ROHM SEMICONDUCTOR

Electronics for the Future

# 온 디바이스 (On-Device) 학습 AI 칩 소개 자료

2022년 9월 27일
로옴 주식회사
마케팅 커뮤니케이션부

* 「tinyMicon MatisseCORE™」「RapidScope™」는 로옴 주식회사의 상표 또는 등록상표입니다.
* 본 자료는 발행일 시점의 정보로, 예고 없이 변경되는 경우가 있습니다.

# AI와 기계 학습의 개요

**인공지능 (Artificial Intelligence)**

인간의 기능 일부를 실현하는
화상 인식 등

**기계 학습 (Machine Learning)**

AI가 기계적으로 학습하는 것

딥러닝 : 학습을 심화시키는 것

**뉴럴 네트워크**
**(Neural Network)**

기계 학습의 일종

## AI의 학습과 추론

고양이와 개의 화상 인식 AI의 경우,

**학습** ・・・ AI가 많은 이미지를 통해, 고양이와 개의 특징을 학습하는 것

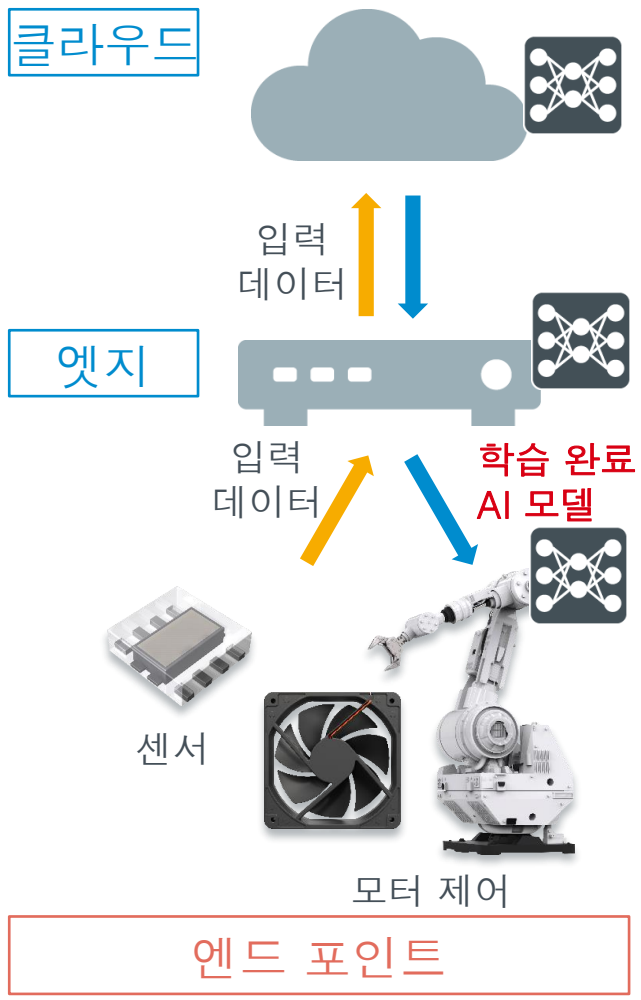
**추론** ・・・ AI가 이미지를 통해, 고양이인지 개인지 판정하는 것

**학습에는 연산 능력이 필요!**

# AI는 클라우드에서 엣지, 엔드 포인트로 확대되고 있다.

| | 기존 클라우드 AI | 엣지 AI | 엔드 포인트 AI |
|---|---|---|---|
| AI 기능 | 학습, 추론 모두 클라우드 | 학습은 클라우드<br>추론은 엣지 | 학습은 클라우드<br>추론은 엔드 포인트 |
| 요구 성능 | • 우수한 학습 능력<br>• 고도의 보안 | • 네트워크 부하 경감<br>• 짧은 응답 시간<br>• 저소비전력 | • 네트워크 부하 zero<br>• 매우 짧은 응답 시간<br>• 초저소비전력 |
| 과제 | • 통신 비용, 전력 증대<br>• 응답 시간 변동 있음<br>• 보안 비용 발생 | • 엣지에 고성능 FPGA나 GPU 필요<br>• 약간의 응답 시간 변동 있음 | • 내장 MCU의 성능에 따른 AI 모델로 한정 |

## AI의 확대 이미지



# 클라우드 타입 AI 시스템과 엔드 포인트 타입 AI 시스템의 비교

| | 클라우드 타입 AI 시스템 | 엣지 타입 AI 시스템 | 엔드 포인트 타입 AI 시스템 |
|---|---|---|---|
| 클라우드 컴퓨터 | AI 칩 | 필요 데이터만 입력 | 필요 데이터만 입력 |
| 엣지 컴퓨터 | 데이터 입력 / AI 출력 | AI 칩<br>데이터 입력 / AI 출력 | 필요 데이터만 입력<br>AI 칩 |
| 엔드 포인트 | 센서<br>모터 제어 | 센서<br>모터 제어 | AI 칩<br>센서<br>AI 칩<br>AI 칩<br>모터 제어 |
| | 클라우드 컴퓨터의 AI에 부하가 집중된다 | 엣지 컴퓨터의 AI에 학습 · 추론의 부하 분산 가능 | 엔드 포인트의 AI에 부하 분산 가능 |

# 엔드 포인트 타입 AI 시스템의 응용 예

이번
주요 타겟

## 화상 인식

복잡한 AI 실현에
고성능 GPU / FPGA 필요

## 각종 기기의 고장 예지

비교적 간단한 AI로 충분 /
고정밀도보다 사이즈, 비용 중시

## 기계의 고장이 먼저 나타나기 쉬운 모터의 고장 예지

**과제**

- 설치기기마다 학습 필요
- 환경 변화가 있으면 재학습 필요
- IC 제품 각각 설계 필요

→ **효율화 필요**

이러한 과제를 **독자 기술 (=온 디바이스 학습)**로 해결

**저전력, 고속 응답, 초소형으로**
**실시간 학습이 가능한 새로운 AI 솔루션 개발**

# 온 디바이스 학습 (디바이스 상에서의 기계 학습)

## 온 디바이스 학습 알고리즘

慶應義塾 Keio University

ROHM SEMICONDUCTOR

## 실제 디바이스에서의 회로 기술

### 온 디바이스 학습

디바이스 상에서 고속으로 AI 학습하는 기술

→ 칩 내에서 학습 가능하여, 학습용 데이터 준비 불필요

→ 클라우드 등에서의 사전 학습 불필요

→ **현장에서 학습이 가능**하여, 편차나 환경 변화에 강하다

+

### AI 액셀레이터 (AxICORE-ODL)

→ **저렴한 하드웨어 회로**로 AI 실현

### 소형 CPU MatisseCORE

→ AI의 구성을 **소프트웨어로 유연하게 변경** 가능

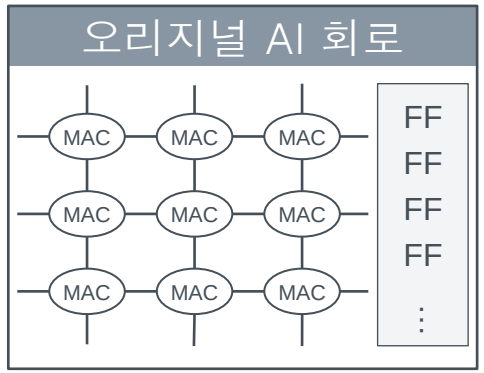
### 특징

① **낮은 비용**
= 엣지 디바이스용으로 상정 (해석적 중시 계산)

