Electronics for the Future

Introduction to ROHM's On-Device Learning AI Chip

(In this, AI Chip means SoC with built-in On-Device Learning AI accelerator)

Nov. 29, 2022
ROHM Co., Ltd.
Marketing Communications Dept.

Overview of Artificial Intelligence (AI) and Machine Learning

Artificial Intelligence

Carry out one of the parts of human functionality
(ex. image recognition and other methods)

Machine Learning

AI learning mechanically (automatically)

Deep learning: Learning on a deeper level

Neural Network

(Deep Neural Network)
A type of machine learning

What is AI Learning and Inference?

Ex.) Image recognition AI for cats and dogs

Learning: The AI looks at many images and learns the characteristics of dogs and cats

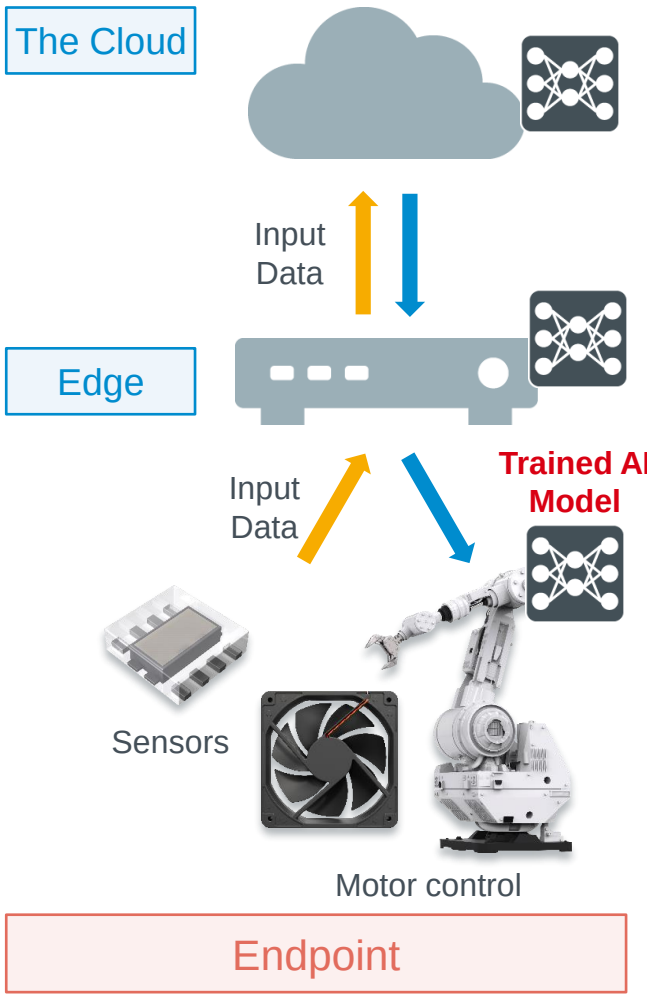
Inference: The AI looks at an image and determines whether it's a cat or a dog

Learning requires computing power

AI is Spreading from the Cloud to the Edge and Endpoints

	Conventional Cloud AI	Edge AI	Endpoint AI
AI Functions	The Cloud for learning and inference.	The Cloud for learning. The Edge for inference.	The Cloud for learning. Endpoints for inference.
Required Characteristics	• Excellent learning capability • Advanced security	• Network load reduction • Short response time • Low power consumption	• Zero network load • Extremely short response time • Ultra-low power consumption
Issues	• Increased communication cost and power • Large variations in response times • High security costs	• Requires high performance FPGAs and GPUs at the edge • Small variations in response times	• Limited to AI models based on embedded MCU performance

AI Expansion Image



Comparing Cloud-Based and Endpoint AI Systems

	Cloud AI System	Edge AI System	Endpoint AI System
Cloud Computer	AI Chip	Send only the necessary data	Send only the necessary data
Edge Computer	Data input / AI output	AI Chip; Data input / AI output	Send only the necessary data
Endpoint	Sensors; Motor control	Sensors; Motor control	AI Chip; AI Chip; AI Chip; Sensors; Motor control
	The load is concentrated on the cloud AI computer	The load for learning and inference can be distributed to the edge AI computer	The load can be distributed to the endpoint AIs

Endpoint AI System Applications

Image Recognition

- Requires high-performance GPU/FPGA to achieve complex AI

ROHM's main targets

Failure Prediction of Various Equipment

- Relatively small AI is sufficient
- Size and cost are more important than accuracy

ROHM's main targets

Failure Prediction of Motors
(Where Machine Failures Tend to Appear First)

Issues	
	● Learning is required for each installed device ● Re-learning is required when there is a change in the environment ● Each IC needs to be individually designed

Need to improve efficiency

Solve these issues utilizing
proprietary technology (on-device learning)

Develop a new fast AI solution that is power-efficient, ultra-compact, and capable of real-time learning

On-Device Learning (Machine Learning on Devices)

On-device learning algorithms By Keio University

慶應義塾 Keio University

Circuit technology in real devices By ROHM

On-device Learning
Technology for fast AI learning on devices

→ Can be learned on the chip, eliminating the need to prepare training data

→ No need for prior learning in the cloud, etc.

→ **Onsite learning** provides resistance to variability and environmental changes

AI Accelerator (AxlCORE-ODL)

→ Achieves **low-cost hardware** circuitry for AI

Compact CPU MatisseCORE

→ **Flexibly change** the AI configuration via **software**

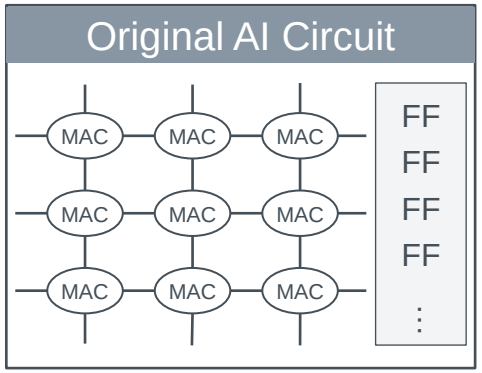
Features

① **Low cost**
= Intended for edge devices (analytical weight calculation)

