ADVANTECH

Enabling an Intelligent Planet

Solution Brief

Enabling Edge AI Inference with Compact Industrial Systems

NVIDIA ELITE PARTNER

Executive Summary

Artificial intelligence (AI) technology breaks away from static rule-based programming, substituting inference systems using dynamic learning for smarter decisions. Advanced AI technology combined with IoT technology is now redefining entire industries with smart applications.

AI + IoT

Smarter, mission-critical applications:

- Agriculture
- Material Handling
- Cities
- Monitoring
- Healthcare
- Robotics
- Manufacturing
- Transportation

An important trend across these industries is the shift of AI inference systems toward the edge, closer to sensors and control elements, reducing latency and improving response. Demand for edge AI hardware of all types, from wearables to embedded systems, is growing fast. One estimate sees unit growth at 20.3% CAGR through 2026, reaching over 2.2 billion units. *

The big challenge for edge AI inference platforms is feeding high bandwidth data and making decisions in real-time, using limited space and power for AI and control algorithms. Next, we see how three powerful AI application development pillars from NVIDIA are helping Advantech make edge AI inference solutions a reality.

* Data Bridge Market Research, Global Edge AI Hardware Market – Industry Trends and Forecast to 2026, April 2019.

Solution Highlights

Advantech brings decades of real-world experience in industrial computing to AI inference systems featuring:

- Full range of NVIDIA® Jetson™ system-on-modules, each combining Arm® processor cores with NVIDIA GPU cores
- Comprehensive software infrastructure including the NVIDIA JetPack SDK, enabling many machine learning frameworks and models, and several multimedia, scientific, and computer vision libraries.

Designed for mission-critical operation, easy deployment, and long lifecycles, use cases for these edge AI inference platforms include IoT sensor and control processing, multi-channel image classification, motion control, and more smart automation needs.

What is AI Inference?

There are two types of AI-enabled systems: those for training, and those for inference. Training systems examine data sets and outcomes, looking to create a decision-making algorithm. For large data sets, training systems have the luxury of scaling, using servers, cloud computing resources, or in extreme cases supercomputers. They also can afford days or weeks to analyze data.

The algorithm discovered in training is handed off to an AI inference system for use with real-world, real-time data. While less compute-intensive than training, inference requires efficient AI acceleration to handle decisions quickly, keeping pace with incoming data. One popular option for acceleration is to use GPU cores, thanks to familiar programming tools, high performance, and a strong ecosystem.

Traditionally, AI inference systems have been created on server-class platforms by adding a GPU card in a PCIe expansion slot. Most AI inference still happens on AI-enabled servers or cloud computers, and some applications demand server-class platforms for AI acceleration performance. Where latency and response are concerns, lower power embedded systems can scale AI inference to the edge.

Edge AI inference supports what is happening today in an application – and looks ahead months and years into the future as it continues learning.

Advantages of Edge AI Inference Architecture

Edge computing offers a big advantage in distributed architectures handling volumes of real-time data. Moving all that data into the cloud or a server for analysis creates networking and storage challenges, impacting both bandwidth and latency. Localized processing closer to data sources, such as preprocessing with AI, can reduce these bottlenecks, lowering networking and storage costs.

There are other edge computing benefits. Personally identifiable information can be anonymized, improving privacy. Security zones reduce chances of a system-wide breach. Local algorithms enforce real-time determinism, keeping systems under control, and many false alarms or triggers can be eliminated early in the workflow.

Extending edge computing with AI inference adds more benefits. Edge AI inference applications scale efficiently by adding smaller platforms. Any improvements gained by inference on one edge node can be uploaded and deployed across an entire system of nodes.

If an edge AI inference platform can accelerate the full application stack, with data ingestion, inference, localized control, connectivity, and more, it creates compelling possibilities for system architects.

Extended lifecycle support

NVIDIA Jetson modules come in commercial versions with 5-year guaranteed availability, and select industrial versions offer 10 years.

NVIDIA Jetson Nano

NVIDIA Jetson TX2

NVIDIA Jetson Xavier NX

NVIDIA Jetson AGX Xavier

Flexibility of CPU+GPU Engines for the Edge

NVIDIA developed the system-on-chip (SoC) architecture used in the NVIDIA Jetson system-on-module (SoM). As applications for them grew, these small, low power consumption SoCs evolved with faster Arm® CPU cores, advanced NVIDIA GPU cores, and more dedicated processor cores for computer vision, multimedia processing, and deep learning inference. These cores provide enough added processing power for end-to-end applications running on a compact SoM.

AI inference can be implemented many ways. There are single-chip AI inference engines available, most with 8-bit fixed point math and optimized for a particular machine learning framework and AI model. If that framework and fixed point math works, these may do the job.