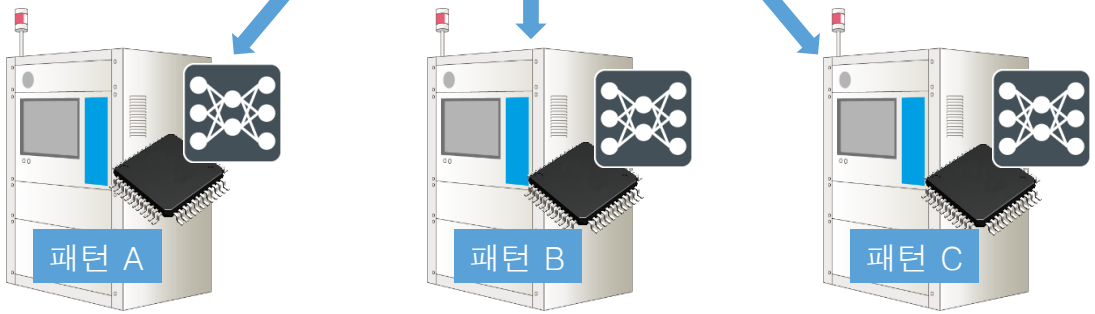
② **추종 능력**
= 변화하는 패턴에 대응 (경량 망각 기구)

③ **고정밀도**
= 정상 패턴이 여러 개 있어도
정밀도 유지 (앙상블 방법)

④ **안정 동작**
= 탑재되어 상시 가동
(과도 학습 억제와 출력의 안정화)




**설치 장소에 따라 달라지는 환경을 현장에서 학습 가능**
**(현장 학습)**



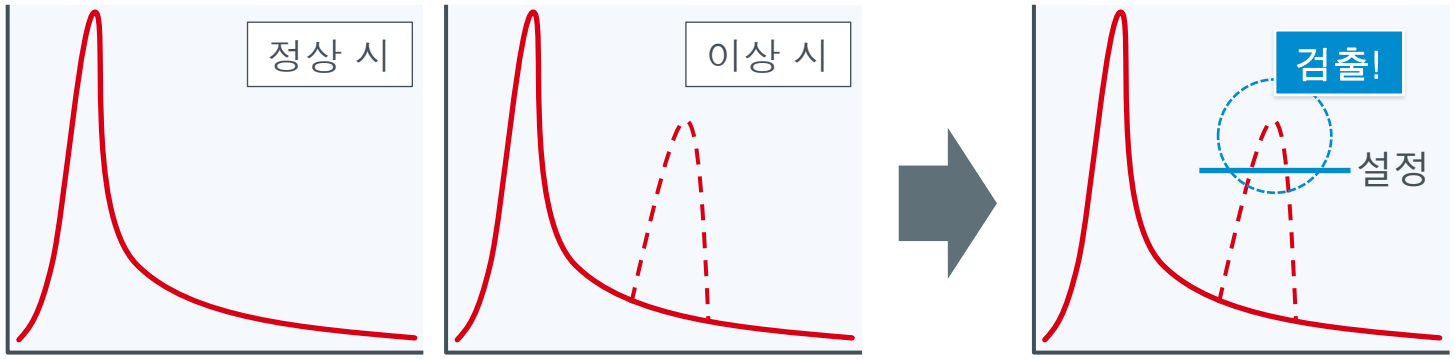

설치 장소 A　　설치 장소 B　　설치 장소 C

장소에 따른 사전 데이터 수집 불필요!
각 디바이스에 다운로드 불필요!

# 보충 AI를 통한 고장 예지 (고장 징후 검출), 이상 검출이란?

## AI는 정상 동작 시에서의 변화를 수치화함으로써, 미지의 이상 상태를 검출할 수 있다

### 기존 방법을 통한 고장 예지




이상 시의 변화 (특정 피크 출현 등)를 알고 있으면 검출 가능

어떤 데이터가 입력되는지,
이상 시의 변화가 어떤 것인지를
사람이 설정하지 않으면 검출 불가능

### AI를 통한 고장 예지


예상 밖의 이상 상태, 또는 이상 시의 변화를 알고 있지 않아도, 이상 검출 가능

어떤 데이터가 입력되더라도,
AI가 정상 상태를 학습함으로써,
이상 상태를 검출 가능 (추론 가능)

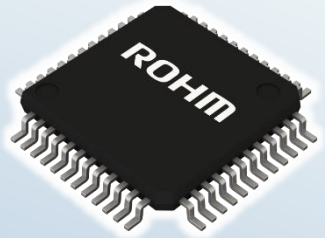
**온 디바이스 학습 알고리즘은 실시간으로 현장에서 실현 (클라우드 서버 불필요)**

# 로옴의 엔드 포인트용, 프로토타입 온 디바이스 AI 칩 「BD15035」 개요

## 온 디바이스 학습에 필요한 AI 액셀레이터, CPU, 입력 I/F를 1chip화

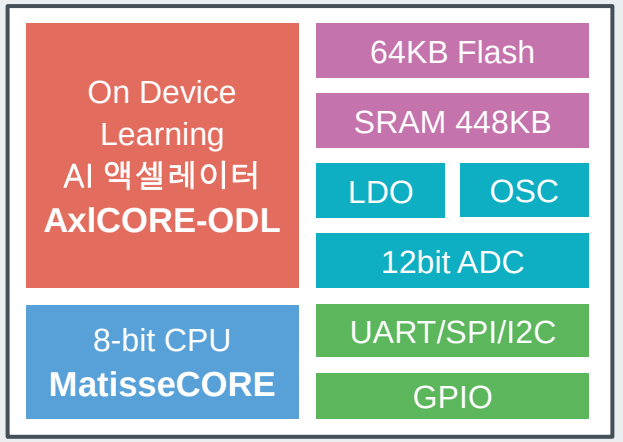
### 주요 회로 · 기능

- AI 액셀레이터 「AxlCORE-ODL」 탑재
  - AI의 베이스에 온 디바이스 학습 알고리즘 채용 (3층 뉴럴 네트워크)
  - FFT, 필터 처리 가능
- 8-bit CPU 「tinyMicon MatisseCORE™」 탑재
- 입력 I/F에 UART / SPI / I2C, 12bit ADC 탑재