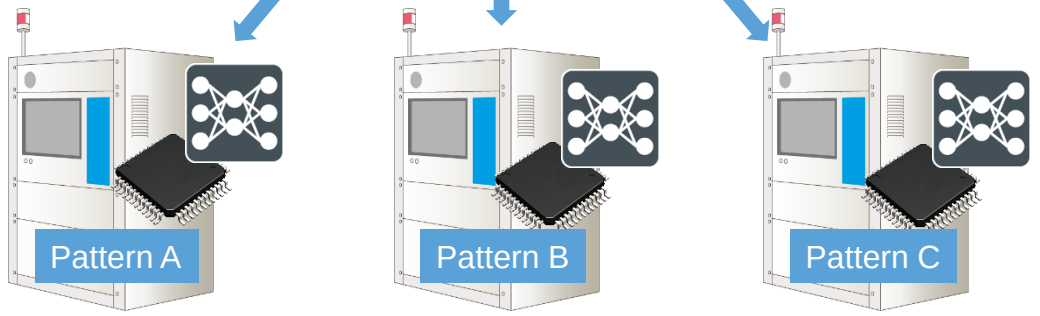
② **Adaptability**
= Responds to changing patterns (lightweight and forgetting mechanism)

③ **High accuracy**
= Maintains accuracy even with multiple normal patterns (ensemble method)

④ **Stable operation**
= Built-in and always ON (suppresses over-learning and stabilizes output)



Different environments for each installation can be learned in-situ (onsite learning)



Installation Location A | Installation Location B | Installation Location C

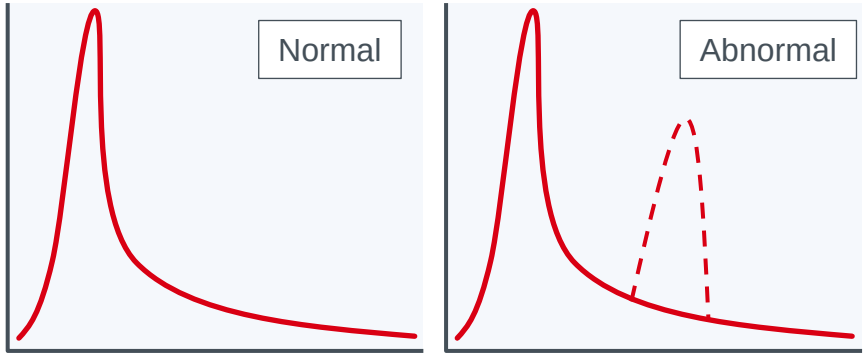
No prior data collection required for each location
No need to download to each device

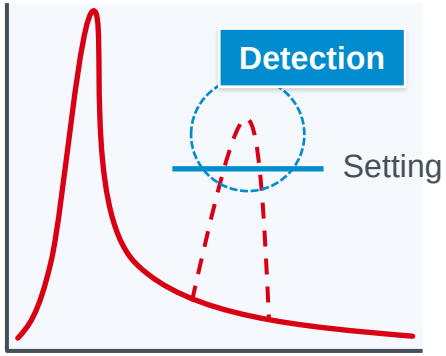
Supplement

What is AI-Based Failure Prediction (Predictive Failure Detection) and Anomaly Detection?

AI can detect even unknown abnormalities by quantifying changes from normal operation

Failure prediction by the conventional method

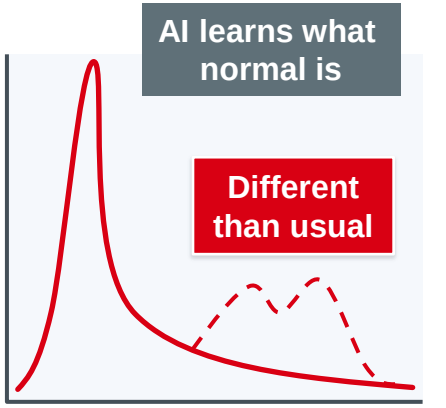




Detection possible if changes during anomalies (i.e. the appearance of specific peaks) are known

Cannot be detected without a person setting up what kind of data will be input and what changes will occur in the event of anomalies

Failure prediction by AI



Detects anomalies even when unexpected or when changes during anomalies are unknown

No matter what data is input,
the AI can detect (infer) abnormalities
by learning normal

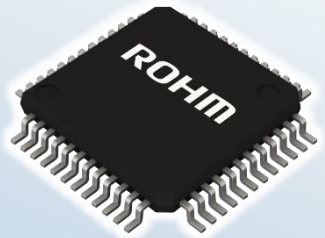
On-device learning algorithms are achieved onsite in real time (no cloud server required)

Overview of ROHM's On-Device Prototype Endpoint AI Chip: BD15035

Integrates an AI accelerator, CPU, and input I/F required for on-device learning on a single chip

Primary Circuits and Functions

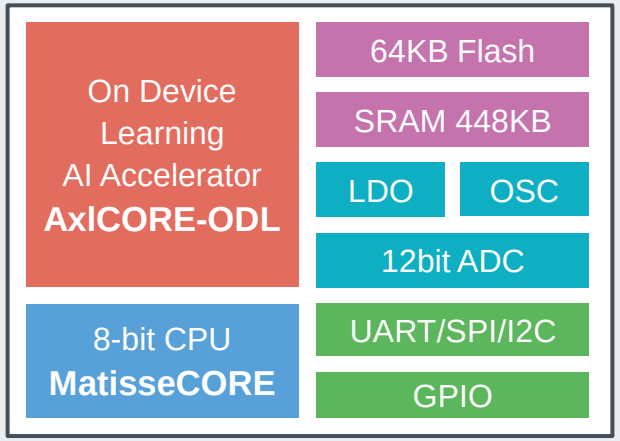
- Equipped with the AI accelerator 'AxlCORE-ODL'
  - Utilizes on-device learning algorithms as a base for AI (3-layer neural network)
  - FFT, filtering possible
- Built-in 8-bit 'tinyMicon MatisseCORE™'
- Incorporates UART/SPI/I2C I/F along with 12bit ADC



