으며, 이를 통해 실시간 비전 프로세싱과 심층학습 추론을 동시에 지원한다. 세부적으로는 12MP급 카메라 입력 사양과 저조도 환경을 지원하는 고성능 AI ISP를 내장하여 HDR, 노이즈 제거, 렌즈 왜곡 보정 등의 고부하 비전 전처리를 하드웨어 도메인에서 실시간으로 수행한다. 아울러 PCIe Gen3, 기가비트 이더넷, USB 3.1 및 다중 MIPI CSI-2 인터페이스를 통해 센서 연동성과 시스템 확장성을 확보하는 한편, 시큐어 부트, Arm TrustZone, 난수생성기(TRNG: True Random Number Generator) 기반의 하드웨어 보안 기능 등을 제공하여 코드와 데이터 보호를 지원한다.

소프트웨어 플랫폼 측면에서는 칩 내 하드웨어 리소스 효율을 극대화하기 위해 전용 Hailo 데이터플로우 컴파일러와 런타임 엔진 HailoRT를 제공한다. PyTorch, ONNX 등으로 학습된 AI 모델을 칩 내부 데이터플로우 기반 하드웨어 토폴로지에 맞게 매핑하고, 고정소수점 양자화를 수행함으로써 컴파일 단계에서부터 연산 오버헤드를 줄이고 실질적인 하드웨어 활용 효율을 높인다. 생태계 호환성 측면에서는 Yocto 리눅스 BSP를 제공하고, OpenCV와 멀티미디어 프레임워크 GStreamer 파이프라인과 직접 연동함으로써 센서 입력, 비전 전처리, AI 추론, 비디오 인코딩에 이르는 전체 파이프라인을 소프트웨어 수준에서 효율적으로 구성할 수 있다.

요약하면, Hailo-15는 데이터플로우 NPU, ISP, 비전 DSP 기반의 저전력 아키텍처와 다중 인터페이스의 확장성, 하드웨어 보안, 전용 컴파일러 중심의 소프트웨어 생태계를 결합하여 실시간 비전 및 저전력 엣지 환경에 최적화된 특화 플랫폼이다.

6. Renesas RZ/V2H

산업용 플랫폼인 Renesas RZ/V2H는 산업용 로봇, 머신비전, 스마트 카메라 등을 목표로 하는 고성능 · 저전력 온디바이스 AI SoC이다[14-16].

하드웨어 플랫폼 관점에서는 응용 구동용 쿼드코어 Arm Cortex-A55, 실시간 제어용 Cortex-R8, 시스템 관리용 Cortex-M33, 그래픽 처리용 Mali-G31 GPU와 함께 독자적인 동적 재구성 가속기인 DRP(Dynamic Reconfigurable Processor)-AI3를 단일 다이에 통합한 하이브리드 이종 아키텍처를 채택하고 있다. 메인 AI 연산을 담당하는 DRP-AI3는 알고리즘 특성에 맞춰 내부 연산 경로를 동적으로 재구성하

고 데이터플로우 방식으로 처리함으로써 메모리 접근 오버헤드를 최소화하고 전력 대비 연산 성능을 극대화한다. 이를 통해 열 제약이 엄격한 엣지 환경에서도 고속 비전 AI 추론과 결정론적 실시간 제어 성능을 동시에 보장한다. 나아가 내장 하드웨어 ISP와 다채널 MIPI CSI-2 입력을 통해 고해상도 영상 전처리를 실시간으로 수행하며, 외부 시스템과의 고속 · 고신뢰성 통신을 위해 PCIe Gen3, TSN 기반 기가비트 이더넷, CAN-FD 인터페이스를 탑재하여 정밀 로봇 제어 및 산업용 실시간 네트워킹 요구사항을 충족한다.

