CRESSEM
CREATIVE ASSEMBLY

# 오픈모델 기반 온디바이스 AI 반도체 기술 동향 및 시장 전망 보고서

| | |
|---|---|
| **문서번호** | CRSM-AI-2026-AUTO |
| **작성일** | 2026-07-21 |
| **작성** | Cressem AI 시스템 (자동 생성) |
| **보안등급** | 사내 비밀 (Confidential) |
| **버전** | v1.0 |

# 목 차

**오픈모델 기반 온디바이스 AI 반도체 기술 동향 및 시장 전망 보고서 3**

개요: 온디바이스 AI 패러다임의 전환 . . . 3

오픈소스 모델의 온디바이스 최적화 기술 동향 . . . 4

차세대 온디바이스 AI 칩셋 및 아키텍처 분석 . . . 6

■ 사내 자료 기준 보충 . . . 9

시장 전망 및 산업별 적용 사례 . . . 9

결론 및 시사점 . . . 11

참고 자료 . . . 12

# 오픈모델 기반 온디바이스 AI 반도체 기술 동향 및 시장 전망 보고서

2026년 현재, 클라우드 의존도를 낮추고 보안과 저지연성을 강화한 온디바이스 AI(On-Device AI)가 반도체 산업의 핵심 동력으로 부상하고 있습니다. 본 보고서는 Llama, Qwen 등 오픈소스 LLM의 온디바이스 최적화 기술과 이를 지원하는 차세대 NPU/AI 칩셋의 기술적 진화, 그리고 이에 따른 반도체 패키징 및 검사 기술의 변화를 심층 분석합니다.

## 개요: 온디바이스 AI 패러다임의 전환

### 1. 클라우드 중심 AI에서 온디바이스 AI로의 구조적 전환

지난 수년간 인공지능(AI) 산업을 주도해 온 패러다임은 거대 모델을 클라우드 서버에 구축하고, 사용자의 요청을 네트워크를 통해 처리하는 '클라우드 AI(Cloud-based AI)' 방식이었습니다. 그러나 2026년 현재, AI 기술의 중심축은 데이터 센터를 넘어 사용자의 손안에 있는 단말기(End-device)로 급격히 이동하고 있습니다. 이러한 변화를 상징하는 핵심 개념이 바로 **온디바이스 AI(On-Device AI)**입니다.

클라우드 AI는 강력한 연산 자원을 바탕으로 초거대 언어 모델(LLM)을 구동할 수 있다는 장점이 있으나, 데이터 전송 과정에서 발생하는 **지연 시간(Latency)**, 개인정보 유출에 대한 **보안 우려(Privacy & Security)**, 그리고 지속적인 서버 운영에 따른 **막대한 네트워크 비용 및 에너지 소모**라는 태생적 한계를 지니고 있습니다.

반면, 온디바이스 AI는 기기 자체에 탑재된 전용 가속기(NPU, Neural Processing Unit)를 활용하여 데이터를 로컬에서 직접 처리합니다. 이는 다음과 같은 세 가지 측면에서 혁신적인 가치를 제공합니다.

- **초저지연성(Low Latency):** 네트워크 연결 상태와 관계없이 실시간 추론(Inference)이 가능하여, 자율주행, 실시간 음성 통역, AR/VR과 같은 즉각적인 반응이 필요한 서비스의 품질을 결정짓는 핵심 요소가 됩니다.
- **데이터 프라이버시(Data Privacy):** 민감한 개인 정보나 기업의 기밀 데이터가 외부 서버로 전송되지 않고 기기 내부에서만 처리되므로, 보안 요구 수준이 높은 엔터프라이즈 및 개인용 서비스에서 필수적인 요건으로 자리 잡았습니다.
- **비용 및 에너지 효율성(Cost & Energy Efficiency):** 클라우드 서버로의 데이터 트래픽을 최소화함으로써 통신 비용을 절감하고, 대규모 데이터 센터의 전력 소모 문제를 완화하는 대안으로 주목받고 있습니다. [출처: lgbr.co.kr]

### 2. 2026년 온디바이스 AI의 시장 위치 및 기술적 성숙도

2026년 현재, 온디바이스 AI는 단순한 기술적 트렌드를 넘어 하드웨어와 소프트웨어가 결합된 하나의 완성된 생태계로 진화하였습니다. 과거에는 특정 기능을 수행하는 '엣지 AI(Edge AI)' 수준에 머물렀다면, 현재는 기기 내에서 멀티모달(Multimodal) 추론과 복잡한 플래닝(Planning)이 가능한 수준까지 도달했습니다. [출처: rownie01.com]

특히 2026년은 온디바이스 AI가 틈새 시장(Niche Market)을 벗어나 주류 시장(Mainstream)으로 완전히 편입된 원년으로 평가됩니다. Apple Intelligence와 같은 고도화된 AI OS가 모바일 환경에 완전히 정착하였으며, 퀄컴(Qualcomm)의 스냅드래곤 X Elite와 같이 45 TOPS(Tera Operations Per Second) 이상의 성능을 갖춘 NPU가 탑재된 AI PC가 보편화되었습니다. [출처: knightk.tistory.com]