BD15035

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

BD15035 Circuit



Features

AI Functions

- On-device learning possible: No pre-training or cloud server analysis needed (3-layer neural network)

Ultra-Low Power Consumption

- Power consumption just a few tens of mW: Supports battery drive or endpoint operation

Compact Chip

- Rebuild AI functions as ultra-compact AI accelerators
- Compact, high-efficiency 8-bit CPU

High-Speed Processing

- High-speed processing via AI accelerators reduces CPU load

Enables real-time failure prediction (predictive failure detection) at equipment location

Performance Comparison of Various AI Chips with ROHM's Endpoint Chip

	Cloud Computer AI Chip	Edge Computer AI Chip	Conventional Endpoint AI Chip	ROHM's Endpoint AI Chip
Required Characteristics	• Excellent learning capability • Advanced security	• Network load reduction • Short response time • Low power consumption	• Zero network load • Extremely short response time • Ultra-low power consumption	• Zero network load • Extremely short response time • Ultra-low power consumption
Hardware Configuration	High-performance GPU/Dedicated machine learning processor	Embedded GPU/FPGA	MCU	AI Accelerator + Matisse-Equipped MCU
Power Consumption	**20W to 200W**	**2W to 10W**	**20mW to 1000mW**	**Approx. 30mW** *Actual measured value for specific application operation
Response Time	**Seconds to tens of seconds**	**Seconds**	**Milliseconds**	**Milliseconds**
Learning	**Possible**	**Not possible** *Uses pre-trained AI models	**Not possible** *Uses pre-trained AI models	**Possible**
Inference	Possible	Possible	Possible	Possible

Learning and inference possible with an AI chip consuming only tens of mW of power
Enables real-time failure prediction at endpoints

AI Mapping

Deep Learning (with Dozens of Intermediate Layers)

*Application examples

- Play Go or Shogi without a human opponent
- Predict the weather
- Identify people in surveillance videos and images

3-Layer Neural Network

*Application examples

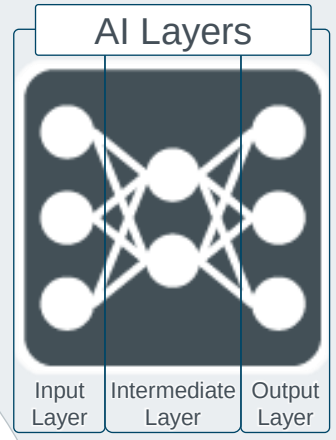
- Identify human movements

Ex: Using an image sensor to determine the degree to which a person is lying down or awake

3-Layer Neural Network AI Chip (BD15035)

(Depending on the memory and other spec constraints)

- Failure prediction by identifying acceleration, current, voice, etc.

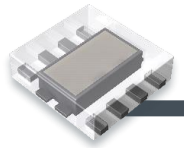


Processing and Division of Roles of the BD15035 AI Chip

Flow from input data to AI chip output

Learning is performed on a large amount of input data to quantify 'anomaly'

Input data



Various sensor signals, etc.

Input I/F
- 12bit ADC
- UART/SPI/I2C

Pre-processing Circuit
- FFT
- Filter Processing

AI Learning and inference (AI Circuit)
- 3-Layer Neural Network

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

Post-processing Circuit
- Output judgment

Edge Computer Side

Performing pre-processing makes it easier for AI circuits to compute

The output can be filtered by anomaly thresholds set in Matisse

AI Accelerator 'AxICORE-ODL'

Responsible for functions requiring computing power
- Pre-processing input data
- AI calculation

BD15035

- On Device Learning AI Accelerator **AxICORE-ODL**
- 8-bit CPU **MatisseCORE**
- 64KB Flash
- SRAM 448KB
- LDO
- OSC
- 12bit ADC
- UART/SPI/I2C
- GPIO

8-bit CPU「MatisseCORE」

Responsible for functions that need to be changed for each application
- AI model settings (no. of neural networks, etc.)
- Data input to pre-processing circuits (auxiliary)
- Switching and control of learning and inference
- Setting the threshold for output judgment

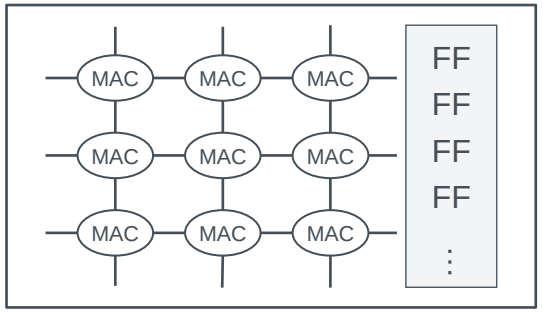
Details

ROHM Technology: Greater Miniaturization and Speed with AI Accelerator 'AxlCORE-ODL'

The on-device learning circuit (AI circuit) provided by Keio University has been redesigned with an AI accelerator to reduce the number of gates

Original AI Circuit

5 million gates

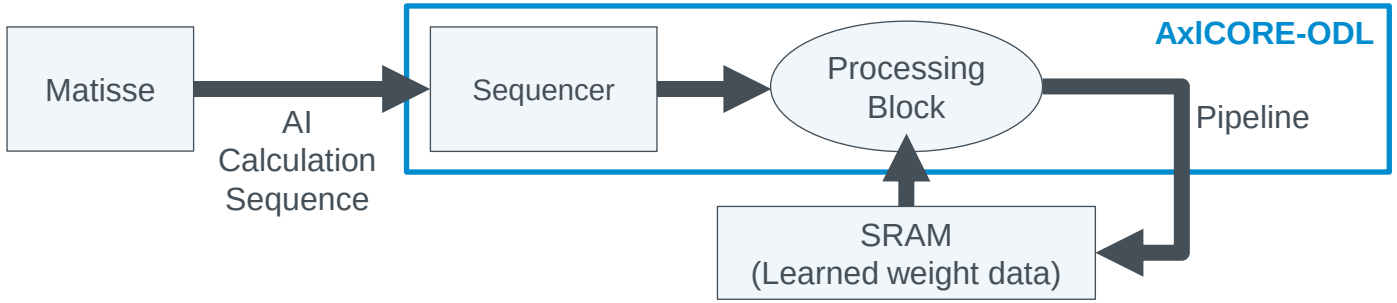


Gate count reduced to just 0.4% (by 250×) of the size

ROHM AI Chip (AI Accelerator + Matisse)