소프트웨어 아키텍처 측면에서 Renesas는 RZ/V2H를 위해 DRP-AI Translator 기반 AI 컴파일 환경을 제공하여 PyTorch, ONNX 등으로 학습된 모델을 DRP-AI3 실행 환경에 맞게 최적화한다. OS 레이어에서는 대규모 연산용 리눅스 BSP와 미션 크리티컬 제어용 실시간 운영체제 환경을 동시에 지원하는 하이브리드 소프트웨어 구조를 구현하여, 비전 AI 워크로드와 실시간 제어 워크로드를 단일 플랫폼에서 분리하거나 병행하여 운용할 수 있도록 한다. 생태계 호환성 측면에서는 GStreamer 및 OpenCV 인터페이스와 밀접하게 연동되어 영상 캡처-전처리-AI 추론 파이프라인을 효율적으로 구성할 수 있고, ROS 2 기반 로봇 소프트웨어 및 ONNX 기반 AI 모델 개발 환경과의 연계를 통해 산업용 로봇 및 비전 응용을 위한 개발 생태계를 지원한다.

요약하면, RZ/V2H는 DRP-AI3 가속기와 Cortex-A/R/M CPU, TSN/CAN-FD 인터페이스 기반의 하드웨어 구조, 리눅스 · 실시간 운영체제 하이브리드 OS 및 전용 소프트웨어 스택을 결합함으로써 고속 비전 AI와 결정론적 제어를 동시에 요구하는 산업용 로봇 · 비전 도메인에 적합한 플랫폼이라 할 수 있다.

7. 기타 글로벌 플랫폼 및 진영 동향

앞서 고찰한 주요 플랫폼 외에도 글로벌 온디바이스 AI 반도체 시장에서는 범용 프로세서 생태계를 기반으로 다각화된 플랫폼을 전개하는 전통 프로세서 기업과, 컴퓨터 비전 등 특정 응용 분야에 집중하는 특화 기업이 서로 다른 시장 접근 전략으로 경쟁력을 확보하고 있다.

전통 칩 메이커인 AMD는 자일링스(Xilinx) 인수를 통해 확보한 FPGA 프로그래머블 로직과 AI 연산 코어를 결합한 어댑티브 SoC 아키텍처를 구축하고 있다[17]. 본 구조는 하드웨어 수준의 재구성 기능을 바탕으로 알고리즘 변화에 유연하게 대응하면서도 예측 가능한 저지연 처리를 보장하는 것이 특징이다. 특히 우주용 방사선 내성(Radiation-Tolerant) 기술이 적용된 Versal 어댑티브 SoC 제품군은 위성 및 우주 임무 등 고신뢰성 임베디드 영역에서 실증 사례를 확보하였으며[25], 이를 기반으로 항공우주와 국방 등 고신뢰성 시장에 전략적으로 집중하고 있다[26].

인텔은 x86 기반 CPU, GPU, NPU의 이종 컴퓨팅 구조와 OpenVINO 소프트웨어 생태계를 결합하여 다양한 AI 워크로드를 지원하는 범용 AI 플랫폼 전략을 전개하고 있다[18]. 이러한 전략은 기존 x86 개발 환경과의 높은 호환성을 바탕으로 AI 모델의 최적화 및 배포 과정을 간소화하고, 이미 보급된 PC 및 산업용 엣지 컴퓨팅 자산을 활용한 AI 기능 확장을 가능하게 함으로써 범용 엣지 시장에서의 기술적 진입 장벽을 낮추고 기존 생태계와의 연속성을 유지하게 한다.

한편, 암바렐라 CV 시리즈는 고성능 ISP와 독자 CVflow AI 가속 아키텍처를 통합한 비전 특화 온디바이스 AI SoC이다[19,20]. 저전력 하드웨어 설계로 제한된 에너지 환경에서도 고해상도 다채널 카

메라와 레이더 등 이종 센서 데이터의 실시간 동기화 및 융합 처리가 가능하다. 특히 최신 자율주행용 CV3-AD 라인업은 ISO 26262 기능 안전 표준을 충족하기 위해 고부하 AI 연산을 담당하는 주 연산 영역은 ASIL-B 등급으로, 안전 감시 및 핵심 제어 기능을 담당하는 세이프티 아일랜드는 최고 등급인 ASIL-D로 설계하고, 두 영역을 하드웨어적으로 분리하였다. 이를 통해 엣지 환경에서 요구되는 저전력 가속 성능과 시스템의 기능적 신뢰성을 동시에 달성하고자 하였다.