이러한 시장의 성숙은 하드웨어 아키텍처의 변화와 맞물려 있습니다. 단순히 CPU/GPU의 성능을 높이는 방식에서 벗어나, AI 연산에 최적화된 전용 NPU 아키텍처와 이를 뒷받침하는 고대역폭 메모리(HBM), 그리고 칩렛(Chiplet) 기반의 이종 집적(Heterogeneous Integration) 기술이 반도체 설계의 핵심으로 부상하였습니다. [출처:

mckinsey-electronics.com]

### 3. 온디바이스 AI 확산에 따른 반도체 산업의 변화 요약

온디바이스 AI 패러다임의 전환은 반도체 가치 사슬 전반에 걸쳐 새로운 요구사항을 창출하고 있습니다. 아래 표는 클라우드 AI 시대와 온디바이스 AI 시대의 주요 기술적 요구사항 차이를 비교한 것입니다.

| 구분 | 클라우드 AI (Cloud AI) | 온디바이스 AI (On-Device AI) |
|---|---|---|
| **주요 연산 장치** | GPU, TPU, 고성능 가속기 | **NPU (Neural Processing Unit)**, Mobile GPU |
| **핵심 성능 지표** | FLOPS (연산 성능), 대규모 메모리 용량 | **TOPS/W (전력 대비 성능)**, 저지연성 |
| **데이터 처리 방식** | 중앙 집중식 (Centralized) | **분산/로컬 처리 (Distributed/Local)** |
| **주요 과제** | 서버 인프라 비용, 전력 효율 | **모델 경량화, 온디바이스 최적화, 발열 제어** |
| **연결성 요구** | 고대역폭 네트워크 (High Bandwidth) | **끊김 없는 연결성(Seamless Connectivity)** |

표 1. 온디바이스 AI 확산에 따른 반도체 산업의 변화 요약

[출처: mdia.tech.informa.com, neuralcoretech.com 종합 분석]

결론적으로, 2026년의 온디바이스 AI는 단순한 '기기 내 실행'을 넘어, 개인화된 지능(Personalized Intelligence)을 구현하는 핵심 인프라로 기능하고 있습니다. 이는 반도체 설계 단계부터 소프트웨어 프레임워크, 그리고 이를 구현하기 위한 초미세 패키징 기술에 이르기까지 산업 전반의 기술적 난이도를 상향 평준화시키며 새로운 성장 동력을 제공하고 있습니다.

## 오픈소스 모델의 온디바이스 최적화 기술 동향

최근 온디바이스 AI(On-Device AI) 시장의 급격한 성장은 클라우드 기반의 거대 모델(LLM)을 엔드 디바이스(End-device)의 제한된 컴퓨팅 자원 내에서 구현하려는 시도와 맞물려 있습니다. 특히 Llama, Qwen, Gemma와 같은 고성능 오픈소스 모델(Open-source LLM)의 등장은 특정 기업의 폐쇄적 생태계를 넘어, 다양한 하드웨어 플랫폼에서 범용적으로 동작할 수 있는 AI 생태계를 구축하는 촉매제가 되었습니다. 2026년 현재, 이러한 오픈소스 모델을 모바일, 웨어러블, 엣지 컴퓨팅 기기에서 실시간으로 구동하기 위한 핵심 기술은 단순히 모델의 크기를 줄이는 것을 넘어, 하드웨어 가속기(NPU/GPU)와의 정합성을 극대화하는 '최적화(Optimization)' 단계로 진입하였습니다.

### 1. 모델 경량화 및 압축 기술 (Model Compression & Quantization)

오픈소스 모델을 온디바이스 환경에 이식하기 위한 가장 근본적인 기술은 모델의 파라미터(Parameter) 수를 줄이거나 데이터의 정밀도를 낮추어 메모리 점유율과 연산량을 감소시키는 것입니다.

**양자화(Quantization)** 기술은 현재 가장 활발하게 적용되는 분야입니다. 기존의 FP32(32-bit Floating Point) 정밀도로 학습된 모델을 INT8, INT4, 심지어는 최근 연구되는 1.58-bit(Binary/Ternary) 수준까지 낮춤으로써 모델 크기를 획기적으로 줄입니다. 양자화는 크게 두 가지 방식으로 구분됩니다.

- **PTQ (Post-Training Quantization):** 학습이 완료된 모델의 가중치를 사후에 변환하는 방식입니다. 추가적인 학습 비용이 거의 들지 않아 효율적이지만, 정밀도가 급격히 낮아질 경우 모델의 추론 성능(Perplexity)이 저하될

수 있습니다.

- **QAT (Quantization-Aware Training):** 모델 학습 단계에서 양자화로 인한 오차를 반영하여 학습하는 방식입니다. PTQ보다 높은 정밀도를 유지할 수 있으나, 막대한 컴퓨팅 자원과 학습 데이터가 필요하다는 단점이 있습니다.

**가지치기(Pruning)** 기술은 모델의 성능에 기여도가 낮은 뉴런이나 연결(Weight)을 제거하여 희소성(Sparsity)을 확보하는 기술입니다. 이를 통해 모델의 연산량을 줄이고 메모리 대역폭(Memory Bandwidth) 요구사항을 완화할 수 있습니다. 또한, **지식 증류(Knowledge Distillation)** 기법을 통해 거대한 Teacher 모델의 지식을 상대적으로 작은 Student 모델로 전이시켜, 경량 모델임에도 불구하고 높은 수준의 추론 능력을 유지하도록 설계합니다.

