AI "굿바이 클라우드"



게티이미지뱅크

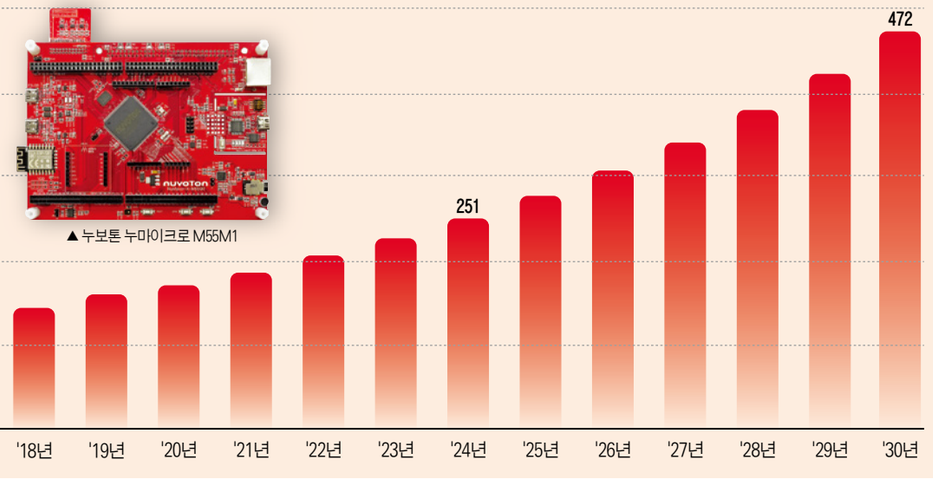
왜 엣지 AI인가?

- 네트워크 끊김·지연 등 걱정 없이 의사결정
- 개인정보 등 민감한 데이터 유출 걱정 차단
- 전력소모 적어 배터리 수명 연장·비용 절감

글로벌 MCU 경쟁

- 스마트폰·가전·전장 등 엣지 AI 활용폭 확대
- 핵심 반도체 MCU 시장도 연 18% 고속 성장
- 美·日·EU 등 앞다퉈 관련법 수립·기술 개발

마이크로컨트롤러(MCU) 시장 규모 (단위: 억 달러), 자료: 옴디아



▲ 누보톤 누마이크로 M55M1

AI는 한동안 데이터센터와 클라우드의 전유물로 여겨졌다. 고성능 GPU와 전력 집약적 연산 자원이 필수라는 인식이 지배적이었지만, IoT 기기 확산과 네트워크 지연 문제, 초저전력·실시간 응답 요구가 커지면서 무게중심은 빠르게 '엣지'로 이동하고 있다.

엣지 AI는 데이터를 서버로 보내지 않고 단말에서 직접 추론해 서버 부하와 통신 비용을 낮추고 개인정보 유출 우려를 줄인다. 특히 웨어러블, 스마트 가전, 자동차 전장처럼 지연에 민감한 영역에서 클라우드 의존도 축소가 뚜렷해지면서 패러다임은 'AI=온디바이스'로 재정의되고 있다.

엣지 AI의 촉매는 마이크로컨트롤러(MCU)다. 수십 년 동안 MCU는 가전과 자동차, 산업용 제어의 '숨은 두뇌' 역할을 맡았지만 복잡 연산에는 한계가 있었다. 최근 초경량 신경망과 전용 AI 가속기(NPU)가 MCU에 집적되면서 단순 제어칩이 AI 추론까지 담당하는 초소형 머신러닝(TinyML) 시대가 열렸다. 불과 몇 년 전만 해도 비현실적으로 여겨지던 '저전력 제어칩 위의 AI'가 산업 표준으로 자리 잡아가는 중이다. MCU 기반 AI는 IoT 기기의 폭발적 증가, 지연이 곧 안전 문제로 이어지는 실시간성 요구, 배터리 구동 기기의 초저전력 한계 등으로 필요성이 커지고 있다.

글로벌 시장조사기 옴디아에 따르면 전 세계 MCU 시장은 2024년 기준 약 200억 달러 규모, 이 중 AI 기능을 탑재한 엣지 MCU는 연평균 18% 성장할 것으로 전망된다. 특히 웨어러블과 스마트홈 분야가 성장을 견인해 2030년에는 전체 MCU 출하량의 30% 이상이 AI 기능을 내장할 것이라는 분석이 나온다. 데이터센터 중심의 AI에서 엣지로의 이동이 단순한 기술적 변화가 아니라 거대한 시장 변화를 수반한다는 의미다.

전통적으로 AI는 시스템 온 칩(SoC)에서 수행됐다. 중앙처리장치(CPU)와 그래픽처리장치(GPU), 디지털 신호처리장치(DSP)를 통합한 SoC는 성능은 높지만 전력·비용 부담이 컸고, 마이크로컨트롤러(MCU)는 저전력·저비용이 강점이지만 자원 제약이 한계로 작용했다. 그러나 NPU를 내재화한 MCU는 CNN 같은 신경망 구조를 압축·양자화 기법과 함께 구동하며 두 영역의 경계를 허물고 있다.