20k gates

2 patents pending



- Fixed-point 32bit
- Large-scale circuit consists of many multiply-accumulators (MAC) an FFs
- Fixed AI structure

✓ **bfloat 16bit floating-point arithmetic** features better accuracy than binary arithmetic (many AI chips lose accuracy by using 1-2bit binary operations for the sake of speed and memory)

✓ **Setting the AI operation sequence with Matisse** allows the computing unit to be consolidated into one

✓ **Variable AI structure (no. of input data, algorithms)**
Makes it possible to improve algorithms while providing an optimum balance processing time and memory usage

✓ **Processing speed is tripled** by pipelining the acquisition, computation, and storage of data from SRAM

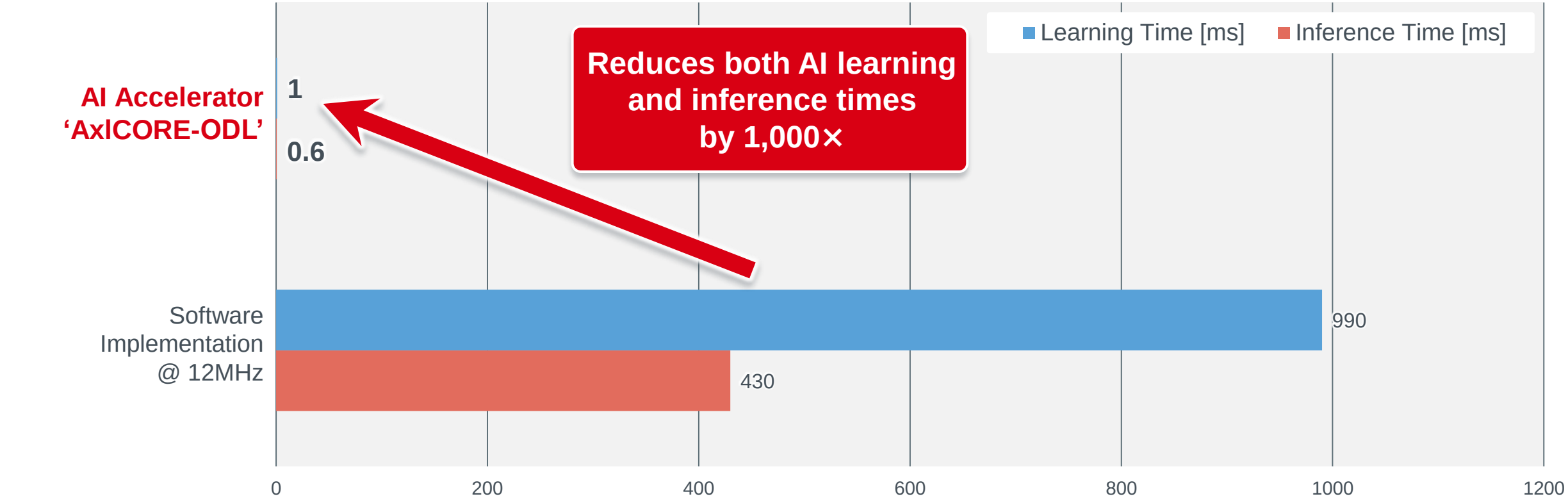
✓ On-device learning algorithms enables **training of 3-layer neural networks** on-chip

✓ Auto-encoder capable of unsupervised learning enables anomaly detection **without pre-training**

ROHM Technology:

Greater Miniaturization and Speed with AI Accelerator ‘AxlCORE-ODL’

Comparison of learning and inference execution times
(neural network setting: 96 nodes as input layer + 12 nodes as intermediate layer)



- ✓ **Requires minimal CPU load**. Sufficient application processor power is ensured even with low-cost 8bit CPUs.
- ✓ **Supports high-speed sampling**. Capable of detecting anomalies that appear in the high-frequency range of around 10kHz.
- ✓ **The built-in AI accelerator can also perform FFT**, which is necessary for pre-processing time-series data.

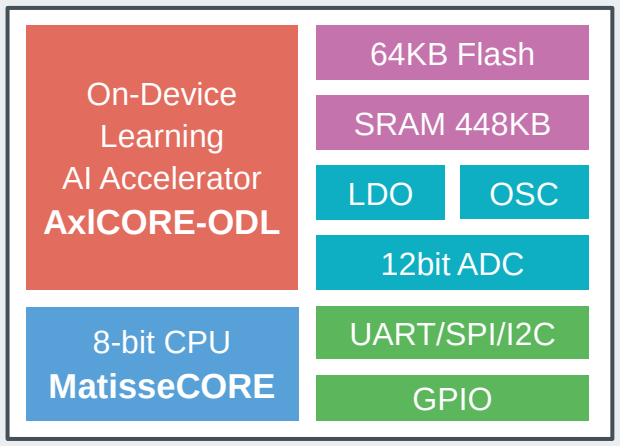
AI Chip (BD15035) Evaluation Board

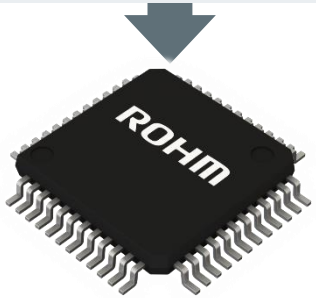
BD15035 Prototype On-Device AI Chip

- Equipped with AI accelerator 'AxlCORE-ODL'
- Built-in high efficiency 8-bit CPU 'tinyMicon MatisseCORE™'

