Realizing High-Speed Deep Learning Inference with AI Edge Computer "AE2100"

Takamitsu Shimada

AI technology represented by deep learning has been rapidly evolving in recent years supported with faster processors and more advanced memory and storage. Additionally, the combination of AI with maturing sensor and sensor networking technologies is advancing the R&D of diverse and captivating systems. As a result, expectations for solving social issues through such technological research have increased. On the other hand, the growing popularity of IoT devices and the accompanying increase in data are increasing the load on networks and the cloud. Furthermore, security threats such as DDoS attacks from malware that infect IoT products are increasing as well.

Amid these circumstances, dedicated AI chips have recently appeared that enables AI edge computing on small terminals with low power consumption. AI edge computing will ensure immediate response, load distribution, and high reliability while in collaboration with the cloud enable data analysis and distribution of updated trained models. It is expected to be the driving force that further promotes social implementation of AI technology.

This article describes the commercialization background and product overview of OKI's new AI Edge Computer AE2100 (**Photo 1**).

Photo 1. AI Edge Computer AE2100

Commercialization Background

Conventionally, processing using AI (hereinafter referred to as AI processing) has been performed in the cloud. However, as the volume of data to be processed grew, the required resources and processing time increased. In order to solve this problem, it is necessary to reduce the burden on the cloud and shorten the processing time by transferring the AI processing performed in the cloud to the edge devices.

There are five advantages to AI edge computing in which AI processing is performed at the edge devices.

(1) High reliability
(2) Real-time processing
(3) Load distribution
(4) Privacy protection
(5) Low communication load

To make these advantages possible, OKI has commercialized the AE2100, an AI edge computer that can perform general-purpose AI processing at the edge with trained models created through various frameworks.

AE2100 Overview

The AE2100 is an AI edge device ideal for performing AI processing of large volumes of sensor data and images, or for performing deep learning inference processing, which is required for real-time and reliable AI processing at the edge.

The AE2100 supports various communication protocols enabling it to work in conjunction with the cloud to provide advanced AI solutions such as visualization of AI processing results at the cloud or batch updates of AE2100 trained models (**Figure 1. left**).

Furthermore, AI processing and highly real-time control such as factory equipment monitoring in an on-premises environment can even be performed on the AE2100 (**Figure 1, right**).

The AE2100 possesses the industry's first computer architecture equipped with Intel's OpenVINO™*1) toolkit, which provides a deep learning inference environment, and a VPU (Vision Processing Unit), a hardware AI accelerator, to provide a high-speed deep learning inference processing environment.

To accommodate various sensors, the device supports

*1) Intel, Intel Atom, Movidius, Myriad and OpenVINO are trademarks of Intel Corporation or its subsidiaries in the United States and/or other countries.

a variety of interfaces and communication protocols including LTE*2), wireless LAN, and OKI's 920MHz multi-hop wireless SmartHop® *3). Therefore, it is possible to accommodate sensor devices that meet the customer's application needs.

Additionally, the device has acquired Microsoft Azure*4) IoT Edge certification and can be used for high value-added solutions in conjunction with cloud services. Moreover, there is a plan to support 5G, which is expected to spread in the future.

Security required for IoT devices has become increasingly important in recent years, and the AE2100 is standardly equipped with security functions such as secure boot, Trusted Platform Module (TPM), and access control functions to sufficiently guard it against malicious attacks from networks.

As described, the AE2100 can accommodate a variety of sensors, operate the AI inference engine for general purposes, and with the outdoor casing described later, it has environmental performance that can withstand harsh outdoor environments. For this reason, it can be used universally in a variety of customer AI use cases.

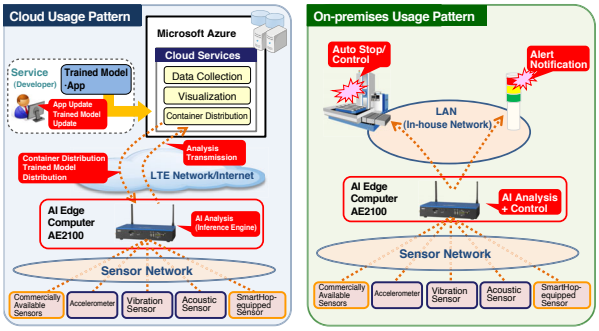

Figure 1. AI Edge Computing System Configuration

Hardware Features

The specifications of the AE2100 are shown in **Table 1**.

