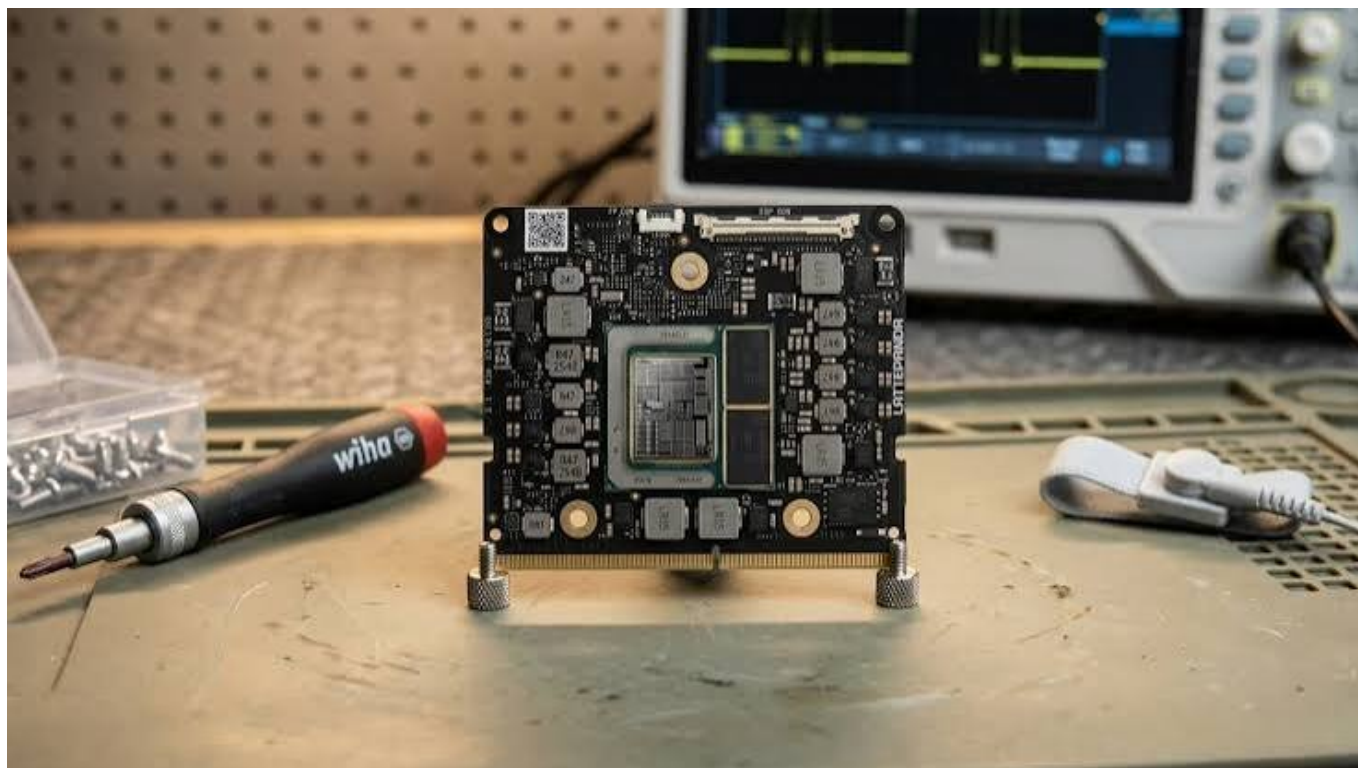


NEWS | TECHNOLOGY

LattePanda Launches Mu Ultra Micro x86 Module with Up to 115 TOPS for On-Device AI

Powered by Intel Core Ultra processors, the compact compute module delivers high-performance local AI, low-power operation and x86 compatibility for robotics, industrial automation, vision systems and intelligent edge applications.

By **Ankit Kankar** | September 14, 2026



Picture Courtesy | Public Domain

LattePanda Team launched LattePanda Mu Ultra, a micro x86 compute module purpose-built for on-device AI. Powered by Intel Core Ultra 5 226V and Ultra 7 256V processors, it delivers up to 115 TOPS of AI performance in a compact 69.6 × 60mm form factor. Built to address the hardware, compute, and software challenges of running AI locally, LattePanda Mu Ultra brings powerful AI computing to engineering teams, system integrators, and OEMs across robotics, industrial automation, portable instruments, vision AI, and a broad range of intelligent edge applications.

