UNTETHER AI

runAI devices

Key Features

502 INT8 TOPs	204MB on-chip SRAM	75W TDP 40W typical	8 TOPs/W
At-Memory Architecture	Scalable voltage and frequency	Low latency, native batch = 1	PCIe Gen4 x16

Overview

The runAI200™ accelerator is designed for real-time deep learning inference and high-performance computing (HPC) applications. Its unique at-memory architecture combines over 260,000 processing elements, 511 custom RISC-V processors, and 204 MB of SRAM into the industry's most efficient chip in its class, delivering 8 TOPs/W. The imAIgine™ software development kit (SDK) enables push-button performance on deep learning networks in standard frameworks, and a custom kernel development flow for high performance computing applications that require arbitrary computation.

Applications

The runAI200 devices are designed to accelerate a multiplicity of AI inference and HPC workloads, such as vision-based convolutional networks, transformer networks for natural language processing, time-series analysis for financial applications, and general-purpose linear algebra for high performance computing applications.

Markets	Application	Networks
Vision	Classification, object detection, semantic segmentation	ResNets, YOLO, SSD, Unets, Pose
Natural language processing	Text-to-speech, speech-to-text, chatbots	RNNs, Transformers, BERT
Financial technology	X-Value adjustments, credit risk, portfolio balancing	TCNs, LSTMs
HPC	Climate modeling, deep packet inspection, simulations	FFTs, BLAS, arbitrary computation

imAIgine Software Development Kit

The imAIgine SDK gives developers powerful automated tools and supporting software to quickly go from pilot model to production. It is organized into three parts.

The imAIgine Compiler

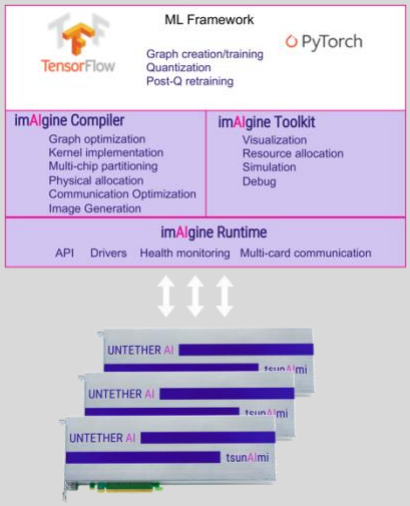
- Import TensorFlow, PyTorch, or ONNX graphs directly
- Automated quantizer and extracts performance without sacrificing accuracy
- Specify performance levels, silicon utilization, and power consumption targets

The imAIgine Toolkit

- Evaluate functionality and performance using the extensive profiling and simulation tools

The imAIgine Runtime

- Provides C-based API for integration into your deep learning environment
- Monitor the health and temperature of the tsunAImi® acceleration cards to ensure proper operation and prevent thermal damage



Familiar frameworks
Quantization and layer optimization done in familiar ML framework

Automated graph lowering
Optimization and allocation algorithms

Extensive feedback
Resource allocations, congestions, cycle-accurate simulation

Easily integrated runtime
Hardware abstraction, communication, and monitoring

Product Specification

Specification	runAI200 processor
Dimensions	47.5x47.5 mm
Process	16 nm
Power	75W TDP, 40W typical
PCIe interface	X16 PCIe Gen4
Clock frequency	Variable, up to 840 MHz
Memory	204 MB on-chip SRAM
Data types	INT8, INT16

Figures of Merit

Metric	runAI200 processor
Compute performance	502 INT8 TOPs
Compute efficiency	8 TOPs/W
On-chip memory bandwidth	251 TBps
System bandwidth	32 GBps via PCIe Gen4

UNTETHER AI