The AE2100 can be chosen from one of the three base models (LAN, LTE and wireless LAN versions) depending on the interface required by the customer's network environment.

The CPU utilizes Intel's Intel Atom®*1) x7-E3950 processor. It is equipped with 32GB (eMMC) of internal storage and a SDXC (UHS-I) card slot is provided for additional storage.

For connection of sensor devices, the AE2100 is equipped with Ethernet*5), USB, serial (RS-232C and RS-485) and contact interfaces (**Figure 2**).

The following options can be installed according to the customer's application needs.

1) SmartHop

OKI's SmartHop is a self-managed IoT wireless technology that supports 920MHz wireless and multi-hop functions to enable long-distance transmissions without requiring an operating license. It can wirelessly accommodate sensors and remote I/O devices equipped with SmartHop that are commercialized by other companies.

The customer can choose between the SmartHop MH series, which is ideal for wireless sensor networks in factories and buildings, and the SmartHop SR series, which is battery-powered and suitable for applications such as those involving social infrastructures where securing power can be difficult.

2) AI accelerator

For hardware AI accelerator option, the Intel® Movidius™ Myriad™*1) X VPU can be selected. This option equips the AE2100 with two Intel Movidius Myriad X VPU chips.

As an AI accelerator for the edge, the Intel Movidius Myriad X VPU has high deep learning inference performance. The inference engine included in the OpenVINO toolkit described later makes full use of the CPU's built-in GPU and Intel Movidius Myriad X VPU. Inference performance is 25 times faster than conventional edge computers equipped only with a CPU. As a result, it is possible to execute a high-level inference process which would conventionally require a server to perform.

3) Heat radiation fin

If the AE2100 is to be installed outdoors or in a factory where the temperature environment is poor, choosing the heat radiation fin option can expand the operating ambient temperature conditions, thus allow the AE2100 to be adapted to various installation environments.

4) Outdoor casing

Environment resistant casing is available should the AE2100 need to be installed at social infrastructures or other outdoor installations. Customer can choose between IP55 (dustproof) and IP66 (waterproof) compliant casings.

*2) LTE is a registered trademark of the European Telecommunications Standards Institute (ETSI). *3) SmartHop is a registered trademark of Oki Electric Industry Co., Ltd.
*4) Microsoft and Azure are trademarks or registered trademarks of Microsoft Corporation in the United States and other countries.
*5) Ethernet is a registered trademark of Fuji Xerox Co., Ltd.

Table 1. AE2100 Specifications

	LAN Version	LTE Version	Wireless LAN Version
CPU	Intel Atom x7-E3950 processor (4 core/1.6GHz)		
Memory	DDR3L 4GB		
Storage	32GB (eMMC) standard / SDXC(UHS-I)×1		
Wired NW	1000BASE-Tx2 (communication use×1 / maintenance use×1)		
LTE	–	LTE supported	–
Wireless LAN	–	–	IEEE802.11b/a/g/n/ac 2×2 supported
920MHz Wireless	SmartHop-equipped (MH series or SR series)*1		
USB	USB 2.0×2		
Serial	RS-232C (D-sub 9-pin)×1/RS-485×1		
Contact	Input×1, Output×1		
AI Accelerator	Intel Movidius Myriad X VPU (2 chips) *1		
Operating Temperature and Humidity	-20 ~ 60°C, 10 ~ 90% RH (no condensation) *2		
Waterproof / Dustproof	IP40 equivalent / IP55·IP66 (with outdoor casing)		
Security	TPM2.0 equipped		
Power	Main unit: DC12V / AC adapter: AC100V		
Dimensions	W250×D156×H47.5mm (excluding heat radiation fins, antennas, screws and other protrusions)		
Weight	1.5kg (excluding heat radiation fins, antennas, etc.)		
Certifications	Radio Act, Telecommunications Business Act		
OS	Yocto Linux 2.5.1		

*1: Factory installed options.
*2: Without options installed. Will vary depending on options installed.

