# 온디바이스 AI를 위한 시스템 소프트웨어 기술 동향

## System Software Technology Trends for On-Device Artificial Intelligence


**강성주** (S.J. Kang, sjkang@etri.re.kr) 온디바이스시스템SW연구실 책임연구원/실장
**고동범** (D.B. Ko, dbko112@etri.re.kr) 온디바이스시스템SW연구실 선임연구원
**신용준** (Y.J. Shin, yjshin@etri.re.kr) 온디바이스시스템SW연구실 선임연구원
**전재호** (J.H. Jeon, jeonjaeho11@etri.re.kr) 온디바이스시스템SW연구실 선임연구원/기술총괄
**정영준** (Y.J. Jung, jjing@etri.re.kr) 온디바이스AI연구본부 책임연구원/본부장



**ABSTRACT**

On-device artificial intelligence (AI) is emerging as a core paradigm for intelligent services that demandreal-time response, energy efficiency, and privacy. Unlike traditional cloud-based models, this approach enables the direct inference of embedded hardware without external server dependency. This study examined the evolution of system software in support of on-device AI, highlighting the shift from resource-constrained embedded systems to AI-centric software stacks. Key themes include mission, accuracy, latency, energy (MALE)-oriented execution, hybrid co-inference with edge/cloud, DevOps-enabled CI/CD, and over-the-air (OTA) update pipelines. Using industry cases in the automotive, robotics, and mobile domains, we analyzed how system software is adapting to enhance the intelligence, autonomy, and scalability of AI-powered devices in the post-cloud era.



**KEYWORDS** AI Framework, Neural Processing Unit, On-device AI, Software-Defined Infrastructure, System Software


## I. 서론

음성비서, 웨어러블, 로봇청소기, 자율주행차, 휴머노이드 등 지능형 디바이스의 확산과 함께 인공지능(AI)의 실행 방식도 근본적으로 변화하고 있다. 기존에는 클라우드에서 AI 연산이 이루어졌다면, 최근에는 단말기 내에서 직접 추론을 수행하는 온디바이스 AI(On-Device AI)가 새로운 패러다임으로 부상하고 있다. 이는 단순한 처리 위치 이동을 넘어, 지연(Latency) 감소, 프라이버시 보호, 네트워크 · 전력 등


* DOI: https://doi.org/10.22648/ETRI.2025.J.400509
* 이 논문은 2024년도 정부(과학기술정보통신부)의 재원으로 정보통신기획평가원의 지원을 받아 수행된 연구 결과임[No. RS-2024-00406245, 미래 모빌리티를 위한 소프트웨어 정의형 인프라스트럭처 기술 개발].

현실적 요구에 대응한 기술 진화의 결과이다[1].

온디바이스 AI는 디바이스 내부에서 데이터의 분석과 의사결정을 수행하는 자율성을 제공하며, 내부적으로 민감 정보를 처리하여 보안성도 확보한다. 또한, 배터리 기반 디바이스에서는 에너지 효율성이 필수이며, 산업현장 등 불안정한 네트워크 환경에서도 독립 작동이 가능해야 한다. Google은 이를 "사용자 디바이스 내 머신러닝 처리로 앱의 신뢰성 · 프라이버시 · 사용자 경험을 향상시키는 기술"로 정의하고, TinyML 재단은 "mW급 전력으로 항상 켜진 초소형 마이크로컨트롤러 기반 추론 기술"로 설명한다. 삼성전자는 "기기 내부에서 데이터 수집 · 분석을 수행해 지연을 줄이고, 네트워크 없이도 AI 기능을 제공하는 방식"이라 정의한다. 즉, 표현은 달라도 저지연 로컬 추론, 프라이버시 보호, 에너지 절감, 오프라인 독립성, 경량 실행 등은 공통된 핵심 특성이다.

하지만 이러한 요구를 충족하려면 단순한 AI 모델의 경량화만으로는 부족하다. 추론 전용 하드웨어(NPU, AI SoC 등)와 이를 정밀 제어 · 지원하는 시스템 소프트웨어의 통합적 구조가 필수적이다. 특히 온디바이스 AI는 GPU 기반 고성능 환경이 아닌 연산 · 전력 · 메모리가 제한된 디바이스에서 동작하므로, 운영체제, 컴파일러, 런타임 프레임워크 등으로 구성된 시스템 소프트웨어 계층이 유기적으로 설계 · 운영되어야만 정확도, 신뢰성, 실시간성을 보장할 수 있다.

또한, 지능형 시스템은 고정 기능 수행을 넘어서, 주변 상황에 따라 판단을 갱신하고 행동을 동적으로 재구성해야 하는 환경에 놓여 있다. 따라서 온디바이스 AI 단독 실행으로는 한계가 있으며, 엣지-클라우드와 연계, 지속적 통합과 배포, OTA(Over-The-Air)를 통한 업데이트 등이 중요해지고 있다.

이에 본고는 온디바이스 AI를 실행할 수 있게 하는 시스템 소프트웨어의 구조와 그 진화 방향을 중심으로 기술적 배경을 조망하고자 한다. Ⅱ장에서는 온디바이스 전용 하드웨어 생태계를 살펴보고, Ⅲ장에서는 운영체제, 런타임, 컴파일러, 프레임워크 등 AI 중심 시스템 소프트웨어 스택의 양상을 분석한다. Ⅳ장에서는 온디바이스 AI의 실행 확장성을 뒷받침하는 기술 동향을 고찰하며, 마지막으로 Ⅴ장에서는 이러한 흐름을 바탕으로 향후 기술 발전과 정책 대응 방향을 제언한다.