### 2. 오픈소스 프레임워크 및 실행 엔진의 발전

경량화된 모델이 실제 하드웨어에서 효율적으로 구동되기 위해서는 모델의 연산 그래프를 하드웨어 아키텍처에 최적화하여 컴파일하는 소프트웨어 스택이 필수적입니다.

최근에는 특정 OS나 하드웨어에 종속되지 않고 다양한 환경에서 오픈소스 LLM을 구동할 수 있는 SDK 및 라이브러리가 주목받고 있습니다. 대표적인 사례로 **RunAnywhere**와 같은 오픈소스 SDK는 iOS, Android 및 크로스 플랫폼 애플리케이션에서 LLM과 멀티모달 모델을 구동할 수 있는 환경을 제공합니다 [출처: GitHub - RunAnywhere]. 또한, **Off Grid** 프로젝트는 Llama, Qwen, Gemma, Phi, DeepSeek 등 최신 오픈소스 모델을 `llama.cpp` 기반의 기술을 활용하여 모바일 기기에서 완전히 오프라인(On-device) 상태로 실행할 수 있도록 지원하며, 음성(Whisper.cpp) 및 비전 기능을 통합하는 추세입니다 [출처: GitHub - Off Grid].

이러한 소프트웨어 스택은 다음과 같은 핵심 기능을 수행합니다:

- **커널 최적화(Kernel Optimization):** NPU의 특정 연산 유닛에 최적화된 연산 커널을 생성하여 연산 효율을 극대화합니다.
- **메모리 관리(Memory Management):** 제한된 RAM 환경에서 모델의 레이어(Layer)를 순차적으로 로드하거나, KV 캐시(Key-Value Cache)를 효율적으로 관리하여 메모리 부족(OOM, Out-of-Memory) 문제를 방지합니다.

### 3. 오픈소스 모델 기반 최적화 기술 비교 분석

현재 시장에서 주로 활용되는 오픈소스 모델과 이를 온디바이스로 구현하기 위한 기술적 접근 방식을 비교하면 다음과 같습니다.

| 구분 | Llama 시리즈 (Meta) | Qwen 시리즈 (Alibaba) | Gemma 시리즈 (Google) |
|---|---|---|---|
| **주요 특징** | 가장 방대한 생태계 및 검증된 성능 | 다국어(특히 아시아권) 및 코딩 성능 우수 | Google의 기술력을 바탕한 효율적 아키텍처 |
| **온디바이스 전략** | 높은 파라미터 효율성 기반의 양자화 적용 | 멀티모달 및 경량화 모델(0.5B~7B) 중심 | TPU 및 모바일 가속기 최적화 지향 |
| **주요 최적화 기술** | GGUF, AWQ, GPTQ 기반 양자화 | FP8/INT4 정밀도 최적화 | MediaPipe 및 TensorFlow Lite 통합 |
| **적용 타겟** | 범용 AI 에이전트, 개인 비서 | 다국어 지원 모바일 기기, Edge AI | 안드로이드 생태계 및 웨어러블 |

표 2. 오픈소스 모델 기반 최적화 기술 비교 분석

### 4. 기술적 과제 및 향후 전망

오픈소스 모델의 온디바이스 최적화는 비약적인 발전을 이루었으나, 여전히 해결해야 할 기술적 난제가 존재합니다.

첫째, **추론 지연 시간(Latency)과 에너지 효율(Power Efficiency)의 트레이드오프(Trade-off)** 문제입니다. 모델을 극단적으로 압축할 경우 연산량은 줄어들지만, 압축된 데이터를 다시 복원하는 과정에서 메모리 접근 빈도가 높아져 오히려 전력 소모가 증가하거나 지연 시간이 늘어나는 현상이 발생할 수 있습니다.

둘째, **멀티모달(Multimodal) 통합의 복잡성**입니다. 텍스트뿐만 아니라 이미지, 음성을 동시에 처리하는 멀티모달 모델은 데이터의 차원이 높아 메모리 요구량이 기하급수적으로 증가합니다. 이를 해결하기 위해 비전-언어 모델(VLM)을 위한 전용 양자화 알고리즘과 하드웨어 가속 기술이 요구됩니다.

셋째, **모델-하드웨어 정합성(Hardware-Software Co-design)**입니다. 소프트웨어 알고리즘이 발전하더라도 이를 뒷받침할 NPU의 아키텍처가 해당 연산(예: Sparse Matrix Multiplication)을 지원하지 못하면 성능 향상은 제한적입니다. 따라서 향후에는 모델의 구조 자체를 특정 NPU의 연산 특성에 맞추어 설계하는 '하드웨어 인지형 모델 설계(Hardware-aware Model Design)'가 핵심 경쟁력이 될 것으로 전망됩니다.

결론적으로, 2026년 현재의 온디바이스 AI 기술은 오픈소스 모델의 유연성과 고도화된 양자화/압축 기술, 그리고 이를 하드웨어 레벨에서 뒷받침하는 최적화 SDK의 결합을 통해 완성되어 가고 있습니다. 이는 클라우드 의존도를 낮추고 보안성과 실시간성을 확보한 진정한 의미의 '개인용 AI(Personal AI)' 시대를 가속화하고 있습니다.