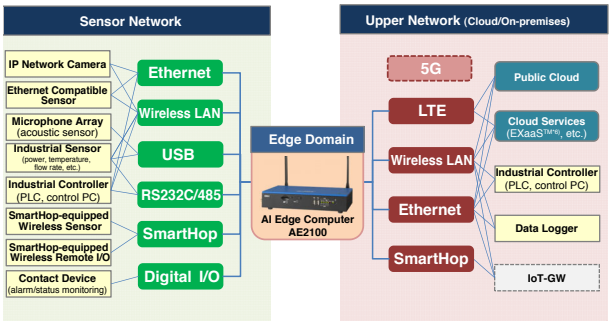

Figure 2. AE2100 Interfaces

Software Features

Figure 3 shows the software architecture of the AE2100.

The AE2100 utilizes embedded Linux®*7) (Yocto Linux) and is equipped with Docker and Azure IoT Edge runtimes that provide a virtualization environment, and a WebUI/SDK that enables configuration and operation of AI edge computers.

A standard container is provided to run on the AE2100 Docker. The customer installs trained models and applications into this container for operation on the AE2100. The container includes Intel's OpenVINO toolkit, an open AI execution environment.

In order to perform deep learning inference on the AE2100, a trained model is created with a standard framework such as TensorFlow, and this is converted to an intermediate model with the model optimizer included in the OpenVINO toolkit before installation into the container. Since the model optimizer supports a standard framework, the customer's trained model can be directly converted to run on the AE2100.

Azure IoT Edge ensures the container containing the trained model is compatible with the distributions and updates from the Azure Cloud. Therefore, even after the operation is started, additional training can be performed in the cloud, and the updated trained model distributed to the edge to improve the function and accuracy of the inference processing.

If the customer wishes to perform processing without using the OpenVINO toolkit, the container prepared by the customer can also run on the AE2100, thereby making effective use of the customer's existing software assets.

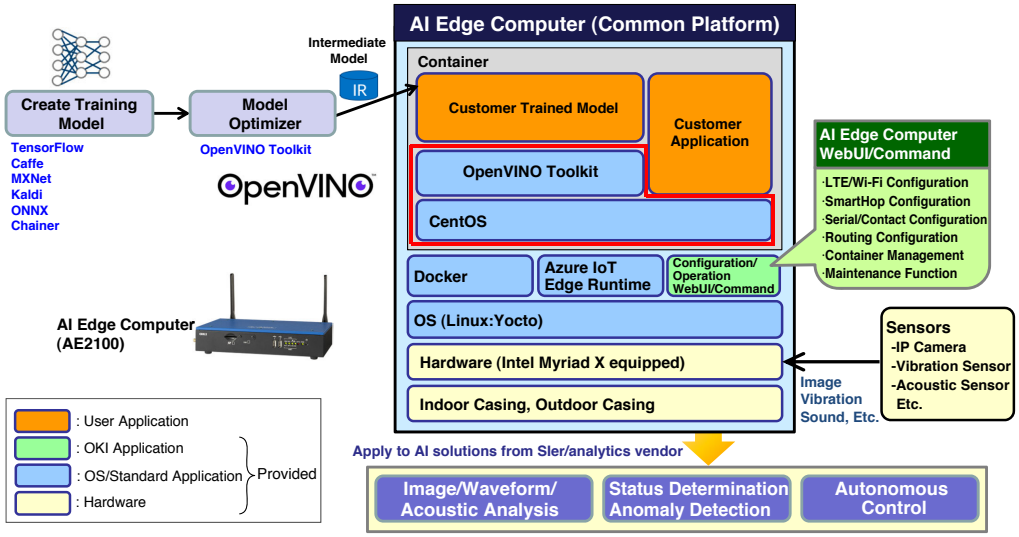

Figure 3. AE2100 Software Specifications

*6) EXaaS is a trademark of Oki Electric Industry Co., Ltd. *7) Linux is a registered trademark of Linus Torvalds in the United States and other countries.
Other company names and product names in this text are, in general, trademarks or registered trademarks of their respective companies.