이 흐름을 뒷받침하는 것은 소프트웨어 기술의 발전이다. 신경망 모델에서 중요도가 낮은 연결을 잘라내 구조를 단순화하는 프루닝, 32비트 연산을 8비트 정수로 줄여 모델 크기와 연산량을 크게 낮추는 양자화 같은 기법이 대표적이다. 여기에 텐서플로우 라이트 마이크로(TensorFlow Lite Micro), ST 엣지 AI 스위트(ST Edge AI Suite) 등 개발 도구가 모델 설계부터 변환, MCU 탑재까지 전 과정을 지원하면서 손쉽게 엣지 AI 개발에 접근할 수 있는 환경이 마련되고 있다.

전력 소모에서도 차이는 뚜렷하다. 최신 모바일 SoC가 AI 추론 시 수백 ㎽를 소비하는 데 비해, NPU가 내장된 MCU는 약 10㎽ 수준에서 연산을 처리한다. 배터리 수명이 중요한 웨어러블이나 센서 기기에서 경쟁력이 되는 이유다.

글로벌 MCU 업체들도 잇따라 대응에 나서고 있다. 유럽 최대 반도체 기업 중 하나인 ST마이크로일렉트로닉스는 전용 AI 가속기 '뉴럴-ART 가속기(Neural-ART Accelerator)'를 내장한 STM32N6를 선보여 MCU 단독으로 비전 AI를 구현했다.

STM32N6은 엣지 AI 구현을 앞당기는 고성능 STM32 마이크로컨트롤러다. Arm Cortex-M55 코어와 ST의 독자적인 NPU 뉴럴-ART 가속기를 탑재, 기존 STM32 시리즈 대비 머신러닝 성능을 최대 600배 향상시켰다. 이를 기반으로 컴퓨터 비전, 오디오 프로세싱, 사운드 분석 등 다양한 기능을 지원하며, 비용과 전력 소모에 민감한 컨슈머 및 산업용 애플리케이션을 엣지단에서 직접 실행할 수 있다.

이 같은 변화의 핵심에는 전용 AI 가속기가 있다. 특히 NPU(신경망처리장치)는 곱셈·덧셈 연산을 동시에 수백 개씩 처리할 수 있는 구조로 설계돼 영상이나 센서 데이터를 분석하는 CNN(합성곱 신경망) 같은 모델을 효율적으로 구동한다. ST마이크로일렉트로닉스의 MCU STM32N6는 이러한 구조를 적용해 수백 개의 연산 유닛을 병렬로 구동, 카메라 입력 이미지를 수ms(밀리초·1ms는 1000분의 1초) 단위로 처리할 수 있다.

엣지 AI 수요 확대는 MCU 설계에도 변화를 불러오고 있다. 기존 MCU는 단일 코어 중심 구조였지만, 최근에는 듀얼 코어나 전용 DSP 블록을 함께 집적하는 사례가 늘고 있다. 일부 벤더는 3D 패키징 기법을 적용해 AI 연산 블록과 메모리를 근접 배치, 지연을 최소화하고 있다. 이는 결국 SoC 못지않은 아키텍처 혁신을 MCU 영역에서도 촉발하는 계기가 된다.

또 다른 움직임도 주목된다. 대만 종합 반도체 기업 누보톤(Nuvoton)은 엣지 AI MCU 시장에 새롭게 합류하며, 경량화된 AI 모델을 온디바이스에서 직접 처리할 수 있는 제품군을 선보이고 있다. 특히 AI 개발 키트와 자체 툴체인을 함께 제공해 스타트업과 중소기업이 손쉽게 AI MCU를 활용할 수 있도록 지원하는 전략을 택했다.

누보톤이 최근 공개한 차세대 AI MCU '누마이크로 M55M1'은 연산 성능이 110GOPS(Giga Operations Per Second)에 달한다.

GOPS는 초당 수행 가능한 연산 횟수를 뜻하며, 1GOPS는 1초에 10억번의 연산을 처리하는 수준이다. 이는 220MHz의 Cortex-M55 CPU와 Ethos-U55 NPU를 기반으로 한다. 기존 고성능 MCU가 1GHz에서 약 1GOPS 수준의 AI 연산 능력을 보이는 것과 비교하면 획기적인 성능 개선을 이룬 셈이다.

또한 내장된 1.5MB SRAM 및 2MB 플래시 메모리와 함께 외부 확장 메모리 지원을 통해 다양한 AI 모델을 지원할 수 있다. 외장형 방식은 부팅 시간이 길고 중간 과정에서도 지연이 잦다는 한계가 있다. 그러나 최신 엣지 AI MCU 솔루션은 기존 MCU 환경과 유사한 사용성을 제공해 이러한 문제를 해결할 수 있다.