115 TOPS of AI Performance in a Compact x86 Module

LattePanda Mu Ultra combines the CPU, GPU, and NPU of Intel Core Ultra 5 226V and Ultra 7 256V processors to deliver up to 115 TOPS (INT8) of AI performance, including 47 TOPS from the dedicated NPU.

Its 69.6 × 60mm form factor brings powerful x86 computing into space-constrained devices, from handheld instruments and smart cameras to robots and other intelligent edge systems.

Run AI Locally on the Device

LattePanda Mu Ultra runs AI workloads directly on the device, enabling low-latency inference, better data privacy, and less reliance on cloud computing.

In testing, LattePanda Mu Ultra achieved text generation speeds of 18 tokens/s with Qwen3.5-9B and 55 tokens/s with Qwen3.5-2B, using INT4 quantization with OpenVINO GenAI on the iGPU. These results demonstrate its ability to support responsive, local conversational AI without cloud inference. This makes it suitable for applications such as local voice assistants, offline document processing, and private knowledge retrieval, keeping sensitive data on the device.

High-Bandwidth Memory for AI Workloads

LattePanda Mu Ultra features 16GB of LPDDR5X-8533 memory, with up to 11.6GB available for graphics memory, providing the bandwidth required for demanding AI workloads across its CPU, GPU, and NPU.

The memory architecture is designed to support workloads ranging from LLMs and multimodal AI to real-time sensor processing. With idle power consumption as low as 2.5W, the module is also well suited to portable and always-on applications where power efficiency matters.

Bring On-Device AI to Existing x86 Systems

Deploying AI directly on devices involves more than choosing the right hardware. Ensuring software compatibility across the AI stack can also add significant development costs.

LattePanda Mu Ultra supports Windows and Linux and maintains compatibility with the established x86 software ecosystem. It also supports popular AI tools and frameworks, including Intel OpenVINO, llama.cpp, and Ollama.

This allows engineering teams and OEMs to build on existing x86 applications and development workflows, making it easier to add on-device AI capabilities without rebuilding their systems from the ground up.

Upgrade with a Modular Design

LattePanda Mu Ultra retains the standardized form factor and connector of the LattePanda Mu family and is largely compatible with existing carrier boards. When paired with the 3.5-inch Lite Carrier Board, Mini Carrier, or M.2 M Key Carrier Board, it enables users to upgrade existing systems without redesigning the entire carrier board.

With interfaces including PCIe 4.0, USB 3.2 Gen 2, USB 2.0, HDMI / DisplayPort, eDP, GPIO, UART, and I2C, LattePanda Mu Ultra can support everything from rapid prototyping and small-batch development to volume production.

Five Key Applications for On-Device AI

Intelligent AI Terminals

Run LLMs locally for translation devices, voice assistants, and private knowledge-based applications where low latency and data privacy are important.

Portable Instruments

Bring high-performance x86 computing to handheld devices such as spectrum analyzers and portable diagnostic equipment for AI-assisted, real-time analysis.

Autonomous Mobile Robots (AMRs)

Support real-time sensor fusion, SLAM, and autonomous navigation in a compact computing platform designed for space-constrained robots.

Service Robots

Run vision, voice, and LLM-based multimodal AI workloads on a single computing module, providing an integrated processing core for service robots.

On-Device Vision AI

Process high-resolution video locally in smart cameras and robotic inspection systems, enabling low-

latency defect detection while reducing data transmission and cloud computing costs.

"Getting AI to run in the cloud is relatively straightforward. The real challenge comes when developers need to deploy it in physical devices—where space is limited, power matters, and migrating existing software can be costly," said Youliang Yu, LattePanda product manager. "LattePanda Mu Ultra addresses these challenges with a compact x86 design, high AI performance, and low power consumption. By combining powerful local computing with compatibility across the existing x86 ecosystem, it gives developers a simpler path to integrate AI into real-world products without rebuilding their systems from the ground up."

Read online: <https://worldbusinesswatch.com/news/13/6/lattepanda-launches-mu-ultra-micro-x86-module-with-up-to-115-tops-for-on-device-ai.html>

© 2026 World Business Watch. All rights reserved.