## 차세대 온디바이스 AI 칩셋 및 아키텍처 분석

2026년 현재, 온디바이스 AI(On-Device AI) 시장은 단순히 클라우드 AI의 기능을 모바일로 옮겨오는 단계를 넘어, 기기 자체에서 고성능 대규모 언어 모델(LLM)과 멀티모달 모델을 실시간으로 추론(Inference)할 수 있는 하드웨어의 시대로 완전히 진입하였습니다 [출처: knightk.tistory.com]. 이러한 패러다임 전환의 핵심은 기존의 범용 프로세서 구조에서 벗어나, AI 연산에 특화된 **NPU(Neural Processing Unit)**를 중심으로 한 **이종 컴퓨팅(Heterogeneous Computing)** 아키텍처의 고도화에 있습니다.

### 1. 주요 칩셋별 NPU 성능 및 아키텍처 비교 분석

현재 온디바이스 AI 시장을 주도하고 있는 Apple, Qualcomm, NVIDIA, AMD 등의 칩셋은 각기 다른 설계 철학을 바탕으로 AI 가속기(AI Accelerator) 성능을 극대화하고 있습니다. 특히 연산 성능을 나타내는 **TOPS(Tera Operations Per Second)** 수치는 칩셋의 경쟁력을 결정짓는 핵심 지표로 자리 잡았습니다 [출처: addthinking.com].

| 구분 | Apple A17 Pro / 차세대 시리즈 | Qualcomm Snapdragon X Elite | NVIDIA/AMD Edge AI Chipsets |
|---|---|---|---|
| **핵심 아키텍처** | Apple Intelligence 통합 설계 | Hexagon NPU 중심 구조 | GPU/NPU 하이브리드 구조 |
| **AI 성능 (NPU)** | Neural Engine 기반 최적화 | 최대 45 TOPS 이상 지원 | 고성능 가속기 탑재 |
| **주요 특징** | 하드웨어-소프트웨어 수직 계열화 | PC 및 플래그십 모바일 범용성 | 개발자 친화적 프레임워크 지원 |
| **주요 타겟** | iPhone, iPad (Apple Intelligence) | Windows on ARM, 플래그십 노트북 | Edge Computing, AI PC, 로보틱스 |

표 3. 주요 칩셋별 NPU 성능 및 아키텍처 비교 분석

**Apple**의 경우, 하드웨어와 소프트웨어(iOS/macOS)의 강력한 통합을 통해 **Apple Intelligence**를 구현하고 있습니다. Apple의 NPU는 단순한 연산 가속을 넘어, 시스템 전체의 전력 효율을 고려한 설계가 특징이며, 모델의 가중치(Weights)를 효율적으로 관리하여 저전력 상태에서도 지속적인 AI 추론이 가능하도록 설계되었습니다 [출처: knightk.tistory.com].

**Qualcomm**은 **Snapdragon X Elite**를 통해 노트북 및 PC 시장의 AI 전환을 주도하고 있습니다. 이 칩셋은 45 TOPS 이상의 강력한 NPU 성능을 제공하며, 이는 Microsoft의 Copilot+ PC 요구사항을 충족하는 수준입니다 [출처: knightk.tistory.com]. Qualcomm의 아키텍처는 CPU, GPU, NPU가 유기적으로 협력하는 이종 컴퓨팅 구조를 통해, 대규모 모델의 로딩과 실시간 텍스트/이미지 생성 작업 시 병목 현상을 최소화합니다.

## 2. 이종 컴퓨팅(Heterogeneous Computing)과 AI 가속기 구조

차세대 온디바이스 AI 칩셋의 핵심은 **이종 컴퓨팅(Heterogeneous Computing)** 아키텍처입니다. 과거에는 CPU가 모든 연산을 처리하거나 GPU가 보조하는 형태였다면, 현재는 특정 연산(예: 행렬 곱셈, 활성화 함수)에 최적화된 **AI 가속기(AI Accelerator)**가 독립적인 영역을 차지합니다.

- **NPU(Neural Processing Unit)의 역할:** 신경망의 핵심 연산인 대규모 행렬 연산(Matrix Multiplication)을 저전력으로 처리합니다. 특히 온디바이스 환경에서는 메모리 대역폭(Memory Bandwidth) 제한이 큰 병목이 되므로, NPU 내부에 데이터 재사용성을 높이는 로컬 메모리 구조를 갖추는 것이 기술적 관건입니다.
- **메모리 계층 구조의 진화:** LLM과 같은 대형 모델을 온디바이스에서 구동하기 위해서는 높은 대역폭의 메모리가 필수적입니다. 이에 따라 칩셋 설계 단계에서부터 고대역폭 메모리(HBM)의 통합이나, 칩과 메모리를 하나로 묶는 2.5D/3D 패키징 기술이 아키텍처 설계의 핵심 요소로 고려되고 있습니다 [출처: mckinsey-electronics.com].
- **데이터 흐름(Dataflow) 최적화:** 모델의 레이어(Layer)가 넘어갈 때 발생하는 데이터 이동을 최소화하기 위해, 연산 유닛(PE, Processing Element) 간의 직접적인 데이터 통신 경로를 설계하는 것이 최신 트렌드입니다.