Click here for the demo video

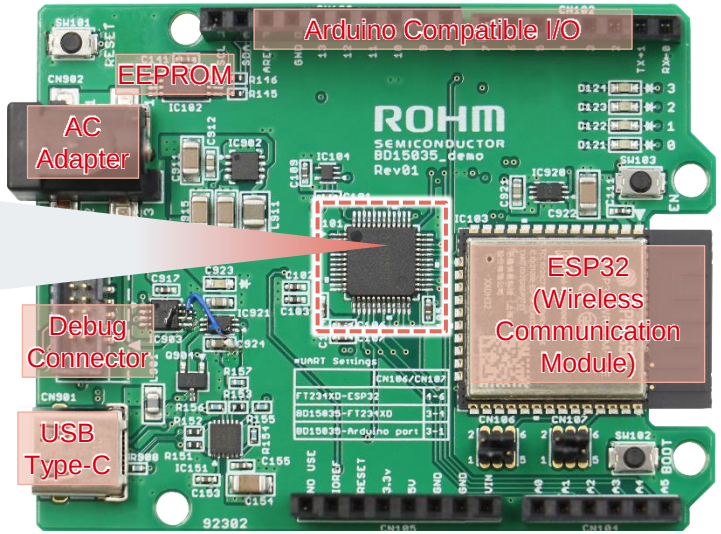
BD15035





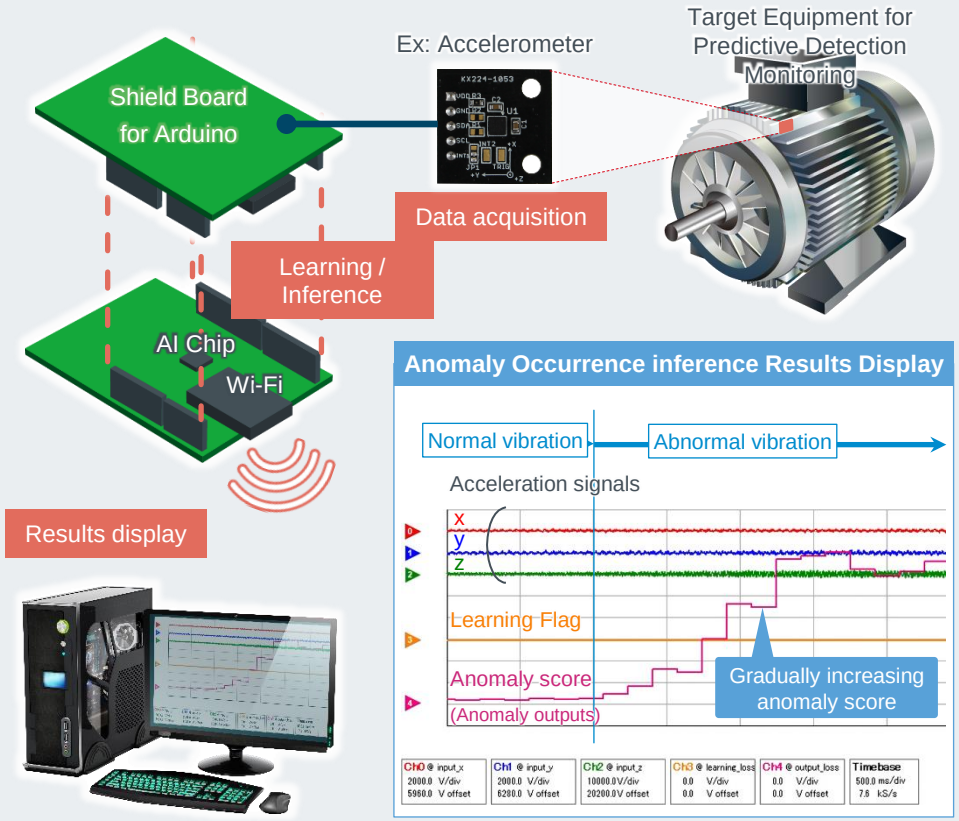
TQFP48V Package
(9.0 mm × 9.0 mm × 1.2 mm)

Evaluation Boards



- Connectable to an Arduino Shield Board
- Onboard Wi-Fi / Bluetooth® Module
- Built-in 64kbit EEPROM

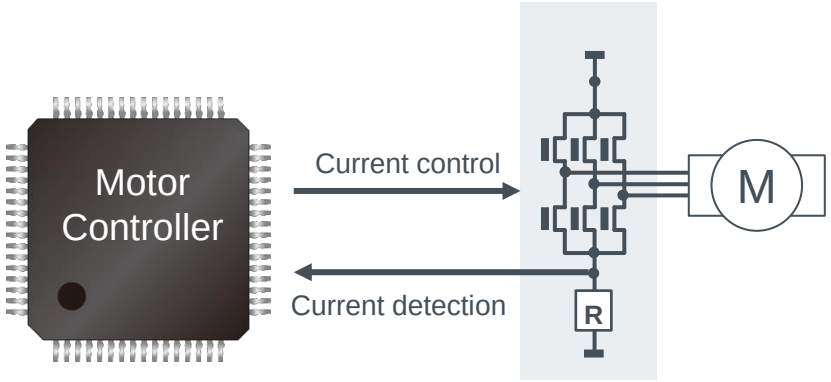
Evaluation Board Usage Diagram
(When using an accelerometer)



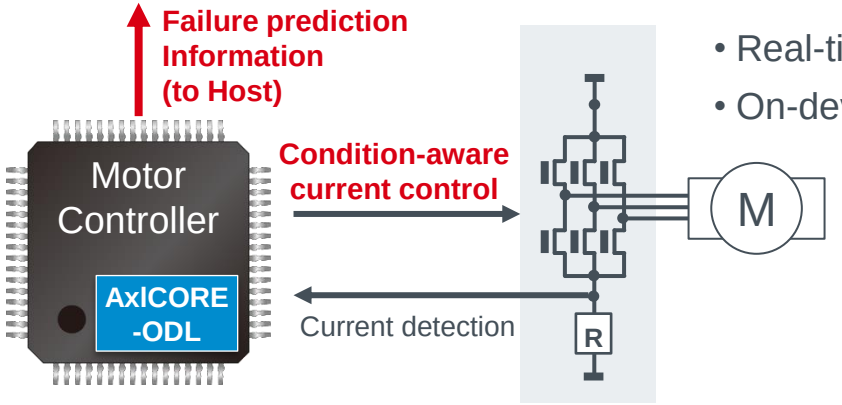
Expected Usage Case 1:

