HighPoint Unveils High-Performance Gen5 AI Compute Platform for Ultra-High Inference Throughput Empowered by Hailo

Featuring the Rocket 1604L Gen5 Retimer technology, this integrated solution achieves 4144 FPS running YOLOv8n, with perfect linear scaling across 4x Hailo-8™ M.2 modules.

HighPoint Technologies, Inc., a pioneer in high-performance PCIe connectivity, today announced an ecosystem partnership with Hailo, a leading innovator in edge AI processors, to deliver a new class of high-performance compute expansion solution: the Enterprise Edge AI Compute Platform.
Purpose-built for high-end modular robotics and industrial AI, this collaboration introduces a Marketplace-Ready hardware platform that enables IT architects to deploy high-density AI clusters achieving a staggering 160 TOPS (INT4). By leveraging readily accessible Hailo-10H™ M.2 acceleration modules, the platform delivers unprecedented throughput, processing up to 4,144 FPS for real-time inference workloads, effectively eliminating the I/O bottlenecks currently facing next-generation autonomous systems.

The Synergy: Building a Localized AI Compute Platform

The Enterprise Edge AI Compute Platform supports both Hailo-8™ and Hailo-10H™ M.2 AI Acceleration Modules with the HighPoint Rocket 1604L Gen5 Retimer AIC. While a single Hailo-10H™ module delivers 40 TOPS, this integrated solution allows users to consolidate up to four modules into a single PCIe x16 slot, delivering a massive 160 TOPS of local inference performance.

Proven Performance: Linear Scaling & Precision

To validate the efficiency of the Gen5 AI Compute Platform, rigorous infrastructure benchmarking was conducted utilizing four Hailo-8 / Hailo-10™ M.2 modules on a single Rocket 1604L AIC. These results demonstrate the critical role of HighPoint's Active Retimer Architecture in maintaining total signal integrity and deterministic I/O across high-density configurations.
The platform achieved industry-leading performance benchmarks, including 4,144 FPS for YOLOv8n and 2,173 FPS for YOLOv5s (Batch size = 1) on the Hailo-8. This unprecedented throughput ensures that high-velocity data ingestion and real-time vision processing occur simultaneously across all four M.2 ports without signal degradation or device-to-device variance, providing a transparent and reliable foundation for mission-critical AI workloads.

- **Massive Throughput**: While a single Hailo-8™ module can achieve up to 1036 FPS, the quad-module AI Compute Platform can deliver a staggering 4100 + FPS, proving its capability for high-speed, multi-stream processing via single PCIe slot.

- **True Linear Scaling**: Performance scaled perfectly as additional modules were added, ensuring that hardware investments translate directly into compute power.
- **Unmatched Precision**: The benchmark recorded an FPS variance between modules of less than 0.01%, a testament to the bit-perfect signal reconditioning provided by the Rocket 1604L's Active Retimer Architecture.

Rocket 1604L: The Hardware Core of the Engine

The Rocket 1604L serves as the mechanical core of the AI Compute Platform, ensuring multi-module configurations operate at peak-rated specifications:

- **Active Retimer Architecture**: Proactively reconditions signals at 32GT/s for a bit-perfect host link.
- **Compact 167mm Footprint**: Allows deployment in standard 1U/2U industrial chassis where full-size GPUs cannot fit.
- **Thermal Stability**: Full-length aluminum heat sink and high-static-pressure fan assembly tuned for 24/7 heavy AI inference.

Marketplace-Ready Deployment

By utilizing standard M.2 form factors and PCIe Gen5 connectivity, HighPoint and Hailo have created a solution that VARs and System Integrators can source through mainstream marketplace channels, removing "Time-to-Market" barriers.

“The Enterprise Edge AI Compute Platform is about moving from experimenting with AI to executing AI at scale," said May Hwang, Director of Marketing at HighPoint Technologies. "By pairing Hailo's efficient processing with HighPoint’s Gen5 connectivity and proven 0.01% variance precision, we are giving architects a powerful, marketplace-ready blueprint to build future-proof localized intelligence hubs.”

"The collaboration with HighPoint demonstrates how scalable, high-performance edge AI can be achieved using standard, accessible hardware building blocks,” said Yaron Ofer, Sales GM at Hailo. “By combining Hailo’s industry-leading efficiency with HighPoint’s Gen5 connectivity, we’re enabling customers to deploy powerful, localized AI systems that deliver data-center-class performance directly at the edge, without the traditional constraints of power, cost, or form factor.”

Learn More

Rocket 1604L PCIe Gen5 x16 4x M.2 NVMe Retimer AIC
Solution: Gen5 AI Compute Platform
The Signal Integrity Revolution: Inside HighPoint’s Gen5 x16 Retimer Architecture

Availability

The HighPoint Rocket 1604L is available immediately through HighPoint's global network of distributors and VARs. Hailo-8™ and Hailo-10H™ M.2 modules can be sourced via Hailo's official marketplace and global distribution channels.

About HighPoint Technologies, Inc.
HighPoint Technologies is a pioneer in high-performance PCIe Gen5 connectivity, specializing in Active PCIe Fabric solutions that eliminate bottlenecks for the Modern Enterprise.