



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 imAagine™ 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

imAagine Software Development Kit

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

The imAagine Compiler

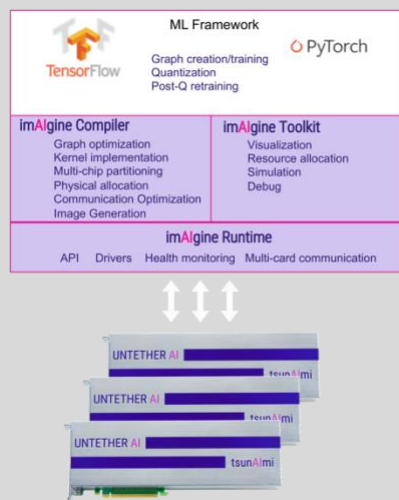
- Import TensorFlow, PyTorch, or ONNX graphs directly
- Automated quantizer and extracts performance without sacrificing accuracy
- Specify performance levels, silicon utilization, and power consumption targets

The imAagine Toolkit

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

The imAagine Runtime

- Provides C-based API for integration into your deep learning environment
- Monitor the health and temperature of the tsunAI mi® 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

Notice

THE INFORMATION DISCLOSED TO YOU HEREIN (THE "MATERIALS") IS PROVIDED SOLELY FOR THE SELECTION AND USE OF UNTETHER AI'S PRODUCTS. TO THE MAXIMUM EXTENT PERMITTED BY APPLICABLE LAW, MATERIALS ARE MADE AVAILABLE "AS IS". UNTETHER AI MAKES NO REPRESENTATIONS OR WARRANTIES, WHATSOEVER WITH RESPECT TO THE MATERIALS OR THE PRODUCTS, INCLUDING BUT NOT LIMITED TO REPRESENTATIONS OR WARRANTIES OF MERCHANTABILITY; SECURITY; RELIABILITY; ACCURACY; QUALITY; INTEGRATION; FITNESS FOR A PARTICULAR PURPOSE, NON-INFRINGEMENT, OR TITLE; THAT THE INFORMATION PROVIDED IN THIS MATERIAL IS SUITABLE FOR ANY PURPOSE; NOR THAT THE IMPLEMENTATION OF SUCH INFORMATION WILL NOT INFRINGE ANY THIRD PARTY PATENTS, COPYRIGHTS, TRADEMARKS, OR OTHER RIGHTS. WITHOUT LIMITING THE GENERALITY OF THE FOREGOING, UNTETHER AI EXPRESSLY DISCLAIMS ANY REPRESENTATION, CONDITION, OR WARRANTY THAT ANY INFORMATION PROVIDED TO YOU HEREUNDER, CAN OR SHOULD BE RELIED UPON BY YOU FOR ANY PURPOSE WHATSOEVER. UNTETHER AI DISCLAIMS ANY AND ALL LIABILITY RELATED TO THIS MATERIAL AND WILL NOT BE LIABLE FOR ANY LOSSES OR DAMAGE CAUSED BY RELIANCE ON THE INFORMATION IN THIS MATERIAL.

No license, either expressed or implied, is granted for any intellectual property rights of Untether AI or any third party through the information in this Material. Untether AI shall not be liable (whether in contract or tort, including negligence, or under any other theory of liability) for any loss or damage of any kind or nature related to, arising under, or in connection with, the Materials (including your use of the Materials), including for any direct, indirect, special, incidental, or consequential loss or damage (including loss of data, profits, goodwill, or any type of loss or damage suffered as a result of any action brought by a third party) even if such damage or loss was reasonably foreseeable or Untether AI had been advised of the possibility of the same. Untether AI assumes no obligation to correct any errors contained in the Materials or to notify you of updates to the Materials or to any products. You may not reproduce, modify, distribute, or publicly display the Materials without Untether AI's prior written consent. You should obtain the latest relevant Material before placing orders and should verify that such information is current and complete. All orders are subject to Untether AI's contract which outlines any applicable terms and conditions for a product.

Trademarks

Untether AI, tsunAI,mi, runAI, imAI, gine SDK are trademarks and/or registered trademarks of Untether AI Corporation in the U.S and other countries. Other company names may be trademarks of the respective companies.

Copyright

© 2023 Untether AI Corporation. All rights reserved.

UNTETHER AI