## Ⅱ. 온디바이스 HW 개발 동향

### 1. 온디바이스 HW 생태계

온디바이스 AI 구현을 위해 각 기업은 신경망 처리장치(NPU: Neural Processing Unit)나 AI SoC 형태의 전용 하드웨어를 개발하고 있으며, 글로벌 대기업부터 국내 스타트업까지 다양한 제품이 출시되고 있다. 이들은 연산 성능, 전력 효율, 아키텍처 구조, 소프트웨어 지원 등에서 차별화를 추구하며, 타겟 애플리케이션도 자율주행, 로봇, IoT 디바이스 등으로 세분화되고 있다. 본 절에서는 NVIDIA, Hailo, EdgeCortix 등 주요 글로벌 기업과 DEEPX, Mobilint, AimFuture 등 국내 기업들의 대표 제품을 중심으로, 성능 지표, 기술 특징, 활용 분야를 비교하고 표 1에 주요 사양과 차별화 요소를 정리하였다.

### 2. Nvidia Jetson 시리즈

Jetson 시리즈는 NVIDIA가 개발한 디바이스용 SoM(System-on-Module) 플랫폼으로, 자율주행, 로봇, 드론, 산업용 장비 등 다양한 온디바이스 AI 응용 분야에서 널리 활용된다. 최신 Jetson 제품군은 고성능(AGX), 중간급(NX), 저전력(Nano)으로 구성되어, 전력 효율과 성능, 폼팩터 측면에서 다양한 요

표 1 주요 온디바이스 AI 하드웨어 비교

| 기업 | 주요 제품 | 성능 지표(연산@전력) | 기술 특징 | 타겟 응용 |
|---|---|---|---|---|
| NVIDIA | Jetson AGX/NX/Nano | 5-275 TOPS@5-60W | ARM+GPU+텐서코어 | 자율주행, 로봇, 드론 |
| Hailo | Hailo-8/10H/15 | 40 TOPS@<3.5W | 분산형 NPU, INT4 | AI카메라, 도시인프라 |
| EdgeCortix | SAKURA-II | 60 TOPS@10W | DNA, 동적 재구성 | UGV, 드론, 스마트시티 |
| Kneron | KL720/530/730 | 1 TOPS@1W | SoC+ISP 통합 | IoT, 보안기기 |
| SiMa.ai | MLSoC | 50 TOPS@<10W | MLSoC 통합형 | 의료기기, 비전AI |
| Mythic | MAP | 25 TOPS@<1W | 아날로그 인메모리 | 웨어러블, 음성I/F |
| 딥엑스 | DX-V1/V3, DX-M1 | 5-25 TOPS@1-5W | 스파스 연산, 압축, Int8 | 카메라, 가전, 로봇 |
| 모빌린트 | REGULUS, ARIES | 10-80 TOPS@3-25W | NPU+ISP, LLM 지원 | 엣지서버, 자율주행 |
| 에임퓨처 | NMP-350/550/750 | 1-16 TOPS@미확인 | MAC 어레이, 로컬메모리 | 가전, IoT, 차량용 |

구를 충족한다. 각 모듈은 ARM 기반 CPU, Ampere GPU, Tensor 코어, NVDLA 딥러닝 가속기, ISP 등을 이기종 통합한 SoC 아키텍처를 채택하고 있다.

Jetson 플랫폼은 CUDA, TensorRT, cuDNN 등과 통합된 JetPack SDK를 통해 개발 생태계를 지원하며, 리눅스 기반 JetPack OS 위에서 PyTorch, TensorFlow 등의 주요 AI 프레임워크를 안정적으로 실행할 수 있다. 실효 추론 성능과 개발 편의성이 뛰어나 연구용을 넘어 산업용 AI 제품 개발의 표준 플랫폼으로 자리매김했으며, 특히 영상처리와 딥러닝 중심의 엣지 AI 응용에서 강력한 성능을 보인다.

## 3. Hailo 고효율 NPU 시리즈

이스라엘 스타트업 Hailo는 2017년 설립된 온디바이스 AI 반도체 전문 기업으로, GPU와는 다른 ANN(인공신경망) 특화 구조를 기반으로 고효율 엣지 추론에 최적화된 NPU 아키텍처를 개발했다[2,3]. 대표 제품 Hailo-8은 단일 칩 기준 26 TOPS 성능과 3 TOPS/W의 전력 효율을 갖추고 있으며, M.2 및 PCIe 모듈 형태로 확장할 수 있다. 최신 Hailo-10H는 INT4 정밀도로 40 TOPS를 3.5W 미만 전력으로 처리해 대형 언어모델(LLM: Large Language Model)의 엣지 실행도 지원하며, Hailo-15는 영상 인식 중심의 비전 AI용 SoC 제품이다.

Hailo의 칩은 연산 유닛과 메모리가 밀접히 결합된 분산형 아키텍처를 채택하여, 병렬 처리와 메모리 접근 효율을 극대화했다. 이로써 스마트 카메라, 도시 인프라, 산업용 로봇, ADAS 등 다양한 응용에 적합한 초저전력 엣지 AI 플랫폼을 제공한다. 또한, 자체 SDK 및 컴파일러 기반 도구체인을 통해 높은 개발 편의성을 갖추고 있으며, LLM 시대에 대응 가능한 초경량 AI 연산 솔루션으로 주목받고 있다.

## 4. EdgeCortix SAKURA 플랫폼

EdgeCortix는 일본 도쿄에 본사를 둔 팹리스 기업으로, “소프트웨어 우선” 설계 철학을 바탕으로 동적 재구성형 AI 가속기(DNA: Dynamic Neural Accelerator) 아키텍처 기반의 SAKURA-II 플랫폼을 개발하였다[4]. 이 칩은 60 TOPS의 INT8 성능을 10W 내외의 전력으로 구현하며, M.2 및 PCIe 모듈 형태로 제공되어 다양한 임베디드 플랫폼과의 연계가 쉽다. 특히 라즈베리파이5와의 결합을 통해 저비용 환경에서도 비전 트랜스포머(ViT)나 멀티모달 모델 실행이 가능하다는 점에서 주목받는다. SAKURA-II

는 16nm 공정 기반으로 6 TOPS/W 이상의 전력 효율을 실현하며, 방위산업, 드론, 스마트시티, 산업 AI 등 고성능 · 저지연이 요구되는 분야에 적합하게 설계되었다.

