AMD and XMPro Partner to Deliver 8-9x Faster, Secure and Predictable AI Inference at the Industrial Edge

Run secure LLMs locally with AMD Lemonade Server, powered by Ryzen™ AI, and XMPro's Composable Intelligence Platform.

XMPro and AMD Partner to Deliver 8-9x faster AI inference at the Industriual Edge

DALLAS , TX, UNITED STATES, July 15, 2025 /EINPresswire.com/ -- The new XMPro–AMD solution allows industrial organizations to eliminate cloud AI costs, reduce latency, and maintain full control over operational data by running powerful AI models directly at the edge using AMD's Lemonade Server and XMPro's composable intelligence platform.

XMPro, a leading industrial intelligence platform provider, has announced a partnership with AMD, a global leader in high-performance computing, to deliver on-site AI processing directly to industrial facilities. The collaboration integrates AMD's open-source Lemonade Server with XMPro's composable platform, enabling manufacturers to run advanced AI models on-site without cloud dependencies, latency issues, or unpredictable costs.

> This collaboration demonstrates the potential of AI at the industrial edge to drive measurable impact in mission-critical environments."
>
> *Pieter Van Schalkwyk - XMPro CEO*

The solution addresses three critical challenges facing industrial AI adoption: unpredictable cloud costs, data sovereignty risks, and network dependency. A typical automotive manufacturing plant processing 50,000 sensor readings per minute can face over $15,000 monthly in cloud AI API costs alone[1]. The new edge-based approach eliminates these recurring expenses while keeping sensitive operational data completely on-site.

"Our collaboration with XMPro is about more than just performance – it's about enabling real-time intelligence where it matters most," said Paul Hartke, Fellow, AMD Research and Advanced Development . "By combining AMD's industrial edge hardware with XMPro's composable AI platform, manufacturers can analyze and act on operational data instantly, without compromising security or data sovereignty. This is the next step in deploying practical, high-impact AI across the industrial landscape."

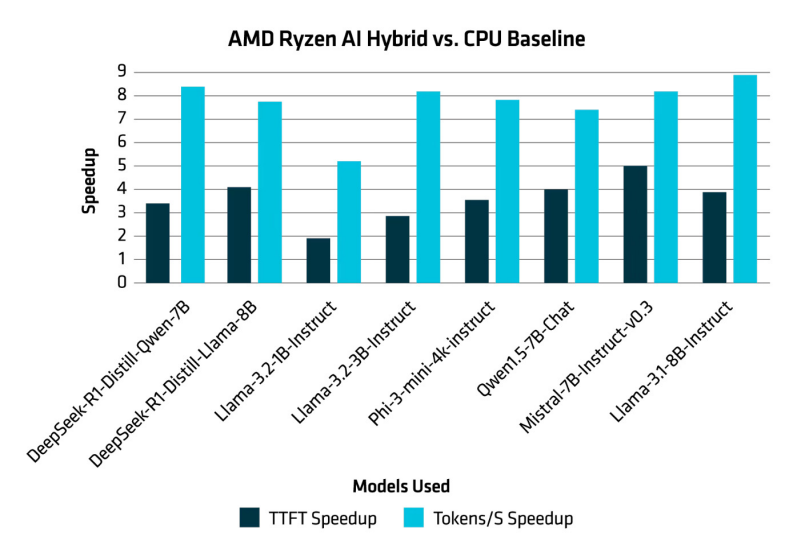

Figure 1. A graph showcasing the speed multiplier of AMD Ryzen™ Ai Hybrid (iGPU + NPU) vs a CPU Baseline for TTFT (time to first token) speedup and Tokens/S Speedup. The CPU baseline represents the “Speed Multiplier” 1 in the graph above.

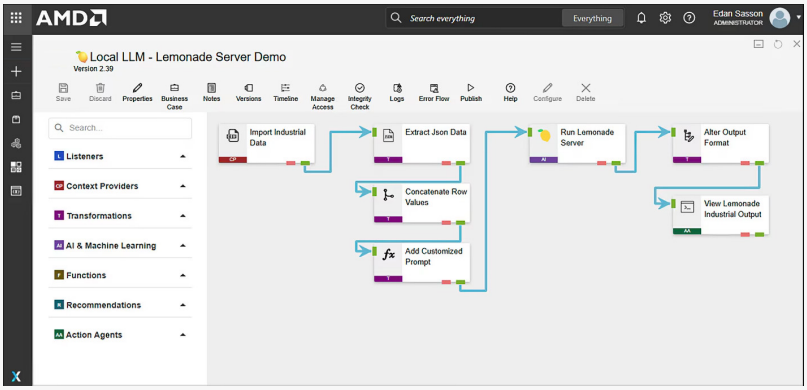

Figure 2. Demonstrating an example of how to use the newly integrated "Lemonade Server" component in XMPro. In this example, we analyze data from a pump and feed it into a local LLM running on Ryzen™ AI.

The integrated solution combines AMD's Lemonade Server, optimized for local large language model deployment, with XMPro's composable industrial intelligence platform. While AMD provides the high-performance local AI processing power, XMPro delivers the critical industrial context and system integration that transforms raw AI capability into actionable business value.

