Edge computing modules add AI capabilities to industrial processing and critical infrastructure

Sponsored by: SECO

Figure 1. Today, the trend in IIoT is implementing ever-evolving, advanced machinery equipped with advanced solutions. Source: Ipopka/Adobe Stock

Today, cloud-based computing allows massive resources to be brought to bear in a range of applications, often using artificial intelligence (AI) for data processing and analysis.

Today, cloud-based computing allows massive resources to be brought to bear in a range of applications, often using artificial intelligence (AI) for data processing and analysis. At the same time, transporting data from where it's collected and used — for instance, the edge — to the cloud and back introduces issues including latency and privacy concerns. One also has to consider the eventuality of remote resource outages, potentially crippling a system that can't function without its cloud overseer.

To help mitigate these challenges, especially in critical industrial and enterprise tasks, engineers may instead choose to implement edge intelligence for time-critical processing. Cloud hardware can still be used as part of an AI implementation, but typically this would involve non-time critical tasks, such as overall data processing and trend analysis.

New chips expand edge AI processing benefits

As the number of deployed internet of things (IoT) devices has exploded over the past decade, this has brought a multitude of benefits to both enterprise users and consumers alike. The basic concept of IoT is that a massive network of small — physically, and in terms of computing power — devices offload processing to a central server, allowing each node to perform well above its individual capability.

The IoT paradigm has found fertile ground in the industrial setting, where connecting machines and systems — creating an industrial IoT (IIoT) ecosystem — allows admins to take advantage of the data that "silent" machines have been producing for years to enhance manufacturing and industrial processes. Today, the trend in IIoT is implementing ever-evolving high-performance machinery equipped with advanced sensors and control electronics, increasingly capable of supporting business intelligence.

The increased sophistication of IIoT networks, however, can lead to latency, network availability and data security issues. Next-generation industrial machinery in particular is producing a large amount of data at a high frequency. Data file sizes (e.g., IMU, audio

Sponsored by:

SECO
Endless ways to the future

Produced by:

GlobalSpec

and image sensors) can be massive, and thus cannot be sent to the cloud, which means they must be processed at the edge with power and performance-efficient processing hardware.

Previous generation industrial CPUs and GPUs do not have the capability to handle this type of massive data analysis, but the upcoming generation of AI chips can take on these tasks locally. With this local computing power, AI chips can process and catalog large amounts of data without (or before) sending it to the cloud. Processing results can be returned directly to machinery, allowing for a quicker response, while reducing transports costs and congestion.

In one potential example, cameras and sensors at various points in the manufacturing process can perform real-time quality inspections, thus immediately detecting defects, reducing waste and improving overall product quality. Also, predictive maintenance can be performed locally by deploying AI algorithms on edge devices installed on industrial machinery. Edge AI can analyze equipment data in real-time, predicting potential failures and optimizing maintenance schedules to prevent costly downtime.

New edge AI chips are in constant development, driven by trends in the consumer market including enhanced computing power, data analytics and real-time capabilities. The carryover into industrial and infrastructure includes options from Intel, NXP, Rockchip, MediaTek and Axelera AI.

For example, NXP's i.MX 8M Plus line of application processors feature a 2.3 trillion operations per second NPU providing customers the option to perform machine learning inference directly on the edge, reducing cloud dependency and latency. NXP's system is appropriate for complex neural network functions like object detection and surveillance. NXP's eIQ inference engine can work with platforms like TensorFlowLite, ONNX Runtime and more for model quantization and conversion.

The 13th Gen Intel Core processors for IoT Edge offer more cores and memory compared to previous generations for compute-intensive edge use cases. Advanced features such as AI hardware acceleration with Intel Distribution of OpenVINO toolkit and Intel Deep Learning Boost push the limit for industrial projects: AI-based industrial process control (AIPC), vision systems, programmable logic controllers and autonomous mobile robots just to name a few.

As another example, AI processing units (APUs) from Axelera AI boast performance up to a staggering 214 trillion operations per second. This APU enables a massive amount of AI computing power at the edge, opening up applications that would have previously required cloud processing and its inherent latency.

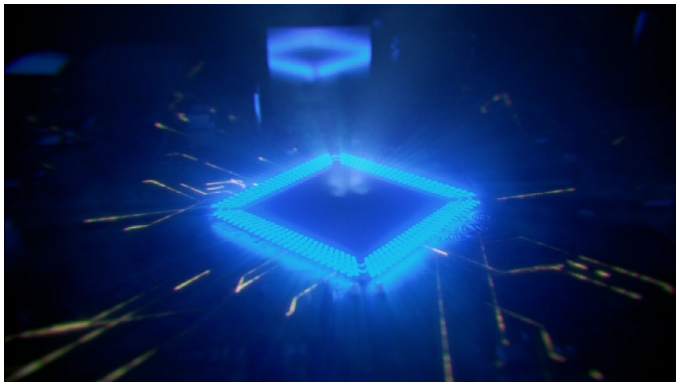

Figure 2. AI chipsets allow for incredible processing abilities, especially when partnered with sophisticated hardware for real-world applications. Source: SECO

Enhanced AI deployment and management

The progressive shift from cloud to edge AI processing, accelerated by advancements in hardware platforms design, is the logical solution for AI-enabled smart applications that need to respond in real time and with low latency. With more and more data being analyzed, processed and stored at the edge, new approaches are rising to deliver what have traditionally been cloud computing capabilities to edge nodes and terminals.