이 플랫폼은 전용 컴파일러 MERA를 통해 ONNX, TensorFlow, PyTorch 기반 모델을 별도 수정 없이 최적화된 실행 코드로 변환할 수 있으며, 실시간 추론에 최적화된 런타임 구조를 통해 다중카메라 스트리밍 환경에서도 낮은 지연과 높은 처리량을 제공한다. 특히 동적 인터커넥트 구조를 활용해 신경망 구조에 따라 내부 연산 경로를 소프트웨어적으로 재구성할 수 있어 다양한 AI 네트워크에 대해 효율적인 병렬 처리가 가능하다.

## 5. 기타 글로벌 온디바이스 HW

온디바이스 AI 시장에서는 독창적 기술을 바탕으로 주목받는 글로벌 스타트업들도 등장하고 있다. Kneron은 컴퓨터 비전과 경량 추론에 최적화된 SoC 기반 NPU 제품(KL720, KL530 등)을 개발했으며, KL720은 약 1W 전력에서 1 TOPS 성능을 제공해 얼굴 인식 · 객체 탐지에 특화되었다. ISP와 이미지 처리 모듈을 내장해 스마트 IoT 카메라에 적합하며, 자체 런타임(KneoRT)과 경량화 툴을 통해 초소형 디바이스에서 효율적인 AI 추론을 지원한다[5].

SiMa.ai는 이미지 센싱부터 추론까지를 단일 칩에서 처리하는 통합형 MLSoC를 개발했다[6]. 이 칩은 Arm Cortex-A CPU, ISP, NPU, DDR 컨트롤러를 통합해 10W 미만에서 50 TOPS 성능을 제공하며, 산업용 비전카메라나 의료기기 등에 적합하다. TensorFlow 등 주요 프레임워크와 호환되며, DevOps 기반 모델 관리도 지원한다.

Mythic은 DRAM이 없는 아날로그 인-메모리 컴퓨팅 구조를 통해 수 mW급 초저전력으로 추론이 가능한 NPU를 구현했다[7]. 연산과 저장을 통합한 구조로 병목을 최소화하며, 1W 이하 전력에서 25 TOPS 성능을 제공해 IoT 센서, 웨어러블 등에서 활용 가능성이 크다.

## 6. 국내 온디바이스 HW 동향

국내에서도 온디바이스 AI 전용 칩과 생태계 구축을 위한 스타트업 활동이 활발하다. 딥엑스(DEEPX)는 DX-V1, DX-M1, DX-V3 등 다양한 NPU 제품군을 보유하고 있으며, 스파스 연산 최적화, 가중치 압축, 고효율 메모리 계층 등을 통해 1~5W 내외 전력에서도 높은 FPS/TOPS 성능을 구현한다. 특히 DX-M1은 $1/TOPS 이하의 원가 경쟁력을 목표로 하며, 카메라, 가전, 로봇 분야의 글로벌 기업들과 협업을 확대 중이다[8].

모빌린트(Mobilint)는 REGULUS(10 TOPS급 SoC)와 ARIES(80 TOPS급 가속기)를 중심으로 엣지 추론 최적화 칩을 개발하고 있다[9]. 자체 SDK와 Transformer · LLM 대응력을 기반으로 산업용 AI 서버 및 자율주행 엣지 노드에 적용되며, 최근에는 MXM 모듈형 제품군을 통해 확장성도 강화하고 있다.

에임퓨처(AimFuture)는 NeuroMosaic 프로세서 기반 NPU 코어와 통합 툴체인을 통해 저전력 엣지용 SoC를 개발하고 있다[10]. 해당 기술은 LG전자 제품 및 IoT 센서에 적용되며, 구조적 Sparsity와 로컬 메모리 최적화를 통한 고효율 연산이 강점이다.

# III. 온디바이스 시스템SW 동향

## 1. AI 중심 SW 스택으로의 전환

AI 시대의 시스템 소프트웨어는 기존의 범용 컴퓨팅 환경과는 다른 방향으로 빠르게 진화하고 있다. 표 2는 온디바이스 AI 실행 환경에서의 시스템

**표 2** 전통 소프트웨어 스택과 AI 중심 온디바이스 스택 비교

| 계층 | 전통 소프트웨어 스택 | AI 중심 소프트웨어 스택 |
|---|---|---|
| 응용 | 범용 앱(C/C++, Java 기반) | AI 추론 앱(PyTorch, TensorFlow, ONNX 등) |
| 런타임 | JVM, OpenMP 등 | AI 런타임(ONNX, TFLite, PyTorch Mobile, HailoRT) |
| 컴파일러 | gcc, clang, javac | AI 컴파일러(XLA, MLIR, Glow, MERA 등) |
| 운영체제 | Linux, Android, RTOS | 경량 실시간 OS(RT-Linux, Zephyr, FreeRTOS, DriveOS 등) |
| 하드웨어 | CPU 중심 범용 SoC | 이기종 AI SoC(NPU, GPU, DSP 등) |

소프트웨어 계층을 기존 구조와 비교해 정리한 것으로, 과거에는 운영체제(OS)와 범용 컴파일러(C/C++, Java 등)가 소프트웨어 스택의 중심이었다면, 현재는 AI 추론을 위한 전용 컴파일러와 프레임워크가 중심으로 자리 잡고 있다. 특히 온디바이스 환경에서는 하드웨어와 긴밀히 연동되는 최적화된 경량 스택 구성이 필수이며, NPU, GPU, DSP 등 다양한 이기종 연산 자원에 적합한 구조가 요구된다.