Many applications call for flexible CPU+GPU engines like those on Jetson modules. With AI models ever changing, accuracy, a choice of frameworks, and processing headroom are important. Inference might need 32-bit floating point instead of 8-bit fixed point math - precision experiments on a CPU+GPU engine are easy. If research suggests an alternative inference algorithm, GPU cores can be reprogrammed easily for a new framework or model. As control algorithms get more intense, a scalable multicore CPU handles increased workloads.

Pillar 1: Scalable System-on-Modules for Edge AI

From entry-level to server-class performance, NVIDIA Jetson modules are the first of three pillars for edge AI inference. Sharing the same code base, Jetson modules vary slightly in size and pinout, with features like memory, eMMC storage, video encode/decode, Ethernet, display interfaces, and more. Summarizing CPU+GPU configurations:

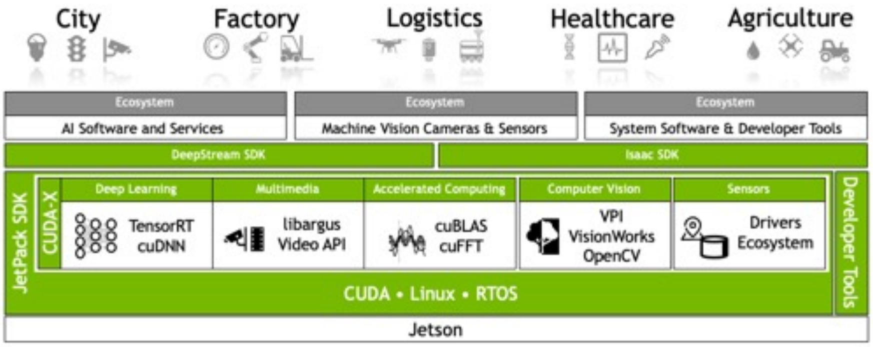
	Jetson Nano	Jetson TX2 Series	Jetson Xavier NX	Jetson AGX Xavier
AI Inference	472 GFLOPS	1.33 TFLOPS	21 TOPS	32 TOPS
GPU	128-Core NVIDIA Maxwell GPU	256-Core NVIDIA Pascal GPU	384-Core NVIDIA Volta GPU, with 48 Tensor Cores	512-Core NVIDIA Volta GPU, with 64 Tensor Cores
CPU	Quad-Core Arm Cortex-A57	Dual-core NVIDIA Denver 2 and Quad-Core Arm Cortex-A57	6-core NVIDIA Carmel Arm-v8.2	8-core NVIDIA Carmel Arm-v8.2

For a complete comparison of NVIDIA Jetson module features, visit: nvidia.com/en-us/autonomous-machines/embedded-systems/

Pillar 2: SDK for Edge AI Inference Applications

The second pillar converts a large base of NVIDIA CUDA® developers into AI inference developers, with a software stack running on any NVIDIA Jetson module for "develop once, deploy anywhere".

The NVIDIA JetPack SDK runs on top of L4T with an LTS Linux kernel. It includes accelerated libraries for cuDNN and TensorRT frameworks, as well as scientific libraries, multimedia APIs, and the VPI and OpenCV computer vision libraries.



JetPack also has a NVIDIA container runtime with Docker integration, allowing edge device deployment in cloud-native workflows. It has containers for TensorFlow, PyTorch, JupyterLab and other machine learning frameworks, and data science frameworks like scikit-learn, scipy and Pandas, all pre-installed in a Python environment.

Developer tools include a range of debugging and system profiling tools including CPU and GPU tracing and optimization. Developers can quickly move applications from existing rule-based programming into the Jetson environment, adding AI inference alongside control.

For a complete description of NVIDIA Jetson software features, visit: developer.nvidia.com/embedded/develop/software

Pillar 3: Ecosystem Add-Ons for Complete Solutions

The third pillar is an ecosystem of machine vision cameras, sensors, software, tools, and systems ready for AI-enabled applications. Over 100 partners work within the NVIDIA Jetson environment, with qualified compatibility for easy integration. For example, several third parties work on advanced sensors such as lidar and stereo cameras, helpful for robotics platforms to determine their surroundings.

Two more SDKs for AI developers

NVIDIA DeepStream SDK
DeepStream SDK is a streaming analytics toolkit for AI-based IoT sensor processing, including video object detection and image classification.

NVIDIA Isaac SDK
Isaac SDK includes building blocks and tools for creating robotics with AI-enabled perception, navigation, and manipulation.