AE2100 Use Cases

1) Image recognition for next-generation traffic

From images of an outdoor camera installed at an intersection, AE2100 AI processing will measure the position and speed of vehicles, pedestrians, etc. to immediately recognize road conditions. As a result, traffic conditions on the road can be quickly grasped and the information can be used to support safe driving and operation of autonomous vehicles.

Through the installation of AE2100 at multiple intersections, it will be possible to grasp traffic conditions and manage safety for an entire region, thus realizing next-generation traffic in an advanced IoT society.

2) Efficiency improvement through anomaly detection using vibration/acoustic data

The AE2100 AI processing can be applied to vibration and acoustic data acquired from sensors installed on a manufacturing equipment to improve productivity of the equipment and efficiency of maintenance.

For example, using the waveform analysis library installed in the AE2100, the status of equipment can be grasped in real time. When an anomaly is detected, a stop command is issued to the equipment, thereby the equipment can be immediately stopped for inspection and proper measures can be taken. As a result, the man-hours required for visual confirmation and cost of repairs can be significantly reduced.

3) Determining sudden natural phenomenon

Sensor devices that combine solar-powered cameras and sensors with low-power wireless communication technology are placed in various locations. Performing AI processing on the data collected from the devices, the on-site AE2100 will determine locations where sudden natural phenomenon is likely to occur. Not only can the communication cost from the site to the monitoring center be reduced, but the results of system's determination can be collected as the phenomenon occurs, making it possible to make quantitative decisions such as evacuation advisories and road closures. If necessary, sensor data and camera images can be obtained for detailed confirmation and decision making.

Ecosystem in AI Edge Domain

OKI has built an ecosystem in the IoT sensor network domain as a sales strategy for the 920MHz multi-hop wireless SmartHop. With the launch of the AE2100, OKI will build a new ecosystem in the AI edge domain in addition to the IoT ecosystem.

In order to meet the customers' problem solving and digital transformation needs at the AI edge domain, OKI is promoting "AI Edge Partnership" with AI-related business partners including SIer and AI vendors providing AI solutions, device vendors providing sensor devices, and distributors handling AI/IoT products.

In "AI Edge Partnership," partners will work together to expand the AI edge market, create various solutions using AI edge computers, and pursue activities to capture business opportunities.

Conclusion

OKI has commercialized an AI edge computer AE2100 that incorporates OpenVINO, a toolkit to provide a deep learning inference environment, and VPU, a hardware AI accelerator, thus enabling high-speed deep learning inference processing.

The AE2100 can be universally used for AI solutions in the edge domain allowing it to provide solutions that suits OKI's focus areas of "transportation," "construction/infrastructure," "disaster prevention," "finance/distribution," "manufacturing" and "marine" fields. OKI is also promoting co-creation with a wide range of partners through the ecosystem to create solutions that meet customer needs.

Using AI edge computing with AE2100 as the core, OKI plans to solve various social issues and contribute to the realization of an advanced IoT society. ◆◆

References

1) Yusuke Takahashi: Anomaly Detection using Vibration Analysis with Machine Learning Technology for Industrial IoT System, OKI Technical Review, Issue 230, Vol.84, No.2, pp.30-33, December 2017
2) Yuhiko Fujiwara: Approach to “Social Infrastructure x IoT” for SDGs, OKI Technical Review, Issue 232, Vol.85, No.2, pp.6-9, December 2018

Authors

Takamitsu Shimada, Smart Communications System Department, IoT Platform Division, ICT Business Group

TIPS【Glossary】

OpenVINO (Open Visual Inference & Neural Network Optimization) toolkit
Free software provided by Intel for computer vision and deep learning. It contains a model optimizer that converts and optimizes learning models, an inference engine, and a library for computer vision.

Microsoft Azure IoT Edge
Service provided by Microsoft that enables the deployment and execution of Microsoft Azure services, AI functions, and custom functions on IoT devices.

WebUI
Interface for configuring devices via a Web browser. AE2100 configuration and management functions can be operated from a Web browser.

TPM(**Trusted Platform Module**)
A security function developed by the TCG (Trusted Computing Group), an international industry organization, that can be used to detect software tampering and authenticate terminals as difficult to impersonate.