IV. 국내 기술 동향

국내 온디바이스 AI 반도체 산업은 자율주행 로봇, 무인기(UAV: Unmanned Aerial Vehicle) 등 피지컬 AI 응용을 목표로 하는 저전력 · 고효율 AI SoC와 전용 소프트웨어 스택 개발을 중심으로 전개되고 있다. 최근에는 NPU 코어의 단순 탑재를 넘어 온칩 ISP, 하드웨어 보안 모듈, 그리고 미들웨어까지 단일 다이에 통합한 플랫폼 형태로 고도화되는 추세이다.

국내 주요 팹리스 기업들은 타깃 응용 도메인에 따라 차별화된 아키텍처 전략을 취하고 있다. 모빌린트(Mobilint)는 자율주행 로봇 및 드론 시장을 겨냥한 저전력 엣지용 REGULUS를 개발하고, 자체 컴파일러 스택인 SDK qb를 통해 이종 연산 자원에 대한 그래프 최적화와 효율적인 모델 배포를 지원한다[21]. 넥스트칩(Nextchip)은 드론, 차량용 및 로봇 비전 분야에 특화된 AI SoC인 APACHE6 기반의 센서 융합 플랫폼을 제공한다. 고성능 온칩 ISP 기술력과 NPU를 결합하여 다채널 비전 데이터를 하드웨어 레벨에서 실시간으로 가속함으로써 악천후 및 야간 환경에서의 객체 인식 신뢰성을 향상시키고 있으며, 최근에는 차량용을 넘어 열화상 및 다중 센서 융합 기술을 기반으로 군용 드론 등 방산 시장으로의 영역 확장을 도모하고 있다[22]. 디퍼아이(Deeper-I)는 고신뢰성 엣지 보안 및 스마트 시티 시장을 겨냥해 하드웨어 기반 보안 기능과 자체 NPU를 통합한 Tachy 시리즈를 개발하고 있으며, 하드웨어 수준의 암호화 및 무결성을 바탕으로 공공 안전 분야의 엣지 보안 시장을 공략하고 있다[23]. 딥엑스(DEEPX)는 초저전력 드론 및 카메라향 DX-M 시리즈부터 다채널 영상 분석 및 로봇 제어에 최적화된 고성능 DX-H 시리즈에 이르는 온디바이스 AI 제품군을 보유한다. 또한, 전용 툴체인인 DXNN SDK를 통해 다양한 엣지 컴퓨팅 자원에서의 추론 최적화를 지원한다[24].

V. 비교 분석

피지컬 AI의 확산과 함께 온디바이스 AI 반도체는 단순한 AI 연산 가속기를 넘어 센서 인식, AI 추론, 제어 및 보안 기능을 통합적으로 수행하는 플랫폼 중심으로 발전하고 있다. 본고에서 분석한 글로벌 주요 기업들은 표 1에서 보듯이 CPU, GPU, NPU를 결합한 이종 컴퓨팅 아키텍처를 기반으로 고성능 AI 연산을 구현하는 동시에, 다중 센서 인터페이스, 기능 안전, 하드웨어 보안, 실시간 운영체제, AI 컴파일러 및 SDK를 통합한 하드웨어 · 소프트웨어 플랫폼을 구축하여 경쟁력을 확보하고 있다. 이는 피지컬 AI 시대의 경쟁력이 개별 NPU의 연산 성능보다 시스템 수준의 플랫폼 통합 역량에 의해 결정되고 있음을 시사한다.

한편, 국내 AI 반도체 기술은 저전력 AI 연산과 비전 처리 중심의 NPU 설계 분야에서는 경쟁력을 확보하고 있으나, 공개자료 기준으로, 국내 제품의 기능 안전과 하드웨어 보안, 다양한 센서 및 제어 인터페이스, 플랫폼 소프트웨어와 개발 생태계 측면에서