## 3. 온디바이스 AI 구현을 위한 기술적 과제: 전력과 성능의 트레이드오프

온디바이스 환경은 클라우드와 달리 전력 소모(Power Consumption)와 발열(Thermal) 문제가 매우 엄격하게 제한됩니다. 따라서 칩셋 설계자들은 다음과 같은 기술적 난제에 직면해 있습니다.

첫째, **연산 정밀도(Precision)의 최적화**입니다. FP32(32비트 부동 소수점) 대신 INT8, FP8, 심지어는 4비트 이하의 양자화(Quantization) 기술을 하드웨어 수준에서 지원함으로써, 정확도 손실을 최소화하면서도 연산량과 메모리 점유율을 획기적으로 낮추는 설계가 요구됩니다 [출처: arxiv.org].

둘째, **모델-하드웨어 정합성(Hardware-Model Alignment)**입니다. 아무리 강력한 NPU를 탑재하더라도, 오픈소스 모델(Llama, Qwen 등)의 연산 패턴이 하드웨어 가속 구조와 맞지 않으면 성능이 저하됩니다. 따라서 최신 칩셋들은 다양한 오픈소스 모델의 커널(Kernel)을 하드웨어 가속기로 직접 컴파일할 수 있는 전용 SDK와 컴파일러 기술을 함께 제공하는 추세입니다 [출처: github.com].