Consider, for example, the need for devices to improve from experience, accessing production data and using it to learn from themselves and adjust actions accordingly. That's where machine learning models come into play. Edge AI allows placing artificial intelligence and machine learning (ML)-powered applications physically closer to data sources to gather insights, identify patterns and initiate actions without relying on traditional cloud networks. While most ML models aren't designed for the edge, as they are processor and memory hungry, they require optimization to suit the capabilities of the specific hardware. This involves a process of fine-tuning the model's parameters, architecture and hyperparameters to harmonize with the computing resources and memory constraints of the edge device, which ensures that the model attains peak performance, minimizes latency and maximizes accuracy.

In the pursuit of efficiently deploying AI models onto resource-constrained edge devices, the technique of model quantization emerges as a beacon of efficiency. Quantization entails the conversion of high-precision floating-point model parameters into lower-precision fixed-point formats, effectively reducing memory and computational requirements. Quantized AI models can operate effectively on various hardware architectures, including CPUs, GPUs and specialized hardware accelerators. This is so that inferences can be made more quickly and power-effectively while preserving high accuracy and performance.

Advances in hardware design make these operations easier, thanks to greater computational capacity than previous generations of chips, and facilitate the execution of Edge Machine Learning Operations (MLOps) techniques. Combining DevOps principles and practices with ML tools, MLOps automate the processes of building, testing, deploying and monitoring ML models. This ensures that the ML model is optimized for production use and can be continuously improved as new data is received. While a major challenge to proper implementation of MLOps is the limited resources of edge devices, latest generation ones are rather well positioned. Not only are they capable of hosting large models, but they are often equipped with specialized processors optimized to run inference. Such processors are also usually backed by SDKs and toolset for artificial intelligence that can be integrated into an MLOps pipeline to facilitate the model deployment process.

Edge AI in manufacturing

Edge AI merges AI capabilities with edge computing, promising advantages in privacy, security and speed. AI is moving to the edge due to latency concerns and the need for real-time decision-making.

Using Edge AI means lowered latency, as data is processed directly on devices, enabling faster real-time decision making. This is crucial in mission-critical manufacturing applications, as slowing down a process for decision making is unacceptable. Keeping processing on-side can increase reliability, also crucial in manufacturing, since immediate dependence on network connectivity and the cloud is reduced or eliminated.

As data doesn't need to be transported off-site, Edge AI reduces bandwidth and storage costs. Power consumption is also reduced, especially important for battery-powered applications. The option to store data locally lessens the risk of data breaches and unauthorized access, increasing data privacy and security. Even though it may not be sent to the to the cloud immediately, edge devices do facilitate capturing a massive amount of production data. This data can thus be accessed as needed for AI model retraining and continuous improvement.

Simplified integration with SECO computing modules

AI chipsets allow for incredible new abilities processing-wise, but they need carrier boards and associated hardware to interface with real-world applications. SECO's wide range of computing modules, single-board computers and full embedded computing packages lets system designers skip the step of creating this basic level of interface hardware. They can instead concentrate on AI software and hardware requirements unique to their particular application, rather than reinventing the proverbial computing wheel.

SECO can also help with the software side of AI implementations. SECO's Clea package is a full-featured IoT platform, enabling edge MLOps, edge and cloud orchestration, AI model deployment, monitoring and validation of updates. Clea offers an excellent way to manage devices from the edge to the cloud with a focus on security. SECO stands ready to assist with developing appropriate AI implementations for customers, potentially involving customized hardware or software as needed.

Contact SECO today to learn how they can help advance your infrastructure.

SECO

Via Achille Grandi, 20
Arezzo, 52100 Italy
Tel: 39 0575-26979

GLOBALSPEC

257 Fuller Road
Suite NFE 1100
Albany, NY 12203
USA
Tel: (518) 880-0200

ABOUT SECO

SECO is a high-tech company that develops and manufactures cutting-edge solutions for the digitalization of industrial products and processes. SECO's hardware and software offering enables B2B companies to introduce edge computing, Internet of Things, data analytics and artificial intelligence in their businesses. SECO's technology spans across multiple fields of application: serving more than 450 customers, operating in sectors like Medical, Industrial Automation, Fitness, Vending, Transportation and many others. Enabling to accurately monitor the functioning of on-field devices, SECO solutions contribute to creating low environmental impact business models thanks to a more efficient use of resources.