**BD15035**
TQFP48V Package
(9.0 mm × 9.0 mm × 1.2 mm)

BD15035 회로




### 특징

**AI 기능**
- 온 디바이스 학습 가능 : 사전 학습이나 클라우드 서버에서의 해석 불필요 (3층 뉴럴 네트워크)

**초저소비전력**
- 수 10mW의 소비전력 : 배터리 구동 및 엔드 포인트에서의 동작 가능

**소형 칩**
- AI 기능을 초소형 AI 액셀레이터로서 재구축
- 소형, 고효율 8-bit CPU

**고속 처리**
- AI 액셀레이터를 통한 고속 처리로 CPU에 대한 부하가 적다

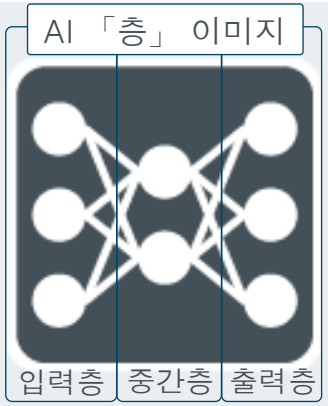
**기기가 설치된 현장에서 실시간 고장 예지 (고장 징후 검출) 실현**

# 각종 AI 칩과 로옴의 엔드 포인트용 AI 칩 성능 비교

| | 클라우드 컴퓨터용 AI 칩 | 엣지 컴퓨터용 AI 칩 | 기존 엔드 포인트용 AI 칩 | 로옴 엔드 포인트용 AI 칩 |
|---|---|---|---|---|
| 요구 성능 | • 우수한 학습 능력<br>• 고도의 보안 | • 네트워크 부하 경감<br>• 짧은 응답 시간<br>• 저소비전력 | • 네트워크 부하 zero<br>• 매우 짧은 응답 시간<br>• 초저소비전력 | • 네트워크 부하 zero<br>• 매우 짧은 응답 시간<br>• 초저소비전력 |
| 하드웨어 구성 | 고성능 GPU/<br>기계 학습 전용 프로세서 | 내장 GPU / FPGA | MCU | AI 액셀레이터<br>＋Matisse 탑재 MCU |
| 소비전력 | 20W ~ 200W | 2W ~ 10W | 20mW ~ 1000mW | → 약 30mW<br>※특정 어플리케이션 동작 시의 실측치 |
| 응답 시간 | 수 초 ~ 수십 초 | 수 초 | 밀리초 | 밀리초 |
| 학습 | 가능 | 불가능<br>※학습 완료된 AI 모델 사용 | 불가능<br>※학습 완료된 AI 모델 사용 | → 가능 |
| 추론 | 가능 | 가능 | 가능 | 가능 |

**소비전력이 불과 수 10mW인 AI 칩으로, 학습 · 추론 가능**
**엔드 포인트에서 실시간 고장 예지 실현**

# 보충 AI 매핑 이미지




## 딥러닝 (수십층의 중간층을 지님)

※응용 예

- 사람 대신 바둑이나 장기를 둔다
- 기상 정보를 예측한다
- 감시 카메라와 화상의 사람을 식별한다

etc

## 3층 뉴럴 네트워크

※응용 예

- 사람의 동작을 식별할 수 있다

예) 이미지 센서로 사람이 쓰러져 있는지, 서 있는지 정도를 식별

### 3층 뉴럴 네트워크 AI 칩 「BD15035」

(메모리 등의 사양에 따른 제약이 있다)

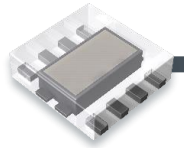
- 가속도 및 전류, 음성 등의 식별을 통한 고장 예지

# AI 칩 「BD15035」의 처리 · 역할 분담

## 입력 데이터에 대해 AI 칩 출력까지의 단계

수많은 입력 데이터에 대해 학습을 실행하여,
「이상 상태」를 수치화

입력 데이터



각종 센서 신호 등

입력 I/F
- 12bit ADC
- UART/SPI/I2C

**전처리 회로**
- FFT
- 필터 처리

**AI 학습 · 추론 (AI 회로)**
- 3층 뉴럴 네트워크

x y

이상치

$$L = \sum_{k=0}^{n} |x - y|^2$$

**후처리 회로**
- 출력 판정

엣지
컴퓨터 측

AI 회로가 연산하기 쉽도록
전처리 실행

Matisse에 설정된 이상 판정의
임계치에 따라 출력 필터링 가능

### AI 액셀레이터 「AxlCORE-ODL」

**연산 능력이 필요한 기능 담당**
- 입력 데이터의 전처리
- AI 연산

### BD15035

On Device Learning AI 액셀레이터 **AxlCORE-ODL**

8-bit CPU **MatisseCORE**

- 64KB Flash
- SRAM 448KB
- LDO
- OSC
- 12bit ADC
- UART/SPI/I2C
- GPIO

### 8-bit CPU 「MatisseCORE」

**어플리케이션에 따라 변경이 필요한 기능 담당**
- AI 모델 설정
  (뉴럴 네트워크의 노드 수 등)
- 전처리 회로에 대한 데이터 입력 (보조)
- 학습 · 추론의 전환 및 제어
- 출력 판정의 임계치 설정

상세

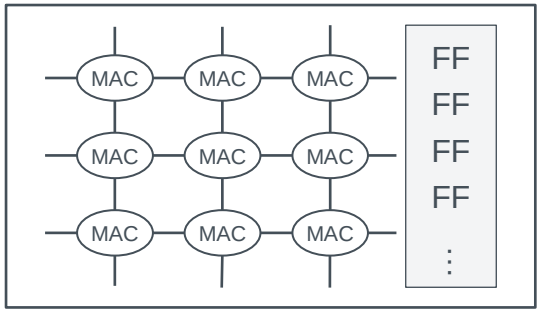
로옴의 기술 :