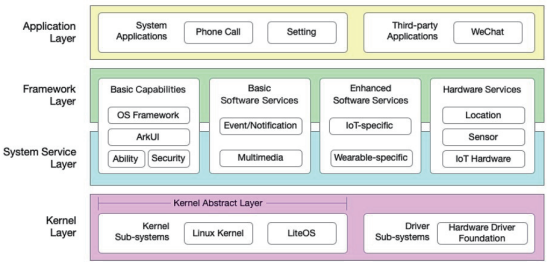
대표적으로 NVIDIA의 Jetson 시리즈는 CUDA, cuDNN, TensorRT, DeepStream SDK 등을 포함한 JetPack 생태계를 통해 고성능 온디바이스 추론 환경을 제공하며, 사실상 업계 표준 플랫폼으로 자리매김하고 있다. Hailo는 HailoRT 런타임과 INT4 전용 컴파일러 체계, 전용 도구체인을 통해 초저전력 기반의 고효율 추론을 실현하고 있으며, EdgeCortix는 MERA 컴파일러와 다양한 프레임워크(ONNX, PyTorch, TensorFlow 등) 호환성을 바탕으로 DNA 아키텍처의 동적 인터커넥트를 최적 활용할 수 있는 구조를 갖추고 있다.

이처럼 주요 기업들은 단순한 NPU 제공을 넘어서, 해당 하드웨어에 최적화된 AI 컴파일러, 런타임, 개발 툴킷이 통합된 소프트웨어 스택을 중심으로 경쟁하고 있다. 이는 온디바이스 AI가 단일 모델 실행을 넘어 복수 모델, 멀티센서, 실시간 대응 등 복잡한 지능형 서비스로 확장되기 때문이다. 각 스택은 단순 코드 실행 외에도 정밀도 변환, 연산 병렬화, 메모리 스케줄링, 인터럽트 처리 등 다양한 최적화 기능을 포함하며, AI 응용의 실효 성능을 좌우하는 핵심 요소로 작용한다.

## 2. 온디바이스 운영체제 기술

온디바이스 AI를 위한 운영체제는 저전력, 실시간성, 이기종 연산 자원 통합, 보안성 등을 동시에 요구받는다. 특히 최근의 온디바이스 환경은 단일 AI 모델을 단말에서 실행하는 수준을 넘어, 복수의 AI 인스턴스, 다중 센서 처리, 지능형 제어 루프를 동시에 만족시켜야 하며, 이에 따라 기존 리눅스나 Android 기반 범용 OS의 한계를 극복하고자 다양한 기업과 연구기관들은 온디바이스 전용 운영체제를 개발하고 있다.

Huawei의 HarmonyOS는 IoT, 모바일, TV, 자동




출처 Reprinted from L. Li et al., "Software Engineering for OpenHarmony: A Research Roadmap," arXiv preprint, 2023. doi: 10.48550/arXiv.2311.01311 [11]

**그림 1** NAS의 모델 구조 탐색 공간 예시

차, 웨어러블 등 다양한 기기를 하나의 플랫폼으로 통합하기 위해 설계된 마이크로커널 기반 분산형 운영체제이다(그림 1)[12]. 자체 AI 프레임워크인 MindSpore를 Ascend AI SoC와 통합하여, 추론뿐만 아니라 학습 · 전이학습도 엣지에서 분산 처리가 가능하도록 구성했다. 또한, NPU · DSP 등 이기종 프로세서를 효율적으로 스케줄링하는 런타임 기반을 통해 온디바이스에서도 멀티모달 지능처리를 지원한다.

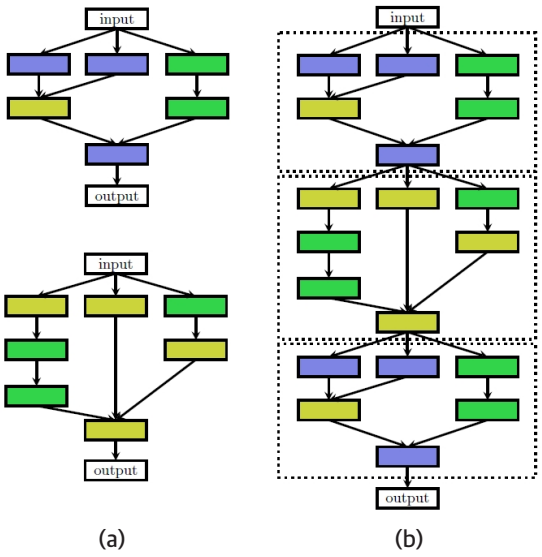
NVIDIA의 Jetson/Drive 플랫폼은 두 가지 OS 스택으로 구성된다. Jetson은 Ubuntu 기반 JetPack OS를 통해 개발자 친화적인 리눅스 환경과 CUDA, TensorRT, DeepStream SDK 등 AI 추론 스택을 제공한다. 반면 Drive OS는 자율주행 SoC용 실시간 운영체제로, ISO 26262 ASIL-D 등 기능 안전 인증을 충족하며, RTOS 커널과 NVIDIA Hypervisor를 기반으로 차량 내 제어와 AI 추론 시스템을 통합 관리한다[13]. 복수 하드웨어 자원의 격리 · 할당도 지원한다.

Apex.AI의 Apex.OS는 ROS2를 실시간 RT-Linux 환경에 맞게 재구성한 운영체제로, 자율주행 및 로봇 시스템에서 실시간성과 기능안전성을 동시에 확보하도록 설계되었다[14]. POSIX 기반 RTOS 커널 위에 구축되며, Zero-Copy 통신, 결정론적 메모리 할당 등 고신뢰 기능을 통해 오픈소스 ROS2의 통신 지연 문제를 해결한다.