## 차세대 온디바이스 AI 칩셋 아키텍처 구성

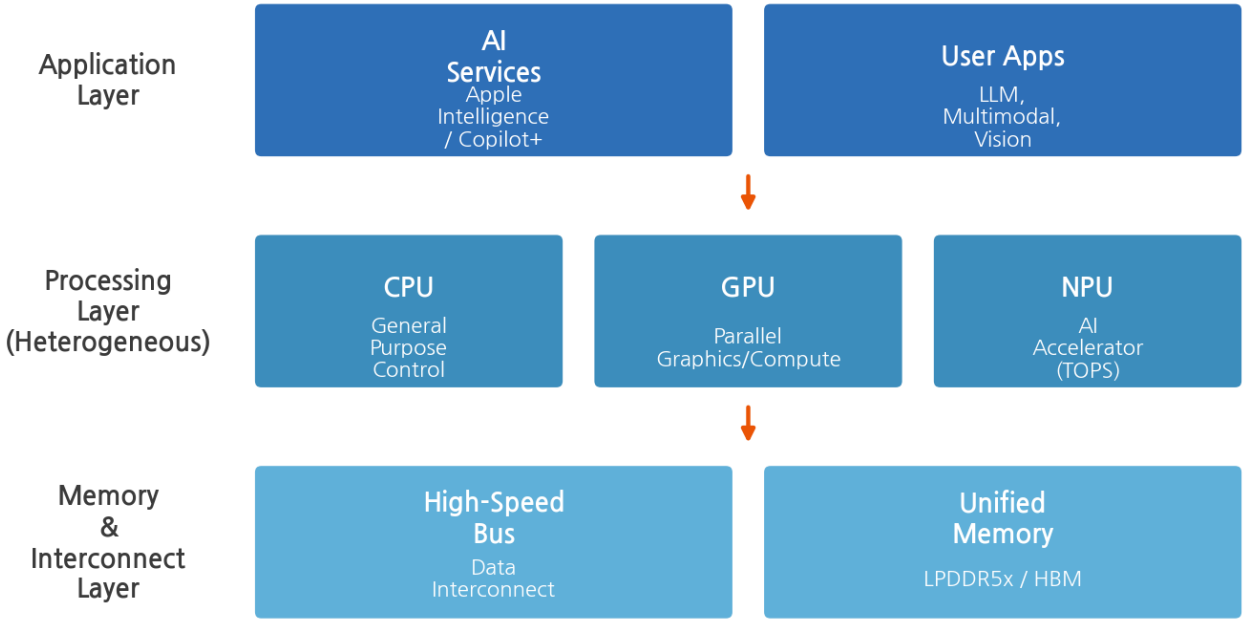


그림 1. 차세대 온디바이스 AI 칩셋 아키텍처 구성

## 온디바이스 AI 통합 스택 최적화 프로세스


그림 2. 온디바이스 AI 통합 스택 최적화 프로세스

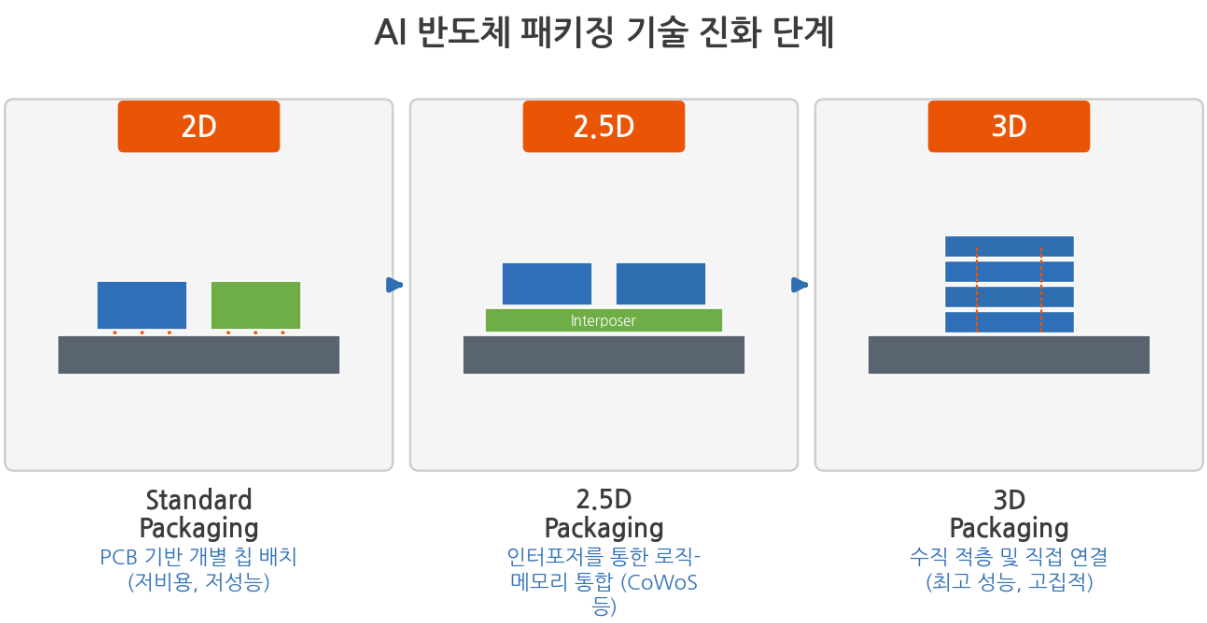


그림 3. AI 반도체 패키징 기술 진화 단계


그림 4. AI 반도체 제조 및 검사 공정 흐름

### ■ 사내 자료 기준 보충

사내 문서에는 크레셈(CRESSEM)의 기술 역량과 관련된 별개의 자료가 있습니다. [1, 2] 사내 자료에 따르면, 크레셈은 HBM 및 FC-BGA 등 고난도 패키징 검사 분야에서 기술적 진입장벽을 확보하고 있으며, Deep Learning 기반 결함 분류를 통해 과검(Over-kill)률을 낮추는 AI 비전 솔루션과 생산 주기를 극대화하는 고속 광학 엔진 기술을 보유하고 있습니다. 또한, 검사 데이터를 분석하여 공정 불량의 원인을 역추적(Root Cause Analysis)함으로써 고객사의 수율(Yield) 향상에 기여하는 스마트 팩토리 솔루션을 지향하고 있습니다.

## 시장 전망 및 산업별 적용 사례

2026년 현재, 온디바이스 AI(On-Device AI)는 단순한 기술적 트렌드를 넘어 소비자 가전(Consumer Electronics) 시장의 근본적인 패러다임을 재편하고 있습니다. 과거 클라우드 의존형 AI 모델이 중앙 집중식 연산에 치중했다면, 현재는 개인화된 데이터 보안과 초저지연(Low-latency) 성능을 확보하기 위해 엔드 디바이스(End

Device) 자체에서 추론을 수행하는 엣지 AI(Edge AI) 하드웨어 시장이 급격히 팽창하고 있습니다 [출처: neuralcoretech.com]. 특히 오픈소스 기반의 경량화된 대규모 언어 모델(LLM)들이 다양한 하드웨어 플랫폼에 이식됨에 따라, 기기별 AI 침투율(Market Penetration)은 기하급수적인 상승 곡선을 그리고 있습니다 [출처: omdia.tech.informa.com].

### 1. 엔드 디바이스별 AI 침투율 및 시장 동향

온디바이스 AI의 확산은 모바일, PC, 웨어러블 기기라는 세 가지 핵심 축을 중심으로 전개되고 있으며, 각 세그먼트별로 기술적 성숙도와 침투 양상이 다르게 나타납니다.

| 구분 | 주요 타겟 시장 | AI 침투율 특성 | 핵심 동력 및 기술 요소 |
|---|---|---|---|
| **모바일 (Mobile)** | 스마트폰, 태블릿 | **성숙기 진입** | NPU(Neural Processing Unit) 내장 칩셋, 온디바이스 LLM, 개인화 비서 |
| **PC (Personal Computer)** | 노트북, 데스크탑 | **급격한 성장기** | Qualcomm Snapdragon X Elite 등 고성능 NPU, 로컬 AI 워크스테이션 |
| **웨어러블 (Wearable)** | 스마트워치, AR/VR | **도입/확장기** | 초저전력(Ultra-low power) 추론, 센서 퓨전, 실시간 음성/시각 인식 |

표 4. 엔드 디바이스별 AI 침투율 및 시장 동향

#### 가. 모바일 및 스마트 디바이스 (Mobile & Smart Devices)

모바일 시장은 온디바이스 AI가 가장 먼저 대중화된 영역입니다. 2026년 현재, 플래그십 스마트폰의 대부분은 Apple Intelligence와 같은 고도화된 온디바이스 AI 기능을 기본 탑재하고 있습니다 [출처: knightk.tistory.com]. 이는 단순히 텍스트 생성을 넘어, 기기 내부에 저장된 개인 데이터를 바탕으로 한 맥락 이해(Contextual Awareness)와 멀티모달(Multimodal) 추론을 가능케 합니다. 특히, 클라우드 연결 없이도 작동하는 온디바이스 언어 모델(On-Device Language Models)의 발전은 개인정보 보호(Privacy)와 데이터 보안을 중시하는 소비자 요구를 충족시키며 시장 점유율을 견인하고 있습니다 [출처: arxiv.org].