# AI 액셀레이터 「AxlCORE-ODL」을 통한, 소형화와 고속화

## 게이오 대학에서 제공받은 온 디바이스 학습 회로 (AI 회로)를 AI 액셀레이터로 재설계하여 게이트 수 삭감

### 오리지널 AI 회로

500만 게이트

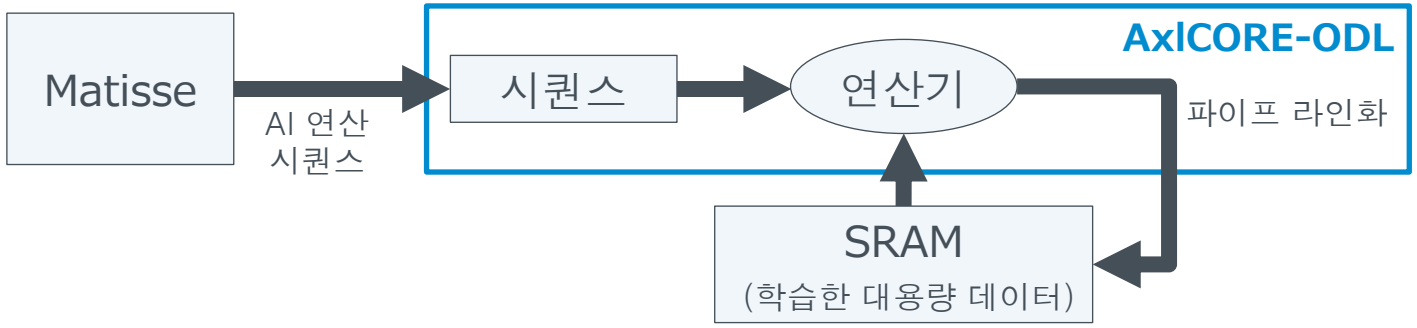



게이트 수 1/250
0.4%까지 소형화

### 로옴의 AI 칩 (AI 액셀레이터＋Matisse)

2만 게이트

특허 2건 출원중




- 고정 소수점 32bit
- 다수의 적화연산기 (MAC)와 FF로 구성되어 회로가 매우 크다
- AI의 구조 고정

✓ **bfloat 16bit 부동 소수점 연산** 채용, 이진연산에 비해 정밀도가 좋다
(대부분의 AI 칩은 고속화 · 메모리 절약화를 위해, 1-2bit의 이진연산을 채용하여 정밀도가 떨어진다)

✓ **Matisse로 AI 연산 시퀀스를 설정**함으로써, 연산기를 1개로 집약

✓ **AI의 구조 (입력 데이터 수, 알고리즘) 가변**
- 처리 시간 및 메모리 사용량의 밸런스, 알고리즘 개량 가능

✓ SRAM에서의 데이터 취득, 연산, 보존을 파이프 라인화하여 **처리 속도 3배** 향상

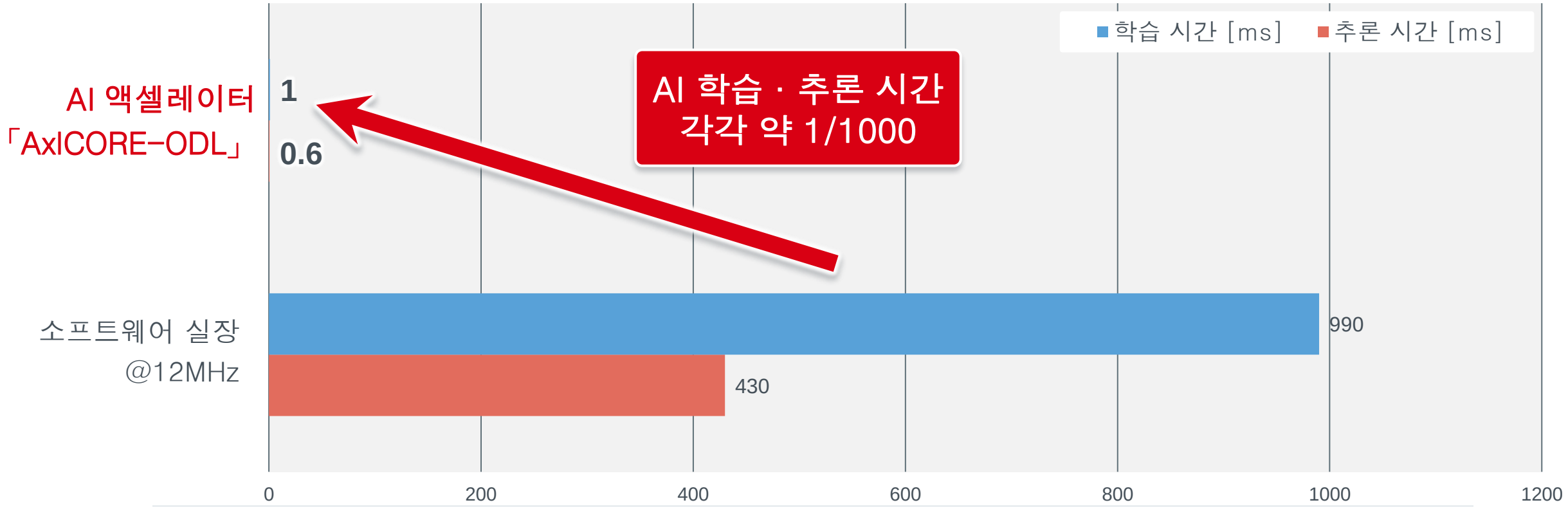
✓ 온 디바이스 학습 알고리즘을 통해, 칩 상에서 **3층 뉴럴 네트워크 학습** 가능

✓ 교사 없이 학습이 가능한 자동 엔코더로, **사전 학습 없이** 이상 검출 가능

상세

# 로옴의 기술 : AI 액셀레이터 「AxlCORE-ODL」을 통한, 소형화와 고속화

## 학습 · 추론 실행 시간 비교 (뉴럴 네트워크 : 입력층 96 노드, 중간층 12 노드를 설정)




- ✓ **CPU 부하가 적다.** 비용이 저렴한 8-bit CPU로도 충분한 어플리케이션 처리 능력 확보
- ✓ **고속 샘플링 대응.** 10kHz 정도의 고주파 대역에 나타나는 이상 검출 대응
- ✓ 시계열 데이터의 전처리로서 필요한, **FFT도 AI 액셀레이터 (회로에 내장)로 실현 가능**

