GW16168 M.2 NXP Ara240 DNPU AI Accelerator for Edge AI Applications

The Gateworks GW16168 is a low-power M.2 AI accelerator card designed to add dedicated AI processing capability to Gateworks VeniceFLEX and Catalina single board computers. Designed and manufactured in the USA, and powered by the NXP Ara240 Discrete Neural Processing Unit (DNPU), the module delivers up to 40 eTOPS of dedicated AI compute, enabling real-time execution of generative AI workloads, including Large Language Models (LLMs), Vision Language Models (VLMs) and transformer-based architectures.

The GW16168 utilizes a standard M.2 M-Key 2280 form factor and integrates via PCIe. The module includes 8 or 16GB of onboard LPDDR4 memory, allowing it to operate independently of host system memory. This architecture supports models up to 16 billion parameters, enabling deployment of AI workloads in power-constrained embedded systems.

Support for industry-standard frameworks, including TensorFlow, PyTorch and ONNX, enables integration into existing AI software environments. The Ara240 delivers high performance-per-watt, enabling efficient execution of edge AI workloads such as machine vision, intelligent automation, predictive maintenance, robotics and local LLM inference.

FEATURES

- NXP Ara240 Discrete Neural Processing Unit (DNPU)
  - Enables real-time AI processing and decision-making
  - Up to 40 equivalent Tera Operations Per Second (eTOPS)
  - Exceptional performance per watt inferences
  - AI-Model Frameworks Supported: TensorFlow, PyTorch, ONNX
  - Security: Secure boot and root of trust
  - Software SDK with extensible compiler for CNNs and LLMs
- 8 or 16GB of LPDDR4 DRAM for large model support
- M.2 2280 M-Key Form Factor
- PCIe Communication: PCIe Gen4 (4x) Signaling
  - Note: Gateworks SBCs typically have 1x PCIe Gen 3
- Compatible with VeniceFLEX, Catalina and other Gateworks SBCs with FSA sockets
- -30°C to +85°C Rated Parts
- Made in USA
- 1 Year Warranty



SPECIFICATIONS

Electrical

Input Voltage
- 3.3V From M.2 Edge Connector

Typical Operating Current
- 3.0A @ 3.3V

Mechanical

Dimensions
- 22x80x23mm (w/ heatsink)

Weight
- 0.22 oz (6.2g)

Environmental

Operating Parameters
- Temperature: -30°C to +85°C Rated Parts
- Humidity (non-condensing): 20% to 90%

Storage Parameters
- Temperature: -40°C to +85°C
- Humidity (non-condensing): 5% to 95%

Ordering Options

- GW16168: 16GB LPDDR4
- GW16168-SP520: 8GB LPDDR4

Development Kit
- Includes GW16168, GW9201 SBC, Power Supply and Accessories
- GW11062-1: 16GB Variant
- GW11062-2: 8GB Variant