Motor Controller with On-Device Learning AI Function

Existing Motor Control Environment



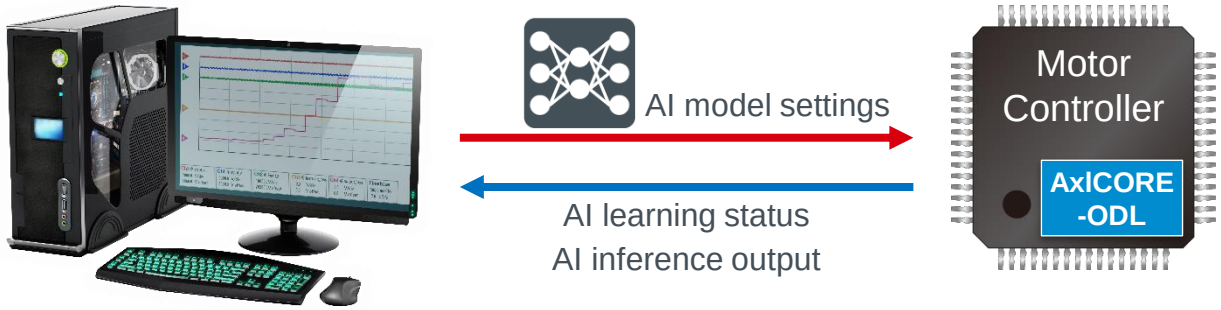
Motor Control Environment with AI Function



- Real-time high-speed inference
- On-device learning

Easily introduce AI functions at a lower cost to conventional motor controller ICs (without additional components)

Tools are also under development to easily achieve everything from AI model construction to evaluation



No need to design complex models or adjust numerous parameters

- Easily build AI models while monitoring AI output
- Tune AI models with minimal parameters (no. of input data, thresholds for anomaly determination)
- Relearn on the device at the touch of a button

Expected Usage Case 2:

General-Purpose MCU for Edge with On-Device Learning AI Function

General-Purpose MCU ← **Add-on endpoint AI**

- General-purpose MCU with peripheral on-device learning AI accelerator
- High-speed AI operations are performed by small hardware, while application functions can be freely implemented with software

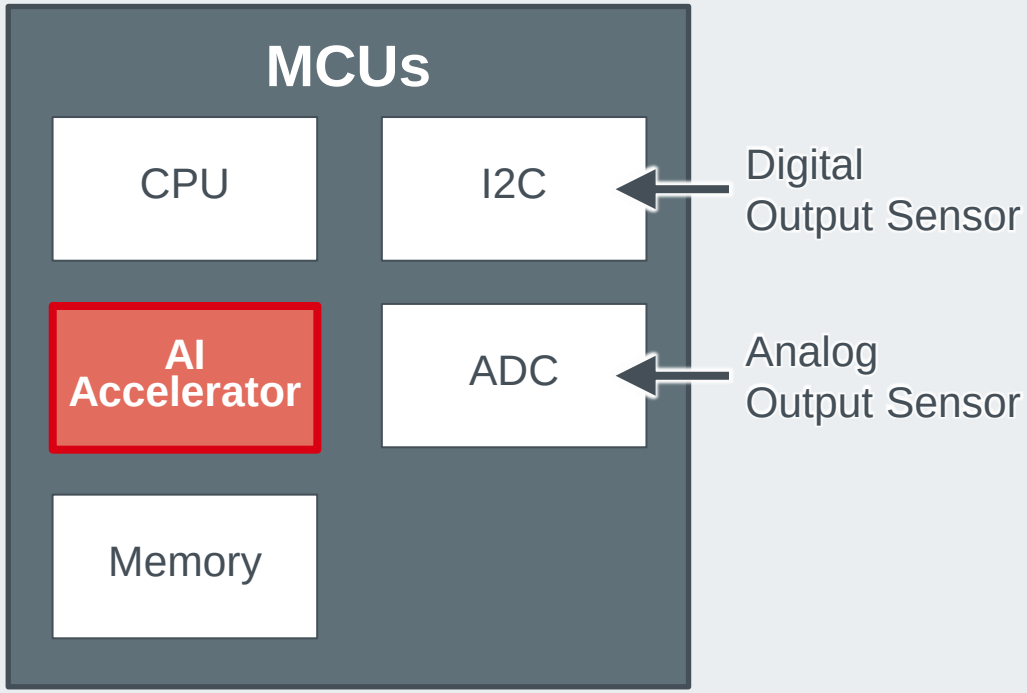
Easily add AI functions (i.e. predictive failure detection) to edge/endpoint MCUs in industrial equipment, automotive systems, home appliances, and more.