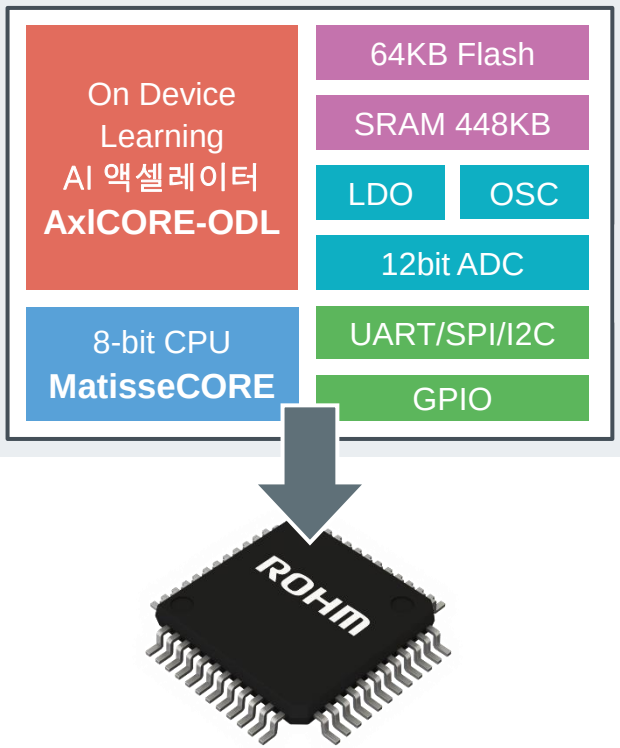
# AI 칩 「BD15035」 평가 보드



## 프로토타입 온 디바이스 AI 칩 「BD15035」

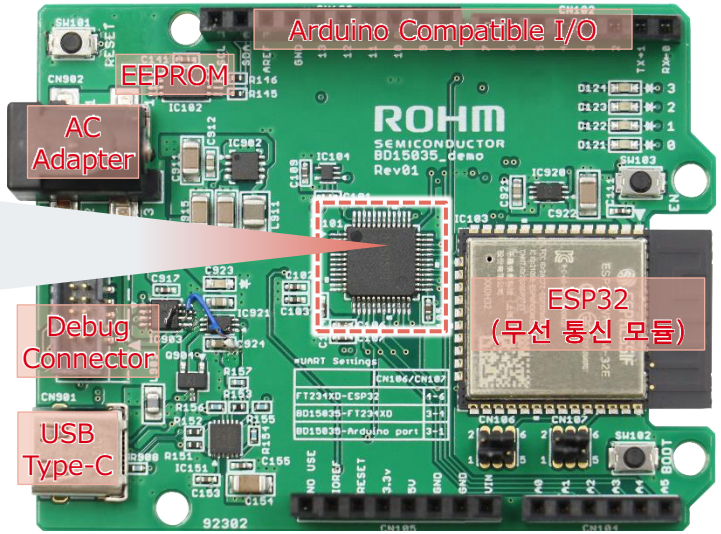
- AI 액셀레이터 「AxlCORE-ODL」 탑재
- 고효율 8-bit CPU 「tinyMicon MatisseCORE™」 탑재

### BD15035



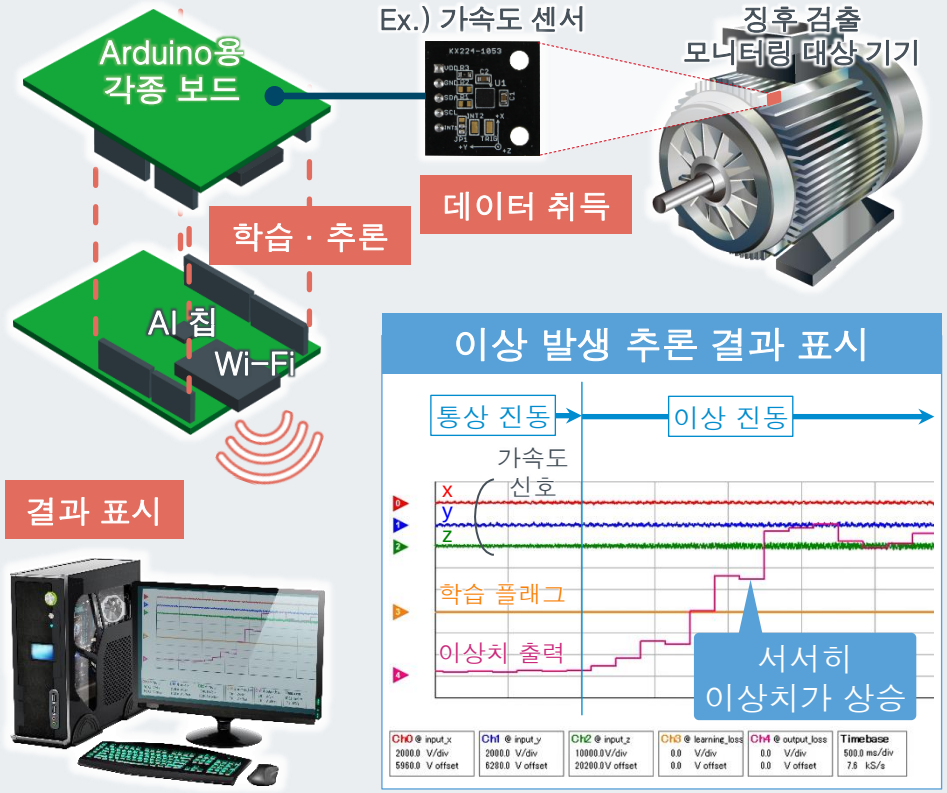

TQFP48V Package
(9.0 mm × 9.0 mm × 1.2 mm)

### 평가 보드




- Arduino용 보드 접속 가능
- Wi-Fi / Bluetooth® 모듈 탑재
- 64kbit EEPROM 탑재

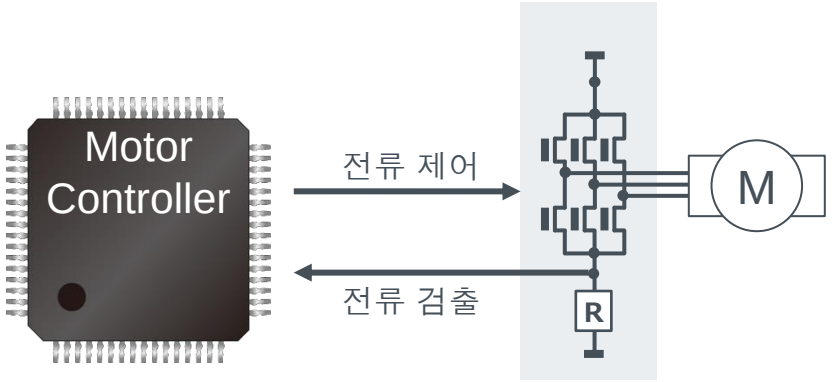
### 평가 보드 사용 이미지
(가속도 센서를 사용한 경우)



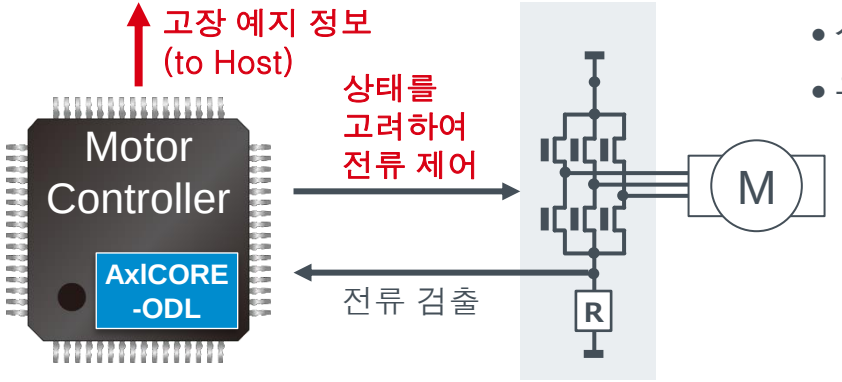
사용 예 1 :