## 3. AI 컴파일러 및 모델 최적화 기술

온디바이스 AI 환경에서는 모델을 실행하기 이전 단계에서, 하드웨어 구조에 맞춘 정적 분석과 최적화 컴파일이 핵심적이다. AI 컴파일러는 모델 구조를 분석하여 대상 NPU 또는 GPU에 적합한 연산 그래프를 생성하고, 연산 정밀도 변환, 연산 병합, 스케줄링 등의 과정을 수행함으로써 실행 성능과 에너지 효율을 동시에 확보한다. 이는 특히 배터리 기반 디바이스나 실시간 제어형 지능시스템에서 필수적인 요건이다.

대표적으로 NVIDIA는 TensorFlow, PyTorch, ONNX 모델을 Jetson 플랫폼에 최적화하는 TensorRT 도구를 제공한다. TensorRT는 레이어 병합, INT8 양자화, 연산 재배치를 자동화하며, JetPack SDK와 통합돼 빠른 추론 속도와 개발 편의성을 제공한다. Hailo는 INT4 기반 NPU에 특화된 Hailo Compiler로 모델 파이프라인 분할, 연산 순서 최적화, 자동 프로파일링을 수행해 높은 수준의 전력 효율을 실현한다. EdgeCortix의 MERA는 정적 컴파일과 동적 구성을 병행해 다양한 환경에서도 일관된 성능을 보이며, 주요 프레임워크와도 높은 호환성을 갖춘다. Intel의 OpenVINO는 operator fusion, constant folding, INT8 양자화 등으로 Intel NPU · CPU 환경에서 경량화된 AI 추론을 지원한다.



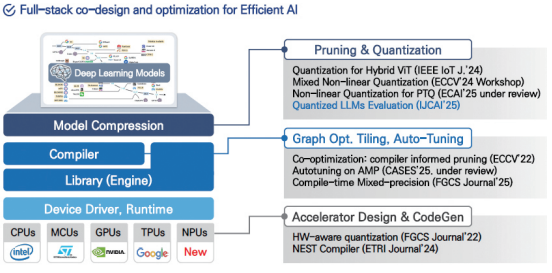

출처 Reprinted from T. Elsken et al., "Neural Architecture Search: A Survey," J. Mach. Learn. Res., vol. 20, 2019, pp. 1-21. [15].

**그림 2 NAS의 모델 구조 탐색 예시: (a) 모델 기초 셀 (b) 셀을 쌓아 생성한 아키텍처**

범용 AI 컴파일러 분야에서는 Google의 XLA, Meta의 Glow, Apache TVM, MLIR 기반 IREE 프로젝트 등이 활발히 개발 중이다. 이들은 하드웨어 맞춤형 모델 재구성, 플랫폼 간 커널 튜닝 등 고유 기능을 통해 다양한 아키텍처에 최적화된 실행 코드를 생성한다. 특히 MLIR은 다양한 연산 표현을 통합 처리하여 이기종 AI 가속기 환경에서의 유연한 코드 생성을 가능하게 하는 핵심 기술로 주목받고 있다.

최근 AI 컴파일러의 주요 발전 흐름 중 하나는 Neural Architecture Search(NAS) 기반의 모델 구조 최적화다. NAS는 그림 2의 예시와 같은 AI 모델의 구조를 정의하는 탐색 공간에서 정확도, 파라미터 수, 연산량, 메모리 사용 등을 종합 고려해 특정 하드웨어에 최적화된 AI 모델을 자동 설계하는 기술로, Google을 시작으로 UP-NAS, ARNAS, BPNAS 등 다양한 경량화 중심 기법이 등장했다. 특히 NAS는 온디바이스 AI에 적합한 경량 모델 자동 설계와 최적화를 가능케 하며, AI 컴파일러와의 통합을 통해 시너지 효과를 내고 있다.

국내에서는 ETRI가 자체 개발한 AI 컴파일러 NEST-C를 중심으로, 그림 3과 같이 효율적인 온디바이스 AI 실행을 위한 풀스택 소프트웨어 기술 체계를 구축하고 있다[16]. 이 구조는 △모델 압축 (Model Compression) △컴파일러 최적화(Graph Optimization, Tiling, Auto-Tuning) △가속기 설계 및 코드 생성(Accelerator Design & CodeGen) 등 전 계층에 걸친 통합적 설계(Co-Design)를 지향한다. 이러한 기술은 실제 현장에서 검증되고 있다. 예컨대, ETRI 컨퍼런스에서는 시각장애인용 사족보행 보조로봇에 NVIDIA AGX Orin 보드를 탑재하고, 온디바이스 AI 기술만으로 최신 비전-언어모델(VLM)과 음성합성기(TTS)를 실시간 실행하는 실증 결과를 시연하였다.




출처 Reprinted from 권용인, "사족로봇 가이드독 사례로 살펴보는 온디바이스AI 최적화 기술," 2025 온디바이스 인공지능반도체 워크숍.

**그림 3 ETRI의 온디바이스AI 최적화 스택**

## 4. AI 런타임 프레임워크

온디바이스 AI 시스템에서 런타임 프레임워크는 컴파일러가 생성한 연산 그래프를 디바이스상에서 실제로 실행하는 핵심 계층이다. 연산자 호출, 순서 조정, 메모리 관리, 가속기 제어, 병렬처리 등을 수행하며, 추론 속도, 지연, 에너지 효율 등 시스템 성능에 직접 영향을 미친다. 특히 하드웨어 구조에 맞춘 최적화된 런타임의 적용이 성능 극대화의 핵심 전략이다.