표 1 피지컬 AI를 위한 주요 온디바이스 AI 반도체 플랫폼 비교

제조사/제품 (국가)	CPU	GPU	NPU (성능)	메모리 (대역폭)	전력 소모	비전·센서 처리	기능 안전 대응	보안	SDK	생태계 연계	주력 응용/ 특징
Tesla/ HW4(AI4) (미국) [4-7]	20-core Cortex-A72	-	3-core 커스텀 NPU (50~240 TOPS)	16GB GDDR6 (224GB/s)	150~200W (보드(2개 칩) 기준)	최대 12개 카메라 (총 48 레인), ISP	Dual SoC, Safety Monitor	Secure Boot(HW3)	Tesla Compiler, FSD Runtime	미확인	실시간 대용량 데이터 처리 자율주행 플랫폼
NVIDIA/ Jetson AGX Orin (미국) [8]	8~12-core Cortex-A78 AE	Ampere GPU (2048-CUDA core +64 Tensor core)	DLA (GPU+DLA 합산 성능: ~275 TOPS)	32~64GB LPDDR5 (204.8GB/s)	15~60W	최대 6개 카메라 (총 16 레인) PVA, MIPI CSI-2Sensor Fusion	ECC 메모리(산업용), Lockstep, HSM 등	Secure Boot, OP-TEE, 디스크 암호화, 롤백방지 등	CUDA, TensorRT, JetPack SDK, Isaac	PyTorch, ONNX, ROS 2 등	GPU 기반 고성능 범용 AI 플랫폼
NVIDIA/ Jetson Orin NX (미국) [9]	Cortex-A78 AE	Ampere GPU	NVDLA-based NPU (GPU+NPU 합산 성능: ~100 TOPS)	16GB/8GB LPDDR5	10~40W	최대 4개 카메라 (총 8 레인), ISP, MIPI CSI-2	ECC	ARM TrustZone			GPU 기반 소형 엣지 AI 플랫폼
Qualcomm/ Dragonwing IQ10 (미국) [10]	18-core Oryon	Adreno	Hexagon (~700 TOPS)	64 GB LPDDR5x	미확인	카메라, LiDAR, Radar, ToF, IMU 등 멀티센서 융합	SIL-3 Safety island 대응	ARM TrustZone	Qualcomm AI Stack, Robotics SDK	PyTorch, ONNX, TensorFlow, ROS 2 등	로보틱스 AI 플랫폼
Rockchip/ RK3588 (중국) [11]	4-core Cortex-A76 + 4-core Cortex-A55 Big.LITTLE	4-core Arm Mali-G610	3-core NPU (~6 TOPS)	~32GB LPDDR4/ LPDDR4X, LPDDR5	12~15W	최대 6개 카메라 (총 16레인), 8K ISP	미확인	Embedded 2 cipher engine	RKNN Toolkit2	PyTorch, TensorFlow, ONNX, ROS 2 등	임베디드, AIoT, 엣지 AI 플랫폼
Hailo/Hailo-15L/15H (이스라엘) [12,13]	4-core Cortex-A53	-	NPU (~20 TOPS)	LPDDR4 / LPDDR4X	미확인	AI ISP(12 MP), 듀얼 MIPI CSI-2, ISP, Vision DSP	미확인	Secure Boot, TrustZone, TRNG, crypto lib. 등	Hailo Dataflow Compiler, HailoRT	PyTorch, ONNX, OpenCL, GStreamer, ROS 2 등	드론, 스마트 카메라, 서비스 로봇/ 비전+ISP 중심
Renesas/ RZ/V2H (일본) [14-16]	4-core Cortex-A55 +2-core Cortex-R8 +1-core Cortex-M33	Arm Mali-G31	DRP-AI3 (8 TOPS/ dense, 80 TOPS/ sparse)	~16GB LPDDR4	~20W	MIPI CSI-2 ISP, WDR, CAN-FD 등	미확인	Secure Boot, crypto, TRNG 등	DRP-AI SDK	PyTorch, ONNX, OpenCL, GStreamer, ROS 2 등	산업용 비전 AI 플랫폼
Ambarella/ CV3-AD (미국) [19,20]	8~12 Cortex-A78 AE+ Cortex-R52 (lockstep)	automotive GPU	NVP(NN)+ GVP(vector) (미확인)	~64GB LPDDR5X / LPDDR5	미확인	최대 3개 카메라, MIPI CSI-2 ISP, CAN-FD 등	dual-core Lockstep, ASIL-B(chip) ASIL-D safety island	HSM, Secure Boot, TRNG, OTP 등	CVflow SDK	PyTorch, TensorFlow, Caffe, ONNX 등	자율주행/ 비전특화
국내 기술 수준 종합 [21-24]	Cortex A 시리즈 등	Arm Mali 등	자체개발 NPU (12~160 TOPS)	2~64GB LPDDR4 / LPDDR4X/ LPDDR5	수십 W 이내	MIPI CSI-2, PCIe G3~ PCIe G5 등	ASIL-B/D목표(넥스트칩 APACHE6) 외 미확인	하드웨어 기반 보안(넥스트칩 APACHE6) 외 미확인	제품별 SDK	PyTorch, TensorFlow, ONNX 등	로봇, 스마트 팩토리, 국방 무인체계, 드론 등