# 온 디바이스 학습 AI 기능 탑재 모터 컨트롤러

## 기존의 모터 제어 환경



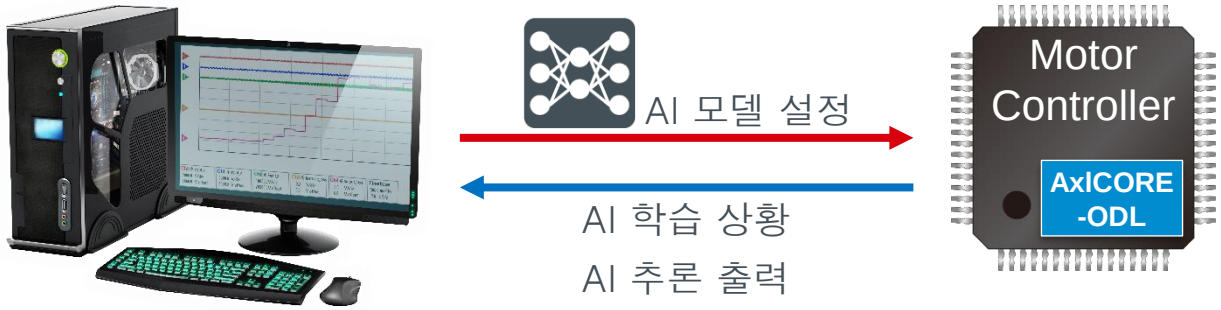

## AI 기능 탑재 모터 제어 환경




- 실시간 고속 추론
- 온 디바이스 학습

**기존의 모터 컨트롤러 IC에 비해,
비용이 저렴하고, 추가 부품 없이 간단히 AI 기능 도입 가능**

AI 모델 구축에서 평가까지, 간단히 실현할 수 있는 툴 개발중




**복잡한 모델의 설계 및 다수의 파라미터 조정 불필요**

- AI의 출력을 모니터링하면서, AI 모델을 간단히 구축
- 최소한의 파라미터로 AI 모델 조정
  (입력 데이터 수, 이상 판정 임계치)
- 버튼 하나로 디바이스 상에서 재학습

사용 예 2:

# 온 디바이스 학습 AI 기능 탑재 엣지용 범용 마이컴

**범용 마이컴** ← 엔드 포인트 AI를 에드온

- 온 디바이스 학습 AI 액셀레이터를 peripheral로 지닌 범용 마이컴
- 고속 AI 연산을 작은 하드웨어로 실행,
  어플리케이션 기능은 소프트웨어로 자유롭게 실장 가능

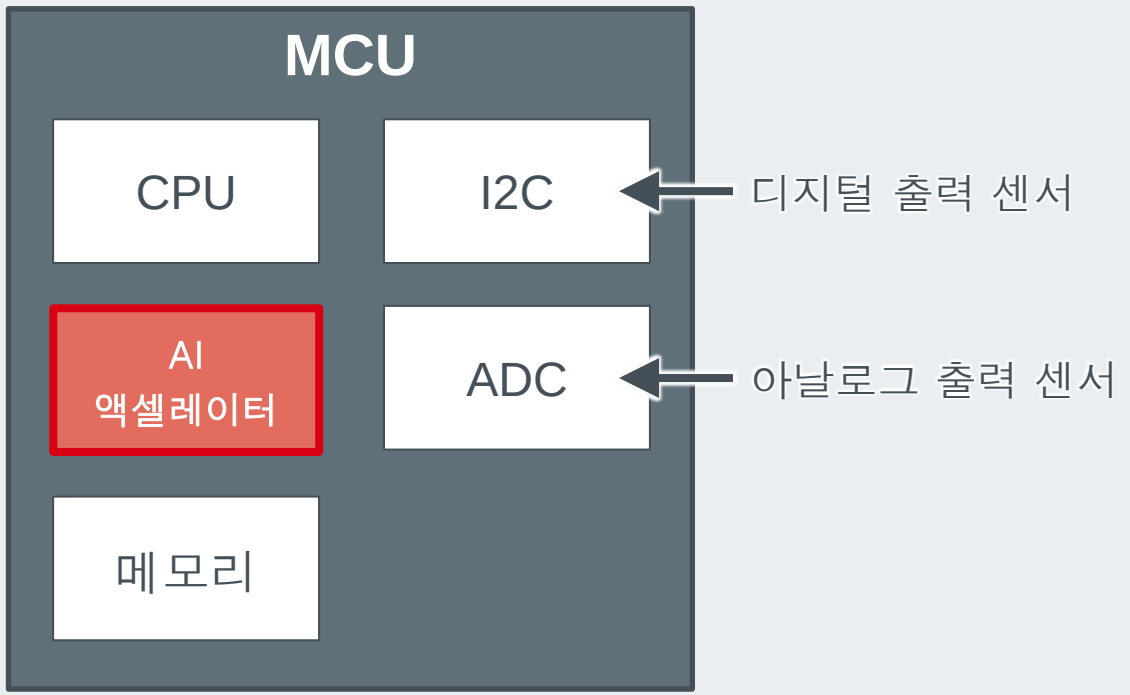
**산업기기, 자동차기기, 가전 등의
엣지 / 엔드 포인트 MCU에
간단히 AI 기능 (고장 예지 검출 등) 추가**

