Embedded AI empowers predictive maintenance and anomaly detection

By Knut Dettmer, Renesas Electronics

With its flexible and scalable e-AI concept, Renesas offers a future-proof real-time, low power artificial intelligence processing solution that is unique in the industry and addresses the specific needs for artificial intelligence in embedded devices at the endpoint.

The evolution of artificial intelligence (AI) technologies such as machine learning and deep learning has been remarkable in recent years. The range of applications is rapidly expanding from cloud-centric applications mainly focused on the IT field to the embedded systems market. AI enables embedded devices to dynamically react and adapt to changes in the operating environment, and to adapt to the constantly changing state and condition of a device or machine.

The trend to move AI processing from centralized cloud processing platforms to the endpoint is motivated by a variety of reasons. First, bandwidth constraints limit the capability to deliver the required data from the observation point to a cloud processing unit to process the related analytics. For many devices and machines, there may even be no cloud connection available. This may be motivated by infrastructure restriction or by data privacy concerns. Still, one might well like to enjoy the many benefits from using AI technology to improve product performance and the overall equipment efficiency. Secondly, even if cloud connectivity with enough bandwidth is available, many AI applications are required to infer data in real-time, within milli- or microseconds. With connectivity technology today, this is not achievable. A cloud connection is unreliable and non-deterministic in terms of latency and introduces more than several dozens of milliseconds delay. Thirdly, data privacy is another motivation to process the analytics at the embedded endpoint and not in the cloud. Many industry segments regard the processed data itself as proprietary and are sensitive to share this data outside their own network. For example, healthcare devices collect personal data about individual health that must be highly restricted in terms of data distribution. In industrial automation, the analyzed data reveals how processes are controlled in a factory and is considered core know-how of the production company. Finally, data protection laws place a lot of restrictions on how end user data can be stored and processed.

Other advantages of working at the endpoint include the possibility to create hierarchical networked AI systems that are robust and scalable while being optimized on an individual use case in terms of performance and power consumption. When looking at these issues, the need for efficient AI inference at the embedded endpoint becomes obvious. It demands efficient endpoints that can infer, pre-process and filter data in real-time. All to optimize device performance and analyze the respective application-specific data points directly at the endpoint, while avoiding all the aforementioned constraints.

A common goal for AI is to improve the overall equipment efficiency, which targets maximizing the device availability, performance and output. Using e-AI methods from Renesas, one can implement predictive maintenance measures that continuously analyze the state and condition of a device to indicate necessary maintenance before the performance of the device degrades, avoiding unplanned downtime at the same time. Through these measurement analytics, the time for maintenance can be optimized individually for a specific device or machine. That results in a dynamic individual maintenance plan which is much more cost-efficient than operating on a static maintenance plan.

It is important to understand that embedded AI processing typically means inference processing. In AI terminology, inference is the process in which captured data is analyzed by a pre-trained AI model. In embedded applications, applying a pre-trained model to a specific task is the typical generic use case. In contrast, creating a model is called training (or learning) and requires a completely different scale of processing performance. Therefore, training is typically done by high performance processing entities, often provided by cloud services. Depending on model complexity, training a model can take minutes, hours, weeks, or even months. e-AI process-