* 미확인: 공식 데이터시트, 제품 설명서, 보도자료 등 공개된 자료에서 해당 정보를 확인하지 못한 항목

** 표에서 Cortex-*는 Arm Cortex-*를 의미함

는 글로벌 선도 기업의 사례를 참고한 지속적인 고도화가 필요한 것으로 분석된다. 특히 AI 모델 최적화 도구, 컴파일러, SDK, 미들웨어 등 소프트웨어 플랫폼은 사용자 접근성과 생태계 확장성을 좌우하는 핵심 요소임에도 불구하고, 국내에서는 하드웨어 중심의 개발이 상대적으로 활발하여 플랫폼 완성도 측면의 경쟁력 확보가 중요한 과제로 남아 있다.

이러한 기술 격차를 해소하기 위해서는 개별 NPU 성능 향상을 넘어 가혹 환경 신뢰성, 하드웨어 보안, 운영체제/미들웨어를 포괄하는 고신뢰성 통합 플랫폼 기술 확보가 핵심 과제다. 다만, 응용 분야별 요구사항과 비용 구조가 상이해 단일 플랫폼으로 모든 도메인에 대응하기는 어려우므로 타깃 도메인별 차별화 전략이 요구된다. 이러한 도메인 차별화 전략의 일례로, 범용 및 로보틱스 시장(민수)에서는 글로벌 표준 프레임워크 호환성과 칩렛 기반 모듈형 설계를 통해 비용 효율성과 확장성을 확보하여 대량 양산 경쟁력을 극대화하는 전략이 유효하다. 반면, 항공우주 및 국방 등 특수 임무 시장(군수)에서는 기술 자립을 바탕으로 통신 단절이나 전자전 환경에서도 안정적인 자율 항법과 표적 탐지가 가능한 고신뢰성 독자 플랫폼 확보에 역량을 집중해야 한다. 이를 위해 검증된 핵심 IP를 상용 및 특수 임무 도메인에 상호 교차 적용(Dual-Use)하여 플랫폼의 신뢰성과 경제성을 동시에 확보하고, 이를 국산 소프트웨어 툴체인 및 차세대 피지컬 AI 응용 기술과 연계함으로써 글로벌 공급망 변화에 대응 가능한 자생적 온디바이스 AI 플랫폼 생태계를 구축할 필요가 있다.

정리하면, 앞서 도출한 기능 안전 및 물리적 보안 요건은 고신뢰성 통합 플랫폼 구축으로 해소하고, 타깃 도메인별 상이한 요구 구조는 범용 · 특수 도메인 겸용 기술을 기반으로 맞춤형 플랫폼을 다각화함으로써 대응하며, 외부 의존적 생태계 문제는 국산 소프트웨어 생태계 조성을 통해 단계적으로 극복해야 할 것이다.

VI. 결론 및 전망

본고에서는 피지컬 AI 구현을 위한 온디바이스 AI 반도체 플랫폼의 핵심 요구사항을 하드웨어와 소프트웨어 측면에서 정립하고, 이를 기준으로 글로벌 주요 기업과 국내 기술 동향을 비교 · 분석하였다. 분석 결과, 피지컬 AI 반도체의 경쟁력은 단일 가속기의 연산 성능만으로 결정되지 않으며, AI 연산 성능, 다중 센서 처리, 기능 안전, 하드웨어 보안, 실시간 소프트웨어 및 개발 생태계를 유기적으로 통합한 플랫폼의 완성도에 의해 좌우됨을 확인하였다. 특히 글로벌 선도 기업들은 CPU, GPU, NPU 기반의 이종 컴퓨팅 구조와 AI 컴파일러, SDK, 개발 프레임워크를 긴밀하게 결합함으로써 생태계 중심의 경쟁력을 확보하고 있으며, 관련 산업 역시 개별 칩 중심에서 플랫폼 생태계 경쟁으로 빠르게 전환되는 추세이다.