## 메리트

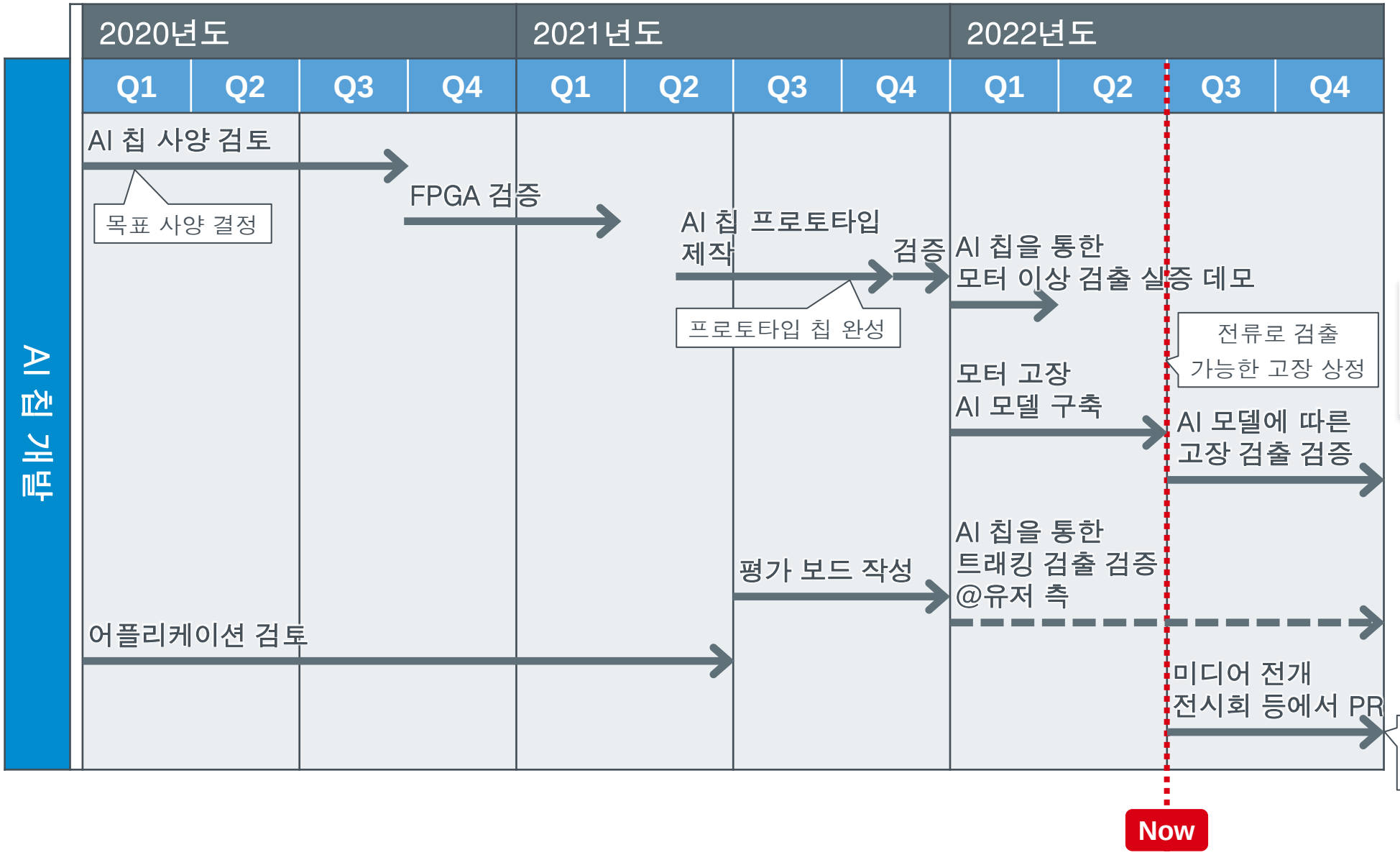
- 필요한 AI 연산은 하드웨어로 연산 가능하므로
  소프트웨어의 부하가 적어 어플리케이션 기능에 제한이 없다
- 사용중인 어플리케이션 마이컴의 대체만으로 AI 기능 추가 가능
- 추론뿐만 아니라 학습도 디바이스 측에서 가능하므로,
  설치 장소에 따른 최적화가 용이

**다양한 타입의 센서 입력을 동시에 처리 가능**

가속도, 전류, 온도, 조도, 마이크



# 프로토타입 칩 개발에서, 향후 제품화까지의 스케줄




## 향후 예정

**2023년도**
로옴 양산 칩 개발 착수

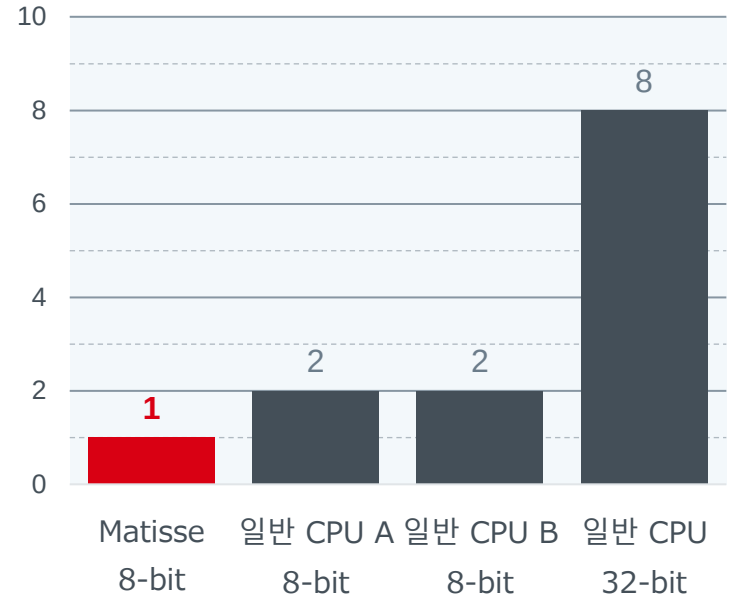
**2024년도**
양산 개시

**모터 · 센서용 IC 제품에 탑재하여, 고장 예지 / 예지 보전에 기여**

# ~ 소형 & 고속 8-bit CPU ~

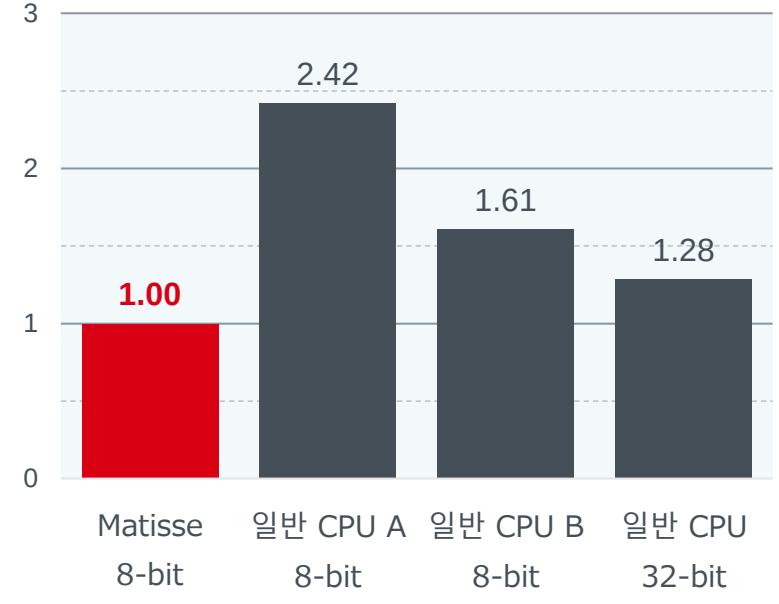
## 일반적인 소형 CPU와의 성능 비교 (Matisse=1로 가정)