특히 Arm 에코시스템을 기반으로 설계돼 개발자들이 기존 MCU 개발 경험을 그대로 이어갈 수 있다. 복잡한 AI 모델 배치 과정도 기존 MCU 개발 환경처럼 단순하게 구현할 수 있도록 최적화된 것이다. 즉, 전문적인 SoC나 서버급 환경에서 필요로 하던 복잡한 절차를 줄이고 센서 연결과 펌웨어 업로드 수준의 친숙한 방식으로 AI 모델을 탑재할 수 있도록 한 것이다.

이외에도 6단계의 저전력 전력관리 모드와 음성 활동 감지(VAD), 스마트 모션 감지, Arm 트러스트존(TrustZone) 기반 보안 기능 등을 갖추고 있어 AI 성능과 전력 효율 및 보안성을 동시에 충족할 수 있는 업계 유일의 MCU 솔루션으로 평가된다.

각국 정부와 표준화 기구도 엣지 AI 확산을 지원하고 있다. 유럽연합(EU)은 일반개인정보보호법(GDPR)을 통해 개인정보를 외부 서버로 전송하지 않는 온디바이스 AI 방식을 장려, 미국과 일본은 IoT 보안 표준을 강화해 MCU 단의 실시간 분석을 의무화하는 흐름을 보이고 있다.

미국 전기전자공학회(IEEE)와 유럽전기통신표준협회(ETSI) 등 국제 표준화 기구 역시 초소형 머신러닝(TinyML) 워킹그룹을 중심으로 모델 경량화와 호환성 확보 방안을 논의 중이다. 정책과 규제가 맞물리며 엣지 AI MCU 수요 확대를 제도적으로 뒷받침하고 있는 셈이다.

툴체인과 개발자 경험(UX)도 시장 확산의 핵심 변수다. ST마이크로일렉트로닉스의 '나노엣지 AI 스튜디오(NanoEdge AI Studio)'와 같이 머신러닝 라이브러리를 자동 생성하는 도구는 GUI(그래픽 사용자 인터페이스) 기반으로 센서 데이터를 불러와 몇번의 클릭만으로 경량 모델을 구현, MCU 활용의 진입 장벽을 크게 낮춘다.

결국 글로벌 경쟁의 향방은 성능·전력 효율, 그리고 소프트웨어 생태계의 완성도에 의해 갈릴 공산이 크다. ST는 폐쇄형 최적화 전략, NXP와 르네사스는 Arm 기반 개방형 전략, Ambiq는 초저전력 특화, 에스프레시프는 커뮤니티 친화성을 앞세워 각각 차별화를 시도하고 있다.

산업별 수요 특성도 뚜렷하다. 제조업 현장에서 활용되는 산업용 IoT 기기는 신뢰성과 장기 공급이 핵심이다. 한번 설치된 장치는 10년 이상 사용되는 경우가 많아 안정성과 호환성이 최우선 과제로 꼽힌다. 반면 웨어러블 기기는 제품 주기가 짧아 전력 효율과 소형화가 경쟁력을 좌우한다. 자동차 전장 분야는 차량 인증 규격 충족과 실시간성 보장이 필수 조건이다. 결국 같은 MCU라도 적용 산업에 따라 요구 조건이 달라지고, 이에 따라 벤더들의 전략도 세분화될 수밖에 없다.

이는 시장 전망에서도 도드라지게 나타난다. 시장조사업체 그랜드뷰리서치에 따르면 글로벌 IoT MCU 시장은 2024년 약 55억달러에서 2030년까지 연평균 16% 이상 성장할 전망이다. 한번 설치되면 10년 이상 사용되는 산업용 장비 특성에 따라 안정적인 공급망과 호환성이 가장 중요한 경쟁 요인으로 꼽힌다.

이미 MCU 기반 AI는 생활가전, 산업 IoT, 웨어러블, 자동차 등 다양한 영역에서 상용화 단계에 접어들었다. △음성인식 리모컨 △AI 청소기 △자세 교정 스마트 미러 △모터 이상 감지 및 예지 보수 시스템 △심박·수면 데이터 분석 스마트워치 △동작 인식 피트니스 트래커 △저속 보조 주행(ADAS) 기능 △졸음 감지 탑승자 모니터링 등이 대표적이다.

전문가들은 향후 엣지 AI MCU 경쟁이 단순한 칩 성능을 넘어 '생태계 전쟁'으로 확산될 것으로 보고 있다. 하드웨어 성능은 빠르게 상향평준화되고 있기 때문에, 얼마나 많은 개발자를 확보하고 레퍼런스 디자인을 제공하느냐가 승부처라는 분석이다. 오픈소스 커뮤니티가 MCU 시장에서 강력한 영향력을 발휘하는 이유도 여기에 있다.

업계 관계자는 "AI는 이제 기술적 가능성이 아니라 제품의 기본 요건이 됐다"며 "이미 다양한 MCU 벤더들이 경량 AI 경쟁에 본격적으로 뛰어든 상황"이라고 말했다. 이어 "이제는 디바이스에 AI를 탑재하지 않으면 경쟁력이 있을지 의문을 제기해야 할 시점"이라며 "이것이 바로 엣지 AI 시대가 본격적으로 열렸다는 신호"라고 강조했다.

정수연 기자 ssu@viva100.com