국내 AI 반도체 기술은 저전력 AI 연산과 NPU 설계 분야에서 지속적인 기술 발전을 이루고 있으나 기능 안전, 하드웨어 보안, 소프트웨어 플랫폼 및 글로벌 생태계 측면에서는 선도 기업과의 격차를 줄이기 위한 집중적인 연구개발이 요구된다. 향후 피지컬 AI의 적용 범위가 로봇, 드론, 자율주행차를 비롯한 미션 크리티컬(Mission-Critical) 영역으로 전방위적으로 확대될 것으로 예상됨에 따라, 온디바이스 AI 반도체는 멀티모달 센서 융합과 온디바이스 학습을 통합 지원하는 고신뢰성 플랫폼으로 진화할 것으로 전망된다. 이에 대응하기 위해 국내 산업은 단일 규격의 성능 경쟁에서 벗어나 공통 핵심 IP 기반의 '범용 · 특수 도메인 겸용 기술'을 확산하고 이를 시장별 요구사항에 맞추어 맞춤형 플랫폼으로

다각화하는 전략적 전환이 필요하다.

결과적으로 국내 온디바이스 AI 산업의 지속 가능한 성장은 하드웨어 · 소프트웨어 공동설계(Co-Design) 역량 강화와 개방형 소프트웨어 툴체인의 고도화에 의해 결정될 것이다. 개발자, 응용 서비스, 하드웨어가 유기적으로 선순환하는 개방형 산업 생태계를 구축함으로써 기술적 외부 의존성을 단계적으로 극복하고 글로벌 엣지 AI 시장에서 독자적인 기술 자립과 선도적 지위를 확고히 다져야 할 것이다.

용어해설

피지컬 AI 실세계와 상호작용을 하며 자율적으로 인지·판단·제어하는 AI 기술

온디바이스 AI 중앙 서버나 클라우드에 의존하지 않고 디바이스 내부에서 AI 추론을 수행하는 기술

IMU(Inertial Measurement Unit, 관성측정장치) 3축 가속도계와 3축 자이로스코프로 구성되며, 드론·로봇·자율주행차의 자세, 방향 및 운동 상태를 실시간으로 측정하는 핵심 센서

참고문헌

[1] NVIDIA, “What is Physical AI?,” https://www.nvidia.com/en-us/glossary/generative-physical-ai/

[2] S. Liu et al., “Edge Computing for Autonomous Driving: Opportunities and Challenges,” Proc. IEEE, vol. 107, no. 8, 2019, pp. 1697-1716.

[3] W. Shi et al., “Edge Computing: Vision and Challenges,” IEEE Internet Things J., vol. 3, no. 5, 2016, pp. 637-646.

[4] J. Mendez et al., “Edge Intelligence: Concepts, Architectures, Applications, and Future Directions,” ACM Trans. Embed. Comput. Syst., vol. 21, no. 5, 2022, pp. 1-41.

[5] AutoPilotReview, “Tesla Hardware 4(AI4) - Full Details and Latest News,” https://www.autopilotreview.com/tesla-hardware-4-rolling-out-to-new-vehicles/

[6] D. Patel and A. Kostovic, “Tesla AI Capacity Expansion - H100, Dojo D1, D2, HW 4.0, X.AI, Cloud Service Provider,” SemiAnalysis, 2023. 6. 28. https://newsletter.semianalysis.com/p/tesla-ai-capacity-expansion-h100

[7] I. Ali, “All that we know about Tesla FSD Hardware 4(HW4) computer, HW3 vs. HW4 comparison,” Tesla Oracle, 2023. 2. 25. https://www.teslaoracle.com/2023/02/25/tesla-autopilot-hw4-is-coming-and-musk-to-announce-big-things-on-investor-day/