#### 나. AI PC 및 컴퓨팅 (AI PC & Computing)

PC 시장은 2025년 말부터 본격화된 'AI PC' 전환기를 지나, 2026년 현재 하드웨어 교체 수요가 폭발하는 단계에 있습니다. 퀄컴(Qualcomm)의 Snapdragon X Elite와 같이 45 TOPS(Tera Operations Per Second) 이상의 강력한 NPU 성능을 갖춘 칩셋이 표준으로 자리 잡으면서, 노트북은 단순한 사무 도구를 넘어 로컬에서 대규모 모델을 구동할 수 있는 AI 워크스테이션 역할을 수행하게 되었습니다 [출처: knightk.tistory.com]. 이러한 고성능 컴퓨팅 환경은 개발자 및 전문가들이 클라우드 비용 부담 없이 로컬 환경에서 AI 모델을 테스트하고 배포할 수 있는 생태계를 조성하고 있습니다.

#### 다. 웨어러블 및 엣지 디바이스 (Wearables & Edge Devices)

웨어러블 및 IoT 기기 영역에서는 'Always-on' AI 기술이 핵심입니다. 스마트워치나 AR/VR 글래스와 같은 기기들은 극도로 제한된 배터리 용량과 발열 제약 내에서 실시간 AI 서비스를 제공해야 합니다. 이에 따라 초저전력 엣지 AI 하드웨어의 중요성이 증대되고 있으며, 음성 인식(Whisper.cpp 등 활용) 및 시각 인식(Vision) 기능을 기기 자체에서 수행하는 기술이 상용화 단계에 이르렀습니다 [출처: github.com].

### 2. 산업별 적용 사례 및 비즈니스 임팩트

온디바이스 AI 기술은 단순 소비자 가전을 넘어 산업 전반의 디지털 전환(Digital Transformation)을 가속화하고 있습니다.

- **개인화 서비스 (Personalized Services):** 사용자의 사용 패턴, 위치 정보, 건강 데이터를 기기 내에서 학습하여 실시간으로 최적화된 피드백을 제공합니다. 이는 의료용 웨어러블 기기에서 실시간 질병 징후 포착이나, 개인 맞춤형 교육 도구의 발전으로 이어집니다.
- **스마트 홈 및 IoT (Smart Home & IoT):** 클라우드 서버를 거치지 않는 로컬 제어를 통해 반응 속도를 극대화하고, 네트워크 단절 시에도 가전제품의 지능형 제어가 가능해집니다. 이는 보안이 중요한 홈 네트워크 시스템의 신뢰성을 높이는 핵심 요소입니다.
- **산업용 엣지 컴퓨팅 (Industrial Edge Computing):** 공장 내 센서나 로봇에 탑재된 AI 칩셋은 실시간 결함 탐지 및 예지 보전(Predictive Maintenance)을 수행합니다. 이는 데이터 전송 지연을 최소화하여 생산 라인의 가동률을 높이고, 대규모 데이터를 클라우드로 전송하는 비용을 절감하는 경제적 이점을 제공합니다.

### 3. 향후 시장 전망 및 과제

2026년 이후의 시장은 하드웨어의 성능 상향 평준화와 함께, 소프트웨어-하드웨어 간의 '최적화된 정합성'을 확보한 기업이 주도권을 잡을 것으로 전망됩니다. AI 모델의 크기가 커짐에 따라 이를 효율적으로 압축하여 구동하는 경량화 알고리즘과, 이를 뒷받침할 수 있는 고대역폭 메모리(HBM) 및 차세대 패키징 기술이 결합된 시스템 수준의 설계(Systems-level design)가 시장의 핵심 경쟁력이 될 것입니다 [출처: mckinsey-electronics.com].

결론적으로, 온디바이스 AI 시장은 단순한 기기 성능의 향상을 넘어, '지능형 엔드 디바이스'가 인간의 일상과 산업 현장에 깊숙이 침투하는 거대한 전환점에 서 있습니다. 소비자들은 더 빠르고, 더 안전하며, 더 개인화된 AI 경험을 요구할 것이며, 이를 충족하기 위한 반도체 설계 및 제조 기술의 혁신은 지속될 것입니다.

## 결론 및 시사점

본 보고서에서 분석한 오픈모델 기반 온디바이스 AI(On-Device AI) 기술 동향을 종합하면, 현재 반도체 산업은 클라우드 중심의 거대 모델 시대에서 개별 디바이스가 스스로 추론하고 학습하는 '엣지 인텔리전스(Edge Intelligence)' 시대로 급격히 전환되고 있습니다. 2026년 현재, 오픈소스 모델(Llama, Qwen, Gemma 등)의 경량화 기술과 이를 뒷받침하는 고성능 NPU(Neural Processing Unit) 아키텍처의 결합은 온디바이스 AI 생태계의 성장을 가속화하는 핵심 동력으로 작용하고 있습니다.

오픈모델 기반 온디바이스 AI 생태계의 지속 가능한 성장을 위한 핵심 성공 요인(Key Success Factors)과 향후 기술적 과제는 다음과 같이 요약됩니다.

