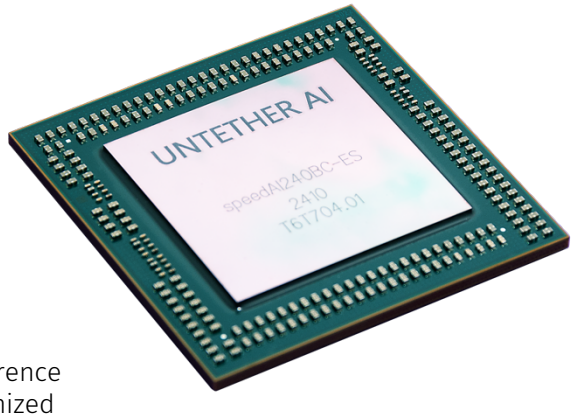
speedAI devices

UNTETHER AI

Unmatched efficiency and scalability with AI inference accelerators



Overview

The speedAI240™ device is designed for real-time deep learning inference applications. This second generation at-memory architecture is optimized for CNNs, attention networks, and recommendation systems.

Key Highlights

Efficiency Boosts Performance

The speedAI240 maximizes performance by dedicating every square millimeter to inference, achieving industry-leading 30 TFLOPS/W in 2000 FP8 TFLOPS.

Flexibility at Scale

With over 1,400 RISC-V cores, it offers coarse-grained control over 370,000 processing elements, interconnected by a high-speed NOC for smooth data flow.

Precision and Accuracy

Leveraging 8-bit floating point data types (FP8) and a custom network-on-chip, speedAI devices minimize data movement costs while ensuring accuracy.

Scalable Architecture

External DRAM supports large models with two LPDDR5 x64 interfaces, while four PCIe Gen5 interfaces facilitate chip-to-chip and card-to-card interconnects for building nodes capable of handling even the largest language models swiftly.

Target Applications

Vision/Object Detection
Smart Retail and City
AgTech
Aerospace and Defense
Finance, Enterprise Datacenter
Speech-to-text
Sentiment analysis
Chatbots
Automotive
Advanced driver assistance systems (ADAS)
In-Vehicle Experience (IVE)

Specifications

Package	Dimensions
Package dimensions	40x40 mm
Interconnect interface	PCIe Gen5 x16 Supports 160 GB/s total PCIe bandwidth
External Memory	Two LPDDR5 x64 interfaces, up to 64 GB capacity Offers 100 GB/s LPDDR5 bandwidth
On-Chip Memory	238 MB on-chip SRAM Up to 1 PB/s of on-chip memory bandwidth

UNTETHER AI

Performance

All estimates use the Untether AI FP8 data type.

Network	Single Chip
ResNet50	80,000
BERT-Large 384	10,000
YoloV7	5,000

For more information:
Visit: www.untether.ai/
Contact: www.untether.ai/about/contact/