XMPro's platform enables companies to rapidly connect AI models to existing industrial systems—ERP, SCADA, IoT sensors—without disrupting current operations. Using drag-and-drop tools, engineering and operations teams can build data streams that unify information across their enterprise, create digital twins of assets and processes, and deploy AI-powered workflows that automatically detect anomalies, predict failures, and trigger intelligent responses.

Companies can now deploy comprehensive AI-powered decision support directly on their factory floor using AMD Ryzen™AI systems, which deliver up to 50 TOPS of AI compute performance through hybrid NPU and integrated GPU architecture.

Performance testing demonstrates significant advantages over traditional CPU-only setups. The

AMD Ryzen™ AI hybrid configuration achieves up to 8-9 times faster token generation and 3-4 times faster initial response time across industry-relevant models[2]. This performance improvement translates to faster AI decision-making on the factory floor, real-time responsiveness for AI-powered operational guidance, reduced hardware load and energy use during AI inferencing, and improved user experience for engineers and operators using AI assistants.

"By combining AMD's industrial edge compute capabilities with XMPro's composable AI platform, manufacturers can now run complex AI inference directly on-site, enabling faster decisions, greater operational autonomy, and complete control over sensitive data," said Pieter Van Schalkwyk, CEO at XMPro. "This collaboration demonstrates the potential of AI at the industrial edge to drive measurable impact in mission-critical environments."

The partnership delivers three key advantages that cloud-based AI cannot match:

• Complete Data Sovereignty: All AI processing occurs on-site, ensuring sensitive operational data never leaves the facility – critical for organizations with strict security requirements or regulatory compliance needs.

• Predictable Costs: One-time hardware investments replace recurring cloud API fees, providing cost certainty and eliminating budget surprises from variable usage.

• Zero Network Dependency: Local processing eliminates latency from cloud round trips and ensures AI capabilities remain operational even during network disruptions.

XMPro's role in the partnership extends beyond platform capabilities to solve the practical challenge of operationalizing edge AI in industrial environments. While many companies can deploy local AI models, XMPro provides the industrial-specific framework to make those models immediately useful for real-world operations. The platform's composable architecture allows organizations to create reusable AI workflows that can be rapidly deployed across multiple sites and adapted to changing business needs, turning edge AI from a technical capability into a scalable business solution.

The solution targets asset-intensive industries including manufacturing, energy, mining, utilities, and logistics, where real-time decision-making and data security are paramount. XMPro's industrial-focused design includes pre-built connectors for common industrial systems, templates for typical use cases like predictive maintenance and process optimization, and governance frameworks that ensure AI decisions remain traceable and auditable, critical requirements for regulated industries.

AMD's Lemonade Server is fully open-source and compatible with OpenAI APIs, ensuring easy integration with existing applications. The server runs efficiently on AMD Ryzen™ AI hardware, leveraging both the Neural Processing Unit (NPU) and integrated graphics processor (iGPU) for

maximum performance and energy efficiency.

The solution is available immediately for pilot deployments. Companies interested in evaluating the technology can contact AMD at lemonade@amd.com or visit XMPro's website for implementation guidance.

About AMD
AMD is a leading provider of high-performance computing solutions, including processors, graphics cards, and data center solutions. The company's Ryzen™ AI processors deliver up to 50 TOPS of AI compute performance through hybrid NPU and integrated GPU architecture, enabling efficient AI processing at the edge.
Learn more --> AMD x XMPro Blog Post

About XMPro
XMPro is an industrial intelligence platform that enables asset-intensive organizations to turn real-time data into decisions and actions. The company's composable architecture allows engineering, operations, and IT teams to rapidly build and scale AI-powered solutions without custom coding. XMPro serves manufacturing, energy, mining, utilities, and logistics companies worldwide, delivering proven operational improvements through intelligent automation and predictive analytics.
Learn More --> https://xmpro.com/amd/

---
[1] Estimate based on 50,000 sensor readings per minute × 60 minutes × ~30 days = 72 million readings per month. At common cloud AI inference rates of $0.20 to $1.25 per 1,000 predictions (as published by AWS, Azure, and Google Cloud), monthly costs range from $14,400 to $90,000. Sources: Google Cloud Vertex AI Pricing, AWS SageMaker Pricing, Azure Machine Learning Pricing.

[2] Performance benchmarks conducted on AMD Ryzen™ AI 9 HX 375 with hybrid NPU+iGPU configuration vs. CPU baseline. All validation, performance, and accuracy metrics collected on HP OmniBook Ultra Laptop 14z with AMD Ryzen™ AI 9 HX 375 W/ Radeon 890M, 32GB RAM, using ONNX TurnkeyML v6.1.1 software. Hugging Face transformers framework used as baseline implementation. All speedup numbers measured with input sequence length (ISL) of 1024 and output sequence length (OSL) of 64. Data collected March 28, 2025.

Wouter Beneke
XMPro - Marketing Lead
email us here

we have missed, please do bring it to our attention. Your help is welcome. EIN Presswire, Everyone's Internet News Presswire™, tries to define some of the boundaries that are reasonable in today's world. Please see our Editorial Guidelines for more information.