TensorFlow Lite는 모바일 및 엣지 디바이스에 최적화된 경량 런타임으로, 양자화 모델과 GPU · DSP · NPU delegate를 통해 다양한 하드웨어 가속기를 지원한다[17]. PyTorch Mobile은 TorchScript 기반 정적 그래프 최적화와 안드로이드/iOS 연동 기능을 갖춰 스마트폰 AI 앱 구현에 적합하다. ONNX Runtime은 다양한 프레임워크 모델을 범용적으로 실행하며, TensorRT, OpenVINO, CoreML 등과의 연동으로 플랫폼별 추론 가속을 지원한다.

NVIDIA TensorRT는 ONNX 및 TensorFlow 모델을 GPU 기반 엣지 디바이스에 최적화해 실행하는 통합형 컴파일러 · 런타임으로, 레이어 병합, INT8 양자화, 스케줄링 등을 통해 고속 · 저지연 추론을


그림 4 온디바이스 AI 시스템의 개발/실행/운영을 위한 소프트웨어 정의형 인프라스트럭처 기술

구현하며 JetPack SDK와 통합되어 개발 효율성도 높다[18]. Hailo의 HailoRT는 INT4 기반 NPU에 최적화된 런타임으로, 파이프라인 분할 및 스케줄링 자동화를 통해 전력 효율(TOPS/W)을 극대화하며 프로파일링 도구도 지원한다.

Intel OpenVINO Runtime은 PTQ Toolkit과 연계해 양자화 정확도와 지연 최소화를 동시에 달성하며, CPU · GPU · VPU 등 다양한 인텔 하드웨어에 최적화된 실행을 지원한다[19].

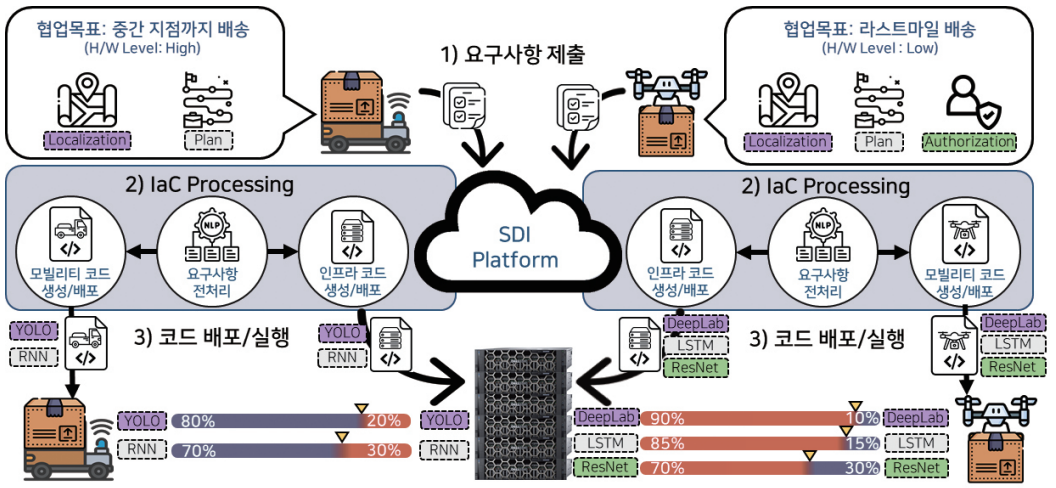
## IV. 온디바이스 AI 연계 기술 동향

### 1. 미션 주도 온디바이스 AI 실행 전략

온디바이스 AI는 단순히 경량화된 모델을 디바이스에 탑재하는 수준을 넘어서, 실질적인 임무(Mission)를 수행하는 지능형 시스템으로 진화하고 있다. 이에 따라 AI의 동작은 정해진 환경과 목적에 최적화되어야 하며, 이를 위한 핵심 설계 기준으로 MALE(Mission, Accuracy, Latency, Energy) 프레임워크가 주목받고 있다. MALE은 AI의 정확도, 반응속도, 에너지 효율 간의 균형을 임무 중심으로 설계하기 위한 기준으로, 실제 사용 시나리오에 따라 각 요소의 우선순위가 달라질 수 있다[20].

예를 들어, 자율주행차 · 드론처럼 실시간 반응이 필수인 시스템은 클라우드 의존 없이 저지연으로 추론 가능한 구조가 필요하며, 이때는 정확도보다 속도와 전력 제약 내 처리 가능성이 중요하다. 반대로, 장시간 작동하는 IoT 환경에서는 에너지 효율성을 우선시하여 AI 연산을 최소화하거나, 상황 발생 시에만 활성화하는 방식이 적합하다. 이를 실현하는 기술로는 저전력 NPU와 서버급 NPU 간 협력 추론(Co-inference), 프루닝 · 양자화 기반 모델 경량화, 조기 종료(Early Exit), 동적 실행모드 등이 있으며, 이는 임무 조건에 따라 유연하게 성능을 절충한다.

특히 ETRI에서는 이러한 MALE 기반 설계 철학을 바탕으로, 그림 4의 SDI(Software Defined Infrastructure) 기술을 개발 중이다[21]. SDI는 다수의 온디바이스 AI 디바이스가 배치된 환경에서 임무 · 자원 · 네트워크 상황을 실시간 판단하여, 각 디바이스가 최적의 방식으로 모델을 분할 · 실행 · 전개할 수 있도록 지원하는 분산 지능 실행 인프라다. 그림 5는 이를 통해 온디바이스 AI 시스템의 동적 협업과 배포/실행 구조를 보여준다. 이는 MALE 요구를 만족시키는 실질적인 시스템 아키텍처로 작동하며, 향후 대규모 협력형 AI 디바이스 생태계의 기반이 될 수 있다.




출처 Reproduced from 전재호 외, "미래 모빌리티를 위한 소프트웨어 정의형 인프라스트럭쳐 기술," 정보과학회지 제42권 제11호, 2024, pp. 16-23.