### 게이트 사이즈 비교



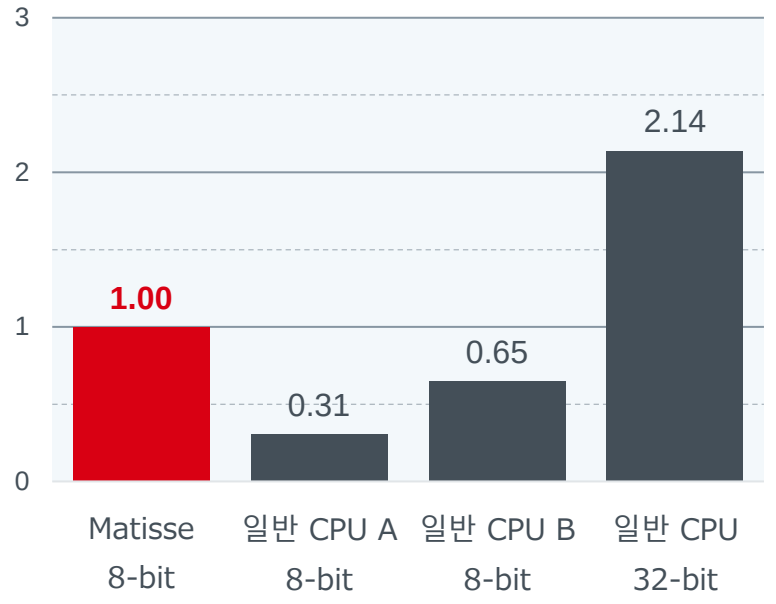

우수한 면적 절약성

### ROM 사이즈 비교 (8-bit 연산 프로그램 시)




콤팩트한
프로그램 코드 사이즈

### 처리 성능 비교 (Dhrystone)



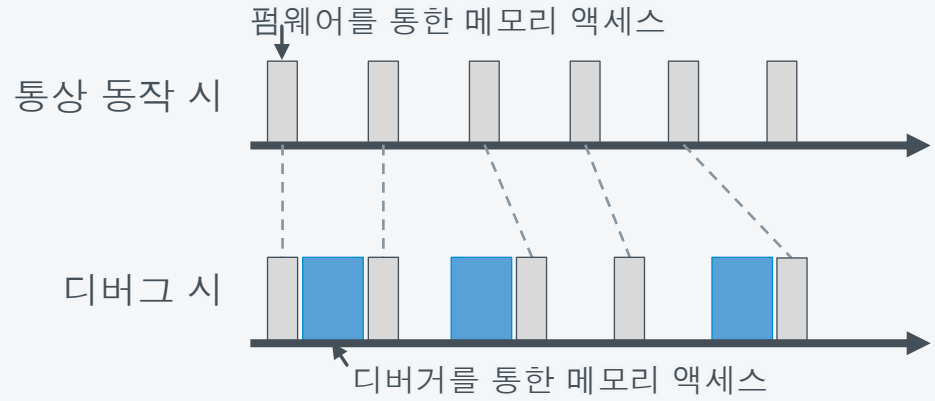

고속 연산 처리

**Matisse는 콤팩트한 칩 면적과 프로그램 코드 사이즈,
고속 연산 처리를 높은 수준으로 실현 (자동차기기 규격 ASIL-D까지 대응 가능)**

참고 : tinyMicon MatisseCORE™

# ~ 내장 사용에 최적인 실시간 디버그 기능 ~

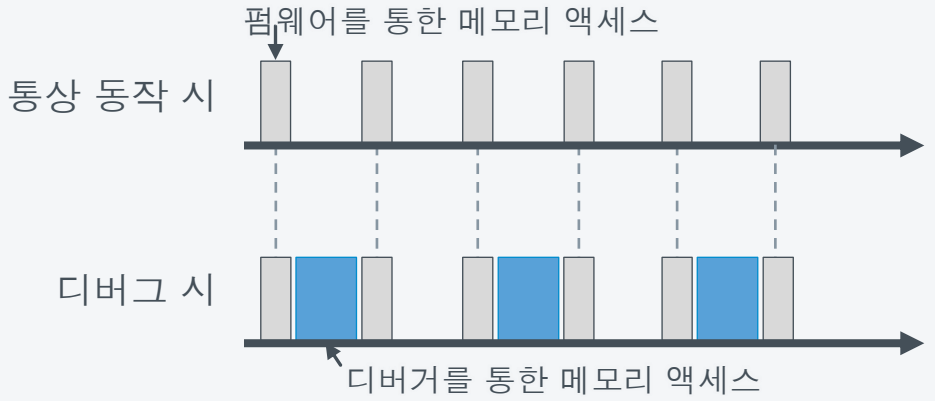
## 기존 방법




디버거를 통한 메모리 액세스가 펌웨어의 동작을 방해

디버그로 인해 프로그램의
동작이 크게 변하는 경우가 있다

## Matisse의 실시간 디버그 기능




펌웨어의 동작을 방해하지 않도록 디버거의 메모리 액세스를 조정

프로그램의 동작을 일절 변경하지 않고
디버그 실행 가능

## Matisse는 손실 없이, CPU 내부 정보의 출력 / 변경 가능

- 실시간 디버그를 통해, **통상 동작 시와 디버그 시의 동작에 변화가 없음**
- 모터 제어 등, **동작을 멈추고 디버그할 수 없는 어플리케이션에서도 간단히 디버그 가능**
- 통상적으로는 볼 수 없는 **IC 내부의 변수를 실시간으로 출력하여 파형 표시 가능**

파형 표시용
소프트웨어
RapidScope™


Electronics for the Future

ROHM Co.,Ltd.

# 주의 사항

- 본 자료에 기재되어 있는 내용은 로옴의 제품（이하, 「로옴 제품」）소개를 목적으로 합니다.
- 로옴 제품 사용 시에는, 별도로 최신 사양서 및 데이터시트를 반드시 확인하여 주십시오.
- 본 자료에 기재되어 있는 정보는, 별도의 보증 없이 제공되는 것입니다.
  만일, 해당 정보의 오류 또는 사용으로 기인하는 손해가 고객 또는 제3자에게 발생하는 경우, 로옴은 일절 책임을 지지 않습니다.
- 본 자료에 기재되어 있는 로옴 제품에 관한 대표적 동작 및 응용 회로 예는 일례로서 제시된 것이며, 이와 관련된 제3자의 지적재산권 및 기타 권리에 대해 권리 침해가 없음을 보증하는 것은 아닙니다.
- 상기 기술 정보의 사용으로 인해 분쟁이 발생하는 경우, 로옴은 해당 책임을 지지 않습니다.
- 로옴은, 로옴 또는 타사의 지적재산권 및 기타 모든 권리에 대해 명시적으로나 묵시적으로 그 실시 또는 이용을 허락하는 것은 아닙니다.
- 본 자료에 기재되어 있는 제품 및 기술 중, 「외국 외환 및 외국 무역법」 기타 수출 규제에 해당하는 제품 또는 기술을 수출하는 경우, 또는 해외에 제공하는 경우에는, 해당 법에 입각하여 허가가 필요합니다.
- 본 자료의 기재 내용은 2022년 9월 현재의 내용이며, 예고 없이 변경되는 경우가 있습니다.