Advantages

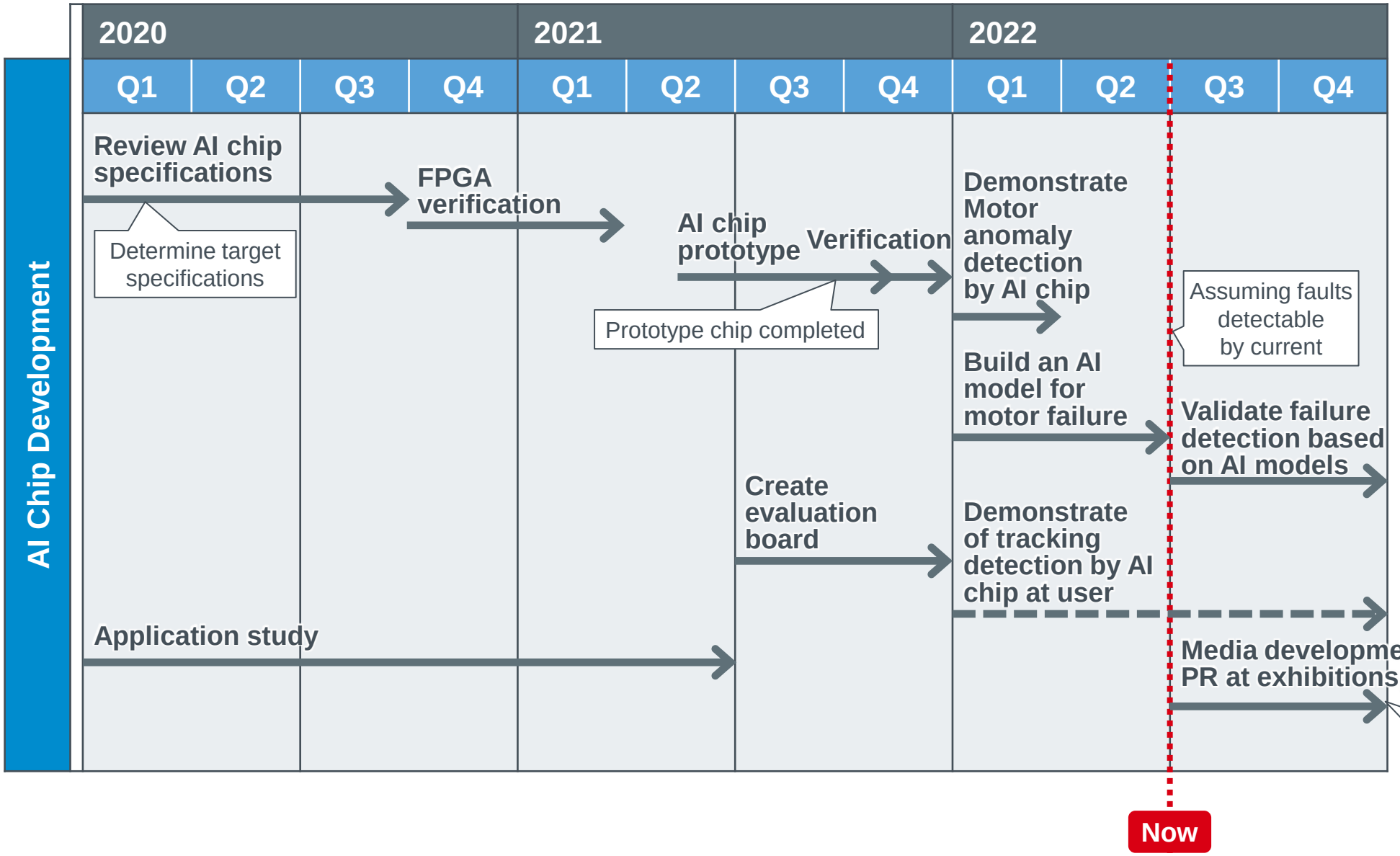
- As all necessary AI operations are computed with hardware, there is less load on the software with no limitations on application functions
- AI functions can be added by replacing existing application MCUs
- Carrying out learning and inference on the device side facilitates optimization at each installation location

Enables simultaneous processing of various types of sensor inputs

Acceleration, current, temperature, brightness, microphone



Schedule for Prototype Chip Development and Future Commercialization



Future Schedule

FY2023
Begin development of product chips for mass production

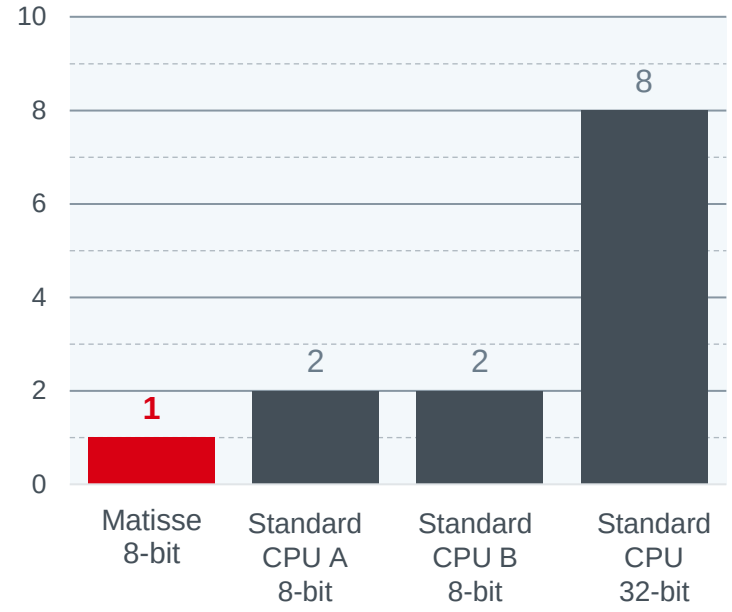
FY2024
Start mass production

Contribute to failure prediction/predictive maintenance by incorporating into ICs for motors and sensors

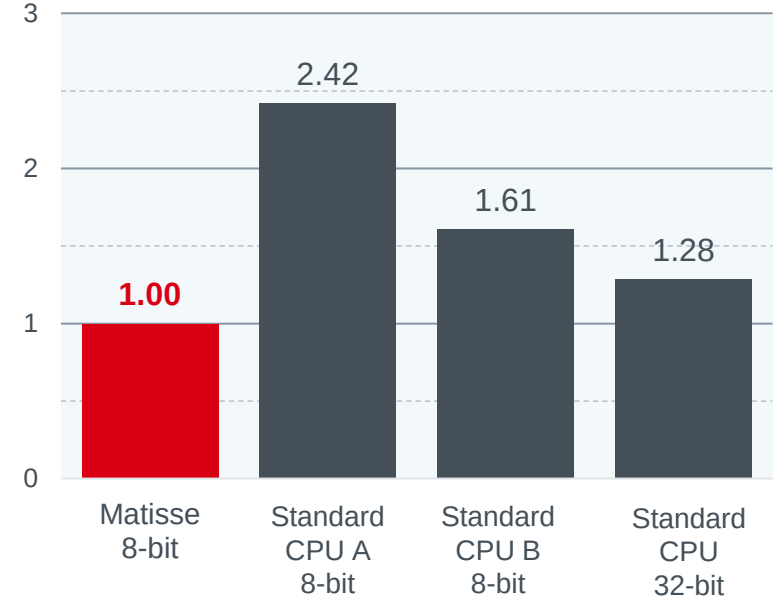
~ Compact and High-Speed 8-bit CPU ~

Performance Comparison vs General Compact CPUs (with Matisse set to 1)