[8] L.S. Karumbunathan, “NVIDIA Jetson AGX Orin Series: A Giant Leap Forward for Robotics and Edge AI Applications,” NVIDIA Technical Brief v1.2, 2022. 7. https://www.nvidia.com/content/dam/en-zz/Solutions/gtcf21/jetson-orin/nvidia-jetson-agx-orin-technical-brief.pdf

[9] NVIDIA, “NVIDIA Jetson Orin NX Series Data Sheet,” NVIDIA, 2026. 2. https://developer.download.nvidia.com/assets/embedded/secure/jetson/orin_nx/docs/Jetson_Orin_NX_DS-10712-001_v0.5.pdf?t=eyJscyI6IndlYnNpdGUiLCJsc2QiOiJkZXZlbG9wZXIubnZpZGlhLmNvbS8_ZmlsZW5hbWU9NDAzLmh0bWwifQ==&__token__=exp=1792974614~hmac=fdba4beae66b376d357633e0dd0009417b0da07d9b209000ba9ab61407fec61a

[10] Qualcomm, “Qualcomm Dragonwing™ IQ10 Series,” https://www.qualcomm.com/internet-of-things/products/iq10-series

[11] Rockchip, “Rockchip RK3588 Datasheet Rev. 1.9,” 2024. 12. 12. https://rockchips.net/wp-content/uploads/2025/03/Rockchip-RK3588-Datasheet-V1.9-20241212.pdf

[12] Hailo, “Hailo-15H AI Vision Processor,” 2026. 1. https://hailo.ai/hailo-files/hailo-15h-product-brief-en/

[13] Hailo, “Hailo-15L AI Vision Processor,” 2025. 12. https://hailo.ai/hailo-files/hailo-15l-product-brief-en/

[14] Renesas Electronics, “RZ/V2H Group Datasheet,” Renesas Electronics, 2026. 7. https://www.renesas.com/en/document/dst/rzv2h-group-datasheet

[15] Renesas Electronics, “RZ/V2H Evaluation Board Kit(Secure type) Hardware Manual Rev. 1.30,” 2026. 2. 2. https://www.renesas.com/en/document/mat/rzv2h-evaluation-board-kit-secure-type-hardware-manual

[16] T. Olujinmi, “Renesas RZ/V2H Cortex-A55/R8/M33 MPU comes with 80 TOPS AI accelerator for robotics and autonomous applications,” CNX Software, 2024. 3. 4. https://www.cnx-software.com/2024/03/04/renesas-rz-v2h-cortex-a55-r8-m33-mpu-80-tops-ai-accelerator-robotics-autonomous-applications/

[17] AMD, “Kria K26 SOM Data Sheet(DS987) Rev. 1.6,” 2026. 3. 25. https://docs.amd.com/r/en-US/ds987-k26-som

[18] Intel, “OpenVINO™ Toolkit for AI PC,” https://www.intel.com/content/www/us/en/developer/tools/openvino-toolkit/ai-pc.html

[19] Ambarella, “Ambarella Expands CV3 Family of Automotive AI Domain Controllers With New CV3-AD685,” 2023. 1. 5. https://investor.ambarella.com/node/14556/pdf

[20] Ambarella, “Ambarella CV5 8K AI Vision Processor Product Brief,” 2022. 11. https://www.einfochips.com/wp-content/uploads/2023/09/Ambarella_CV5_Product_Brief.pdf

[21] Mobilint, “ARIES, Ultra Performance AI Accelerator for On-Premises AI,” https://www.mobilint.com/aries

[22] Nextchip, “ADAS Processor - APACHE6,” https://www.nextchip.com/kr/adas/adas.php?idx=4

[23] Deeper-I, “Our Offerings,” https://deeper-i.ai/

[24] DEEPX, “DEEPX Products,” https://deepx.ai/products/*/

[25] iWave Global, “Versal AI Edge XQRVE2302 SOM for Satellite and Lunar Missions Onboard AI Compute,” 2026. 1. 2. https://iwave-global.com/articles/versal-ai-edge-xqrve2302-som-transforming-satellite-and-lunar-missions-onboard-ai-compute/

[26] AMD, “Space,” https://www.amd.com/en/solutions/aerospace-and-defense/space.html