**그림 5 온디바이스 AI 시스템의 동적 협업과 배포/실행 구조**

## 2. 온디바이스 중심 협업 추론

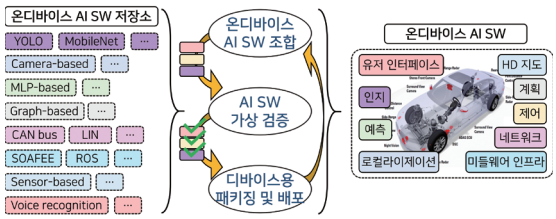
온디바이스 AI 성능 향상을 위한 전략으로 협업 추론(Co-Inference) 구조가 주목받고 있다. 이 방식은 디바이스가 1차 추론을 담당하고, 복잡하거나 불확실한 경우에만 클라우드 자원을 활용하는 하이브리드 구조다. 이를 통해 전체 지연과 전력 소모를 줄이고, 프라이버시도 보호할 수 있다.

협업 추론은 두 가지 주요 방식으로 구분된다. 첫째, 모델 분할 방식은 신경망을 중간에서 나눠 전단은 디바이스, 후단은 클라우드에서 처리하는 구조로, Microsoft의 Edgent가 대표적이다. 이때 피처 전송 효율이 관건이며, Bottlenet++ 등의 압축기법과 동적 분할 전략이 활용된다. 둘째, 결과 보정 방식은 디바이스에서 경량 모델로 1차 결과를 내고, 필요 시에만 클라우드에 보정 요청하는 구조다. Google Assistant, Apple Siri 등이 이 구조를 따른다.

연구 사례로는 Neurosurgeon[22], DeepDecision[23]이 있으며, 최근에는 SLM(소형 언어모델)과 클라우드 LLM을 연계한 하이브리드 방식도 등장하고 있다. 산업적으로는 Tesla 차량이나 스마트 CCTV처럼 대부분의 AI 연산은 온디바이스에서 처리하되, 특정 상황에만 클라우드 협력을 활용하는 방식이 확대되고 있다. 이러한 협업 추론은 클라우드와 디바이스 간 역할 분담을 통해 효율성과 실시간성을 모두 확보할 수 있는 차세대 AI 실행 구조로 주목받고 있다.

## 3. 온디바이스 AI의 CI/CD 및 OTA

온디바이스 AI가 실제 산업에서 안정적으로 작동하려면, 초기 탑재 이후에도 성능 개선, 오류 수정,




출처 Reproduced from 전재호 외, "미래 모빌리티를 위한 소프트웨어 정의형 인프라스트럭쳐 기술," 정보과학회지 제42권 제11호, 2024, pp. 16-23.

**그림 6 ETRI 온디바이스 AI CI/CD 기술 개요**

기능 추가가 지속적으로 이루어져야 한다. 특히 자율주행차, 로봇, UAM처럼 신뢰성과 실시간성이 동시에 요구되는 디바이스에서는 소프트웨어의 정기적 갱신이 핵심 전략이다. 이를 위해 산업계는 클라우드 기반 CI/CD-OTA 체계를 도입하고 있으며, 클라우드에서 학습된 AI 모델을 무선으로 디바이스에 배포해 자동화된 운영환경을 구축하고 있다.

ETRI에서는 그림 6의 온디바이스 AI SW의 지속적인 조합/배포(CI/CD) 기술을 개발하고 있다[21]. 온디바이스 AI 환경의 동적 특성과 SW 스택 요구사항을 고려하여, 최적화된 AI 모델 및 소프트웨어 모듈을 저장소로부터 자동으로 가져오고 조합한다. 이렇게 구성된 소프트웨어는 클라우드 환경에서 성능의 가상 검증 과정을 거친 뒤, 각 온디바이스 기기의 하드웨어 사양과 SW 스택을 고려해 자동으로 패키징되어 배포된다. 이와 같은 자동화된 개발 및 배포 체계는 온디바이스 기기마다 상이한 HW 및 SW 스택에도 일관된 품질의 AI SW를 빠르게 공급할 수 있도록 하며, 개발과 운영 비용을 동시에 절감하는 효과를 기대할 수 있다.

이 밖에도 Tesla는 FSD를 통해 차량 주행 데이터 기반 재학습과 OTA 업데이트를 통해 지속적 성능 개선을 가능케 하며[24], Microsoft의 AVOps는 차량 AI의 전체 수명주기를 클라우드에서 통합 관리하는 방식을 제안한다[25]. 이 구조는 단순 자동화를 넘어, 안전성 검증, A/B 테스트, 단계적 롤아웃 등 정교한 운영 기능도 포함한다. 또한, Mender.io, Sonatus 등의 플랫폼은 산업용 디바이스에 맞춘 OTA 기능을 지원하며, 온디바이스 AI는 클라우드와 연결된 “살아 있는 소프트웨어”로 진화 중이다. 이는 지능형 시스템의 핵심 인프라로 자리 잡고 있다.

## V. 결론

온디바이스 AI는 단순한 클라우드 대체가 아니라, 단말이 독자적으로 AI를 실행하기 위해 필요한 소프트웨어의 실시간성, 디바이스의 구동과 AI 실행과 관련된 에너지 효율, 프라이버시, 독립성 등 복합적인 요구를 만족시켜야 하는 새로운 지능형 컴퓨팅 구조다. 이를 가능케 하기 위해 시스템 소프트웨어는 기존 임베디드 SW에서 벗어나 AI 중심 구조로 재편되고 있다. 본고에서는 이 흐름을 세 가지 축으로 정리하였다.