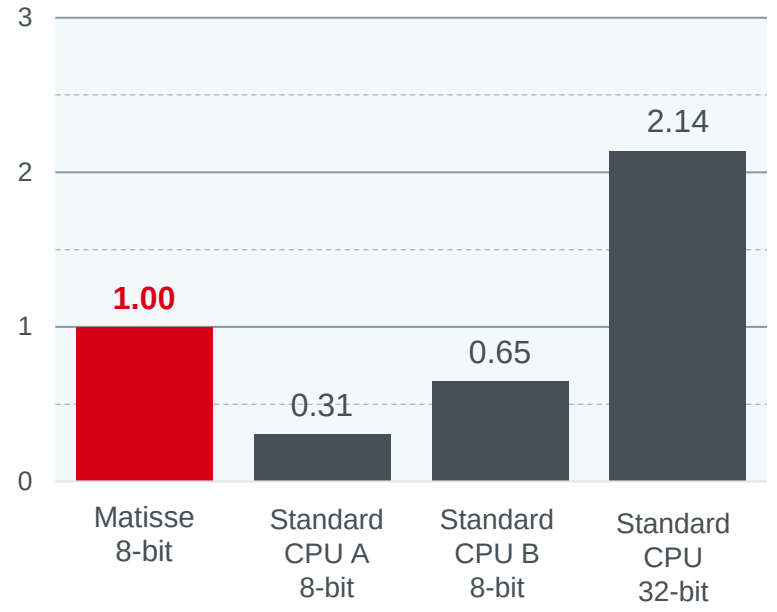
Gate Size Comparison



ROM Size Comparison (With 8bit calculation program)



Processing Performance Comparison (Dhrystone)



Excellent area savings

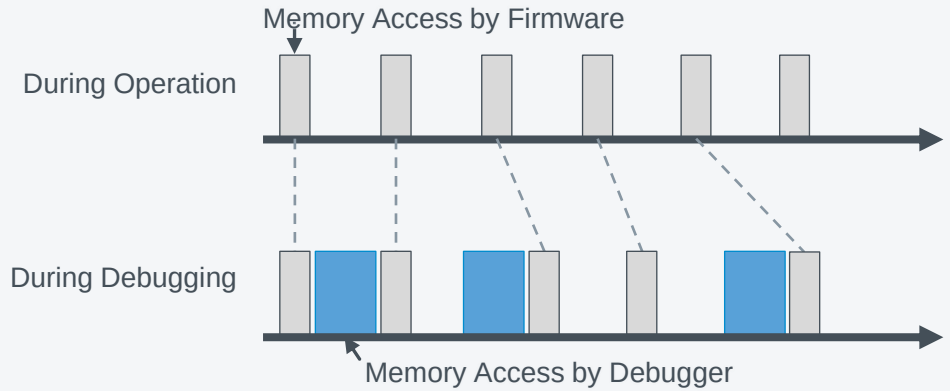
Compact program code size

High-speed arithmetic processing

Matisse offers a more compact chip area, smaller program code size, and fast arithmetic processing (and can be adapted to automotive ASIL-D as process)

~ Real-Time Debugging Functionality Ideal for Embedded Use ~

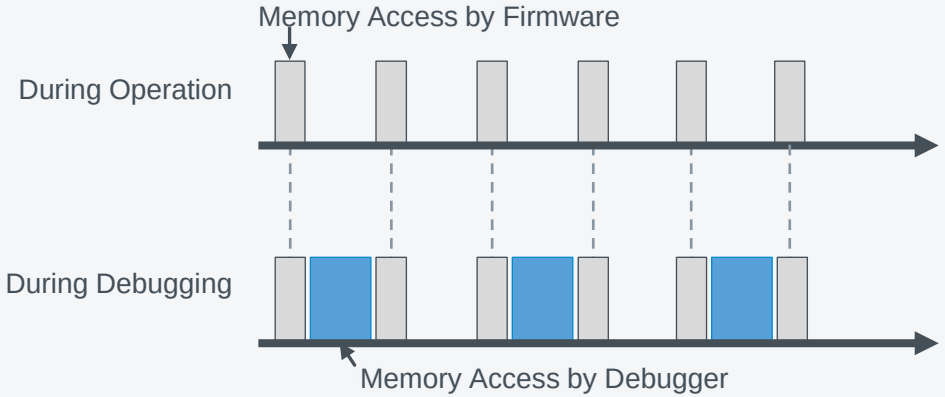
Conventional Method



Memory access by debugger prevents firmware operation

Debugging can significantly change program behavior

Matisse's Real-Time Debugging Capabilities

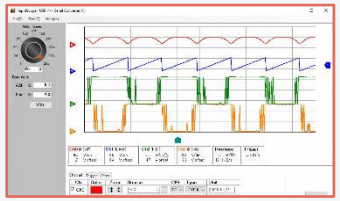


Adjustable debugger memory access eliminates interference with firmware operation

Debugging can be performed without changing program behavior

Matisse can retrieve/modify internal CPU information in a completely lossless manner

- Real-time debugging ensures that **operation is the same during normal operation as during debugging**
- **Enables easy debugging of applications that cannot be stopped and debugged**, such as motor control
- **Variables within the IC** that cannot normally be seen **can be extracted in real time and displayed as waveforms**



Waveform display software RapidScope™

Electronics for the Future

ROHM Co.,Ltd.

Notes

- The content specified herein is for the purpose of introducing ROHM's products (hereinafter "Products").
- If you wish to use any such Product, please be sure to refer to the specifications, which can be obtained from ROHM upon request.
- Great care was taken in ensuring the accuracy of the information specified in this document. However, should you incur any damage arising from any inaccuracy or misprint of such information, ROHM shall bear no responsibility for such damage.
- The technical information specified herein is intended only to show the typical functions of and examples of application circuits for the Products.
- ROHM does not grant you, explicitly or implicitly, any license to use or exercise intellectual property or other rights held by ROHM and other parties.
- ROHM shall bear no responsibility whatsoever for any dispute arising from the use of such technical information.
- If you intend to export or ship overseas any Product or technology specified herein that may be controlled under the Foreign Exchange and the Foreign Trade Law, you will be required to obtain a license or permit under the Law.
- The content specified in this document is correct as of November 2022.