#### 1. 하드웨어-소프트웨어 통합 최적화(HW-SW Co-optimization)

단순히 고성능 칩셋을 제조하는 것을 넘어, 특정 오픈모델의 가중치(Weight)와 연산 구조에 최적화된 하드웨어 가속기 설계가 필수적입니다. SDK(RunAnywhere, Off Grid 등)를 통한 모델-하드웨어 간의 정합성(Alignment) 확보는 온디바이스 환경에서 제한된 전력(Power Consumption)과 메모리 대역폭(Memory Bandwidth) 내에서 LLM(Large Language Model)을 구동하기 위한 최우선 과제입니다.

#### 2. 패키징 및 검사 기술의 고도화

온디바이스 AI 칩셋의 성능이 상향 평준화됨에 따라, 이를 구현하기 위한 2.5D/3D 패키징 및 칩렛(Chiplet) 기술의 중요성이 증대되고 있습니다. 특히 고성능 연산을 지원하기 위한 HBM(High Bandwidth Memory)의 통합과 미세 공정에서의 신뢰성 확보가 시장의 승패를 결정지을 것입니다. 이에 따라 초미세 간극 측정 및 결함 제어를 위한 고정밀 광학 검사 및 AI 비전 솔루션의 역할은 더욱 막중해질 것으로 전망됩니다.

### 3. 데이터 프라이버시 및 보안(Privacy & Security)

클라우드를 거치지 않는 로컬 추론의 가장 큰 강점인 '데이터 주권'을 완벽히 보장하기 위한 보안 아키텍처 설계가 요구됩니다. 이는 온디바이스 AI가 엔터프라이즈 및 개인용 디바이스 시장에 깊숙이 침투하기 위한 필수적인 신뢰 기반이 될 것입니다.

**[Strategic Implications & Future Roadmap]**

| 구분 | 핵심 전략 방향 (Strategic Direction) | 기술적 로드맵 (Technical Roadmap) |
|---|---|---|
| **반도체 설계/제조** | 이기종 통합 시스템(Heterogeneous Integration) 강화 | 3D IC, Hybrid Bonding 도입 및 칩렛 기반 설계 확대 |
| **소프트웨어/모델** | 모델 경량화 및 양자화(Quantization) 기술 고도화 | 온디바이스 전용 소형 언어 모델(sLLM) 최적화 |
| **검사/품질 관리** | AI 기반 스마트 팩토리 및 실시간 결함 분석 | 고속 광학 엔진 및 데이터 기반 수율(Yield) 역추적 솔루션 |

표 5. 결론 및 시사점

결론적으로, 향후 온디바이스 AI 시장은 개별 칩의 성능 경쟁을 넘어, **'오픈모델의 유연성 + NPU의 효율성 + 첨단 패키징의 집적도 + 검사 기술의 신뢰성'**이 유기적으로 결합된 통합 솔루션 제공 기업들이 주도할 것으로 예측됩니다. 기업들은 급변하는 AI 알고리즘의 변화에 즉각 대응할 수 있는 하드웨어 유연성을 확보하는 동시에, 제조 공정의 수율 극대화를 위한 고도화된 검사 인프라 구축에 집중해야 합니다.

## 참고 자료

1. [PDF 생성형 Ai 열풍으로 성큼 다가온 온디바이스 Ai](https://www.lgbr.co.kr/uploadFiles/ko/pdf/busi/LGBRReport240408202428081628192 03.pdf)

2. [온디바이스 Ai 칩셋 Top 5 완벽 분석: 최신 트렌드와 추천 제품](https://addthinking.com/on-device-ai-chipset-trend-recommendation/)

3. [온디바이스 AI 완벽 가이드 2026 - Apple Intelligence·퀄컴 NPU·로컬 LLM ...](https://knightk.tistory.com/331)

4. [2025년 11~12월 기준 최신 AI 기술 동향: Gemini 3, TPU, 가속기·대모델 ...](https://rownie01.com/ai-trend-2025/)

5. [2026 Semiconductor Industry Outlook | Deloitte Insights](https://www.deloitte.com/us/en/insights/industry/technology/technology-media-telecom-outlooks/semiconductor-industry-outlook.html)

6. [Semiconductor Outlook 2026 | AI Growth & Packaging Shifts](https://www.mckinsey-electronics.com/post/strategic-semiconductor-and-electronic-component-trends-to-shape-2026-market-dynamics-technologica)

7. [PDF 2026 Trends to Watch: Semiconductors](https://omdia.tech.informa.com/-/media/tech/omdia/assetfamily/2025/12/04/2026-trends-to-watch-semiconductors/2026-trends-to-watch-semiconductors-pdf.pdf)

8. [Edge AI Hardware 2026: On-Device Intelligence, Architecture & Chip Comparison ...](https://neuralcoretech.com/edge-ai-hardware-2026-chip-comparison/)

9. [On-Device Language Models: A Comprehensive Review](https://arxiv.org/html/2409.00088v1)

10. [Awesome-On-Device-AI-Systems/README.md at main - GitHub](https://github.com/jeho-lee/Awesome-On-Device-AI-Systems/blob/main/README.md)

11. [GitHub - ztachip/ztachip: Opensource software/hardware platform to build edge AI ...](https://github.com/ztachip/ztachip)

12. LGBRReport240408202428081628192O3.pdf (사내 자료)

13. 2026-trends-to-watch-semiconductors-pdf.pdf (사내 자료)

14. 분석_크레셈CRESSEM20260509001.pdf (사내 자료)