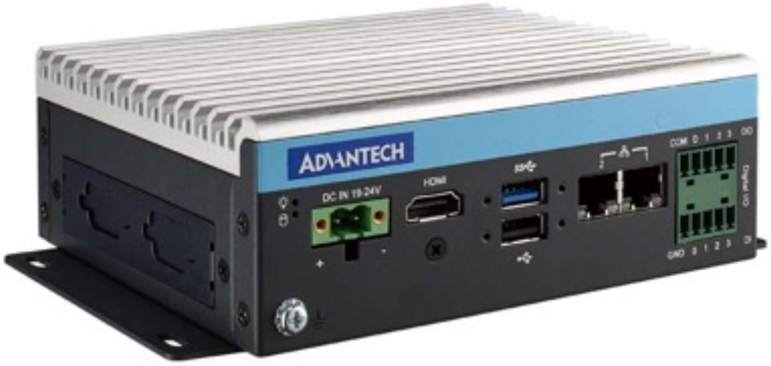
Systems for Mission-Critical Edge AI Inference

Many edge AI inference applications are mission critical, calling for small form factor computers with extended operating specifications. Advantech created the compact MIC-700AI Series systems, targeting two different scenarios with full range of performance options.

Adding AI inference to control

The first scenario is the classic industrial computer, with a rugged form factor installed anywhere near equipment requiring real-time data capture and control processing. These scenarios often have little or no forced air cooling, only DC power available, and DIN rail mounting for protection against vibration.

For this, the MIC-700AI series brings AI inference to the edge. Designed around the low-power NVIDIA Jetson Nano using advanced thermal engineering, the fanless MIC-710AI operates on 24VDC power in temperatures from -10° to +60°C. With an M.2 SSD, it handles a 3G, 5 to 500 Hz vibration profile.



Advantech MIC-710AI AI Inference System

The MIC-710AI features two GigE ports, one HDMI port, two external USB ports, and serial and digital I/O. For expansion, Advantech iDoor modules are mPCIe cards with cabled I/O panels (cutout seen on left side above). iDoor modules handle Fieldbus, wireless, and more I/O.

Advantech PCM-24S2WF iDoor Module with Wi-Fi and Bluetooth

Longevity and revision control

With extended availability of all components, Advantech offers a 5-year lifecycle on all MIC-700AI Series platforms.

Additionally, system revision notification is standard, with full revision control services available.

From off-the-shelf to customized

Advantech can handle local sourcing and integration needs, and customization including bezels, I/O, power, and other needs, creating solutions ready for customer resale. For most needs, there is no minimum order quantity (MOQ).

Mid- to high-performance image classification

The second scenario involves machine vision and image classification, where cameras look for objects or conditions. Systems often use Power over Ethernet (PoE) to simplify wiring.

At the high end with the NVIDIA Jetson AGX Xavier, the MIC-730IVA provides eight PoE channels for connecting industrial video cameras. It also provides two bays for 3.5" hard drives, enabling direct-to-disk video recording. The system runs from 0° to 50°C, using AC power.

Advantech MIC-730IVA 8 Channel AI Network Video Recorder

A portfolio of AI-enabled solutions

All MIC-700AI Series systems run the same software, enabling developers to move up or down and get applications to market faster.

NVIDIA Jetson Module	AI Inference + Control	Image Classification
Jetson Nano	MIC-710AI/MIC-710AIL	MIC-710IVA
Jetson TX2 Series	MIC-720AI	-----
Jetson Xavier NX	MIC-710AIX/MIC-710AIXL	MIC-710IVX
Jetson AGX Xavier	MIC-730AI	MIC-730IVA

The latest MIC-710AIL features a Jetson Nano or Jetson Xavier NX in an ultra-compact enclosure, also with iDoor module expansion

Advantech MIC-710AIL AI Inference System (Lite)

These systems bring AI inference to the edge in reliable, durable platforms ready for a wide range of applications including manufacturing, material handling, robotics, smart agriculture, smart cities, smart healthcare, smart monitoring, transportation, and more.

For more information on Advantech Edge AI systems, visit: https://www.advantech.com/products/edge-ai-system/sub_9140b94e-bcfa-4aa4-8df2-1145026ad613

Advantech Contact Information

Hotline Europe: 00-800-248-080 | Hotline USA: 1-800-866-6008
Email: skyserver@advantech.com
Regional phone numbers can be found on our website at:
http://www.advantech.com/contact/
https://www.advantech.com/nc

NVIDIA, CUDA, and Jetson are trademarks and/or registered trademarks of NVIDIA Corporation in the U.S. and/or other countries.

Arm and Cortex are trademarks or registered trademarks of Arm Limited (or its subsidiaries) in the US and/or elsewhere.