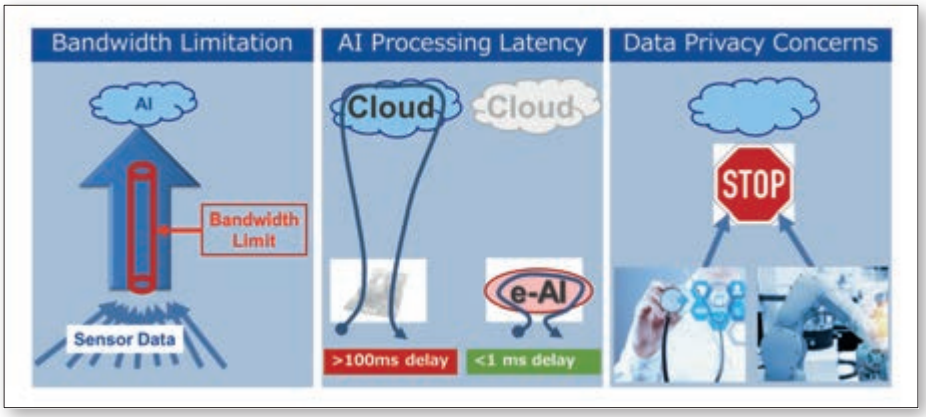

Figure 1. Motivation for embedded AI in the endpoint

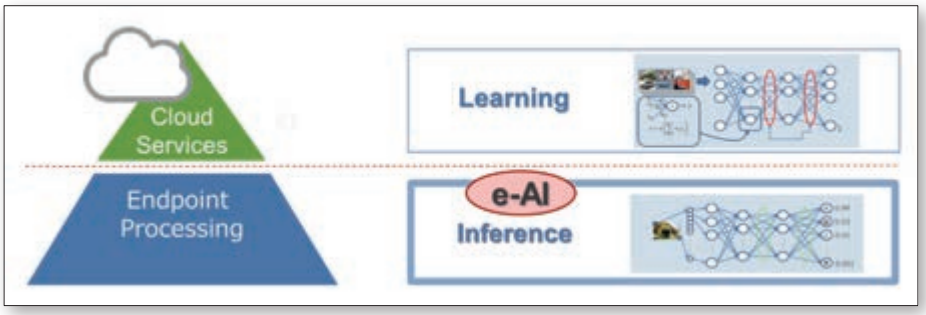

Figure 2. Learning vs Inference

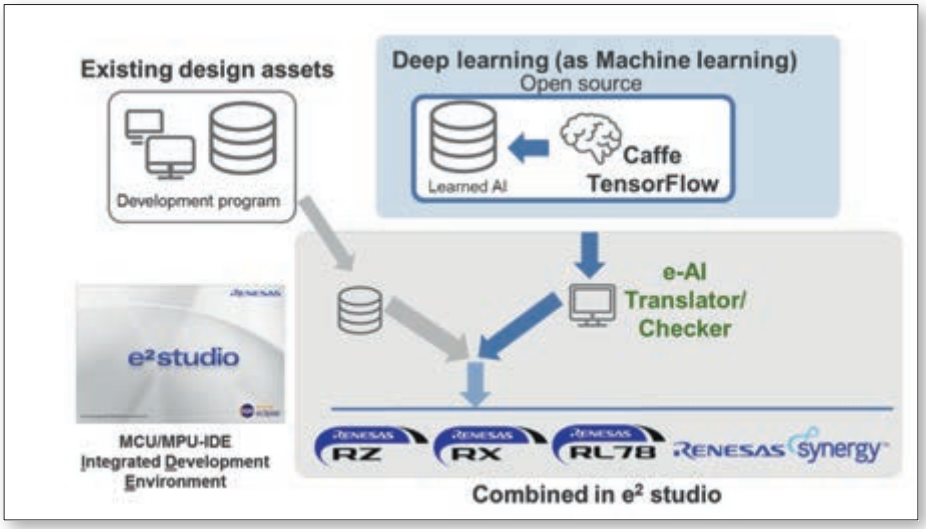

Figure 3. Embedding neural network processing onto Renesas devices

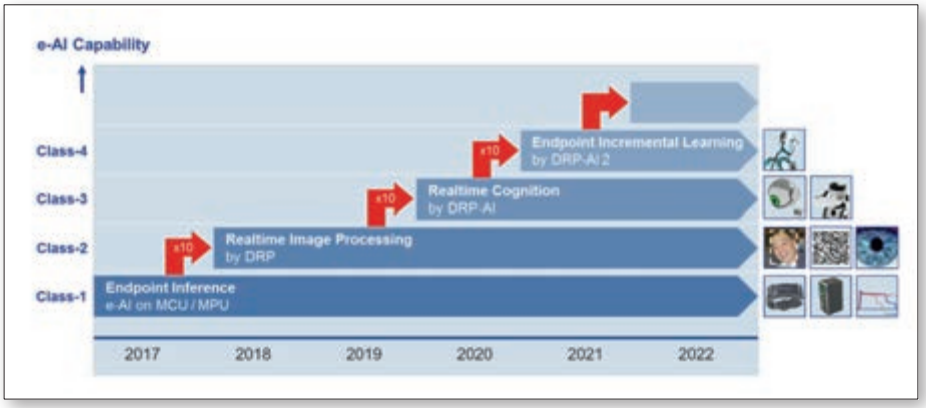

Figure 4. Renesas e-AI capability index

ing does not normally attempt to tackle these kinds of model creation tasks. Instead, it will help to improve the performance of a device using pre-trained models. Taking advantage of the data generated by rapidly increasing data received from sensors, e-AI can ensure that the devices output operate at the ideal state, whether in an industrial drive, a washing machine or a footstep detector. This is where Renesas focusses - endpoint intelligence.

As a leading semiconductor manufacturing company, Renesas has implemented such mechanisms in their factories. The technologies developed from these activities put Renesas in a position to share them with our customers and therefore enable our customers to enjoy all their benefits. More concretely, the company has implemented anomaly detection and predictive maintenance algorithms based on neural network architectures to optimize performance of plasma edging machines in its Naka factory. The results were so convincing that further proof of concepts was implemented with a variety of customers and partners. For example, GE Healthcare Japan Hino Factory is utilizing an AI unit which is based on the Renesas e-AI technology for improving their productivity. Our partner Advantech provides this AI unit as easy retrofit option to implement e-AI technologies for existing machines or devices.

When moving AI processing to the endpoint, power performance becomes of prime importance. An off-the-shelf graphics card or smartphone accelerator will exceed the power and size limitations of industrial applications by many times. Plus, solutions nowadays need to have a platform approach and must scale depending on the application requirements. For instance, some basic algorithms might be processed simply by the software of a microcontroller. Some others may need some basic hardware accelerators. Still others might need significant hardware acceleration to meet the algorithms performance targets.

To address this variety of performance targets, Renesas has developed a dynamically reconfigurable processor (DRP) module, that can flexibly assign resources to accelerate e-AI tasks. DRP uses highly parallel processing to meet modern AI algorithm demands. As the DRP design is optimized for low power consumption, it can support multiple use cases of customized algorithms with rapid inference which will fit most embedded endpoint requirements. This DRP enables high performance at low power, so that it fits ideally into embedded applications. It is dynamically reconfigurable, thus allowing adaption of use cases and/or algorithms within the same hardware.

The good news is that whatever level or acceleration is utilized, the software tools and interfaces will stay the same. Renesas does not provide its own learning frameworks but provides tools that translate neural network models into a format that can be executed on its MCUs and MPUs. The translator takes neural network models from common training frameworks such as Google TensorFlow as an input. These frameworks are typically used to train a neural network model. The inference will be executed by the Renesas MCU/MPU devices, simply by embedding the output of the translator tools into the respective program. It is as easy as that. The initial e-AI/DRP roadmap shows four performance classes, each class adding ten times of neural network performance to the previous class. The unique positioning for endpoint inference in combination with Renesas MCU/MPU technology gives users an unmatched power/performance ratio for AI processing. ■