첫째, 온디바이스 AI의 실행 환경 특성과 하드웨어 발전에 따라, 경량 고성능 AI 추론을 지원하는 NPU 중심 HW 생태계와 이에 최적화된 시스템 SW 요구 조건을 분석하였다. 둘째, 운영체제 · 컴파일러 · 프레임워크 등 시스템 SW 전반이 AI 중심의 스택으로 재구성되는 흐름과 NVIDIA, Hailo, ETRI 등 주요 기업 · 기관의 최신 지원 기술 동향을 살펴보았다. 셋째, DevOps와 온디바이스의 융합, 즉 SDV · 로봇 등 실시간 시스템에서 클라우드 수준의 지속적 업데이트와 협업 추론이 가능한 실행 구조로 확장되는 방향성을 조명하였다.

이처럼 온디바이스 AI는 점차 스마트 디바이스의 표준으로 자리 잡으며, 하드웨어뿐만 아니라 시스템 소프트웨어가 성능 · 응답성 · 에너지 효율성의 핵심으로 부상하고 있다. 특히 자율주행차, 로봇, 드론 등 미션 중심 시스템에서는 온디바이스 중심 실행이 필수이며, 이를 위해 소프트웨어도 AI 적응성과 자동화 역량을 갖춘 지능형 OS · 프레임워크로 진화해야 한다. 정책 및 산업 측면에서도, 온디바이스 AI의 MALE 요구조건을 만족시키는 기술 기반을 조기 확보하고, 하드웨어-소프트웨어 수직 통합 생태계를 구축하는 것이 국가적 경쟁력 확보의 관건이 될 것이다.

**용어해설**

**ONNX** 다양한 딥러닝 프레임워크 간 모델 호환을 가능하게 하는 오픈 포맷. PyTorch, TensorFlow 등에서 학습한 모델을 ONNX로 변환해 다른 플랫폼에서 실행 가능하게 하는 기술

**OTA** 무선 네트워크를 통해 소프트웨어나 펌웨어를 원격으로 업데이트하거나 배포하는 기술. 온디바이스 AI 시스템에서는 모델이나 시스템 소프트웨어의 지속적 개선을 위한 필수 기술 활용

**TOPS** 초당 테라(1조) 연산을 수행할 수 있는 능력을 의미하는 지표로, AI 추론 성능을 평가할 때 사용됨. AI SoC나 NPU의 연산 처리량을 수치화하는 데 자주 사용

## 참고문헌


[1] X. Wang et al., "Empowering Edge Intelligence: A Comprehensive Survey on On-Device AI Models," ACM Comput. Surv., vol. 57, no. 9, 2025, pp. 1-39.

[2] Hailo Website. https://hailo.ai/products/ai-accelerators/hailo-8l-m-2-ai-acceleration-module-for-ai-light-applications/#hailo8lm2-overview

[3] Hailo Community, "How to connect a Hailo-8 to the Raspberry Pi 5," 2024. 3. https://community.hailo.ai/t/how-to-connect-a-hailo-8-to-the-raspberry-pi-5/183

[4] EdgeCortix Website. https://www.edgecortix.com/en/

[5] Kneron Website. https://www.kneron.com/page/soc/

[6] SiMa Website. https://sima.ai/hardware/

[7] Mythic Website. https://mythic.ai/products/m1076-analog-matrix-processor/

[8] DEEPX Website. https://deepx.ai/products/ai-chips/

[9] mobilint Website. https://www.mobilint.com/ko/aries

[10] AiM Future Website. https://aimfuture.ai/ko/products_ko/#neuromosaic-studio

[11] L. Li et al., "Software Engineering for OpenHarmony: A Research Roadmap," arXiv preprint, 2023. doi: 10.48550/arXiv.2311.01311

[12] Awesome HarmonyOS Contributors, "Awesome HarmonyOS," GitHub, 2025. https://github.com/Awesome-HarmonyOS/HarmonyOS

[13] NVIDIA Developer, "NVIDIA DriveOS SDK," 2024. https://developer.nvidia.com/drive/os

[14] D. Pangercic, "How Apex.AI Certified ROS 2 According to ISO 26262 ASIL-D," Automatisiertes Fahren 2021(AWF-TSC April 2021), 2021.

[15] T. Elsken et al., "Neural Architecture Search: A Survey," J. Mach. Learn. Res., vol. 20, 2019, pp. 1-21.

[16] 2025 온디바이스 인공지능반도체 워크숍 웹사이트. https://www.theise.org/eventin/ws250522/

[17] TensorFlow, "TensorFlow Lite," https://www.tensorflow.org/lite/guide?hl=ko

[18] NVIDIA, "TensorRT Documentation," https://docs.nvidia.com/deeplearning/tensorrt/latest/index.html

[19] Intel, "OpenVINO™ toolkit: An open source AI toolkit that makes it easier to write once, deploy anywhere," https://www.intel.com/content/www/us/en/developer/tools/openvino-toolkit/overview.html

[20] J.H. Lee et al., "MALE: A Multi-Objective Evaluation Method for AI Mobility Services across the Cloud-Edge-Device Continuum," in Proc. IEEE Int. Conf. Syst., Man, Cybern., (Vienna, Austria), Oct. 2025.

[21] 전재호 외, "미래 모빌리티를 위한 소프트웨어 정의형 인프라스트럭쳐 기술," 정보과학회지 제42권 제11호, 2024, pp. 16-23.

[22] Y. Kang et al., "Neurosurgeon: Collaborative Intelligence Between the Cloud and Mobile Edge," ACM SIGARCH Comput. Archit. News, vol. 45, no. 1, 2017, pp. 615-629.

[23] DeepDecision Webstie. https://www.deepdecision.ai/

[24] 테슬라코리아, "풀 셀프 드라이빙 구현 기능(수퍼바이즈드)," 2025. https://www.tesla.com/ko_kr/fsd

[25] Microsoft Learn, "Reference architecture for autonomous vehicle operations (AVOps)," 2025. 4. 1, https://learn.microsoft.com/en-us/industry/mobility/architecture/ra-mobility-avops