

Key Features

2 PetaOps of inference performance	800MB of on-chip SRAM	Low latency, batch = 1	80,000 fps ResNet-50 V1.5
12,000 qps Bert-base	32GB/s PCIe bandwidth	Multi-chip partitioning of large neural networks	Thermal monitoring

Overview

The tsunAI mi accelerator cards have the industry's highest compute density at over 2 PetaOps of INT8 performance in a single card. The power efficiency of the runAI200 devices enables 4 devices per card and stay within a 300W Thermal Design Point (TDP). The x16 PCI-Express Gen4 interface supports up to 32GBytes/s of bandwidth, enough for the most demanding AI applications.

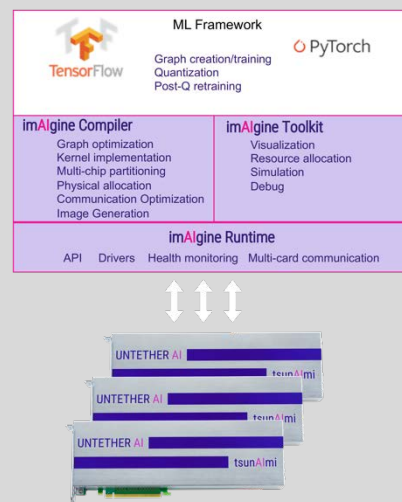
Applications

The runAI200 devices are designed to accelerate a multiplicity of AI workloads, such as vision-based convolutional networks, RNNs or attention networks for natural language processing, and time-series analysis for financial 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, Attention, BERT
Financial technology	X-Value adjustments, credit risk, portfolio balancing	TCNs, LSTMs

imAInge Software Development Kit

The imAInge software development kit (SDK) enables users to quickly create performant AI workloads using industry standard tools. Using TensorFlow or PyTorch, graphs are created, trained, and quantized. If additional accuracy is required, the imAInge SDK provides a post-quantization retraining module to ensure the utmost accuracy. Graphs are converted to program images in the imAInge compiler. Users can specify performance levels, silicon utilization, and power consumption targets to guide the compiler. A full toolset is available for implementation analysis for optimization. The imAInge Runtime provides an easily integrated c-based API for integration into your machine learning environment. The runtime also monitors 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	tsunAI mi accelerator card
Form factor	Double-wide, full height, full length PCIe
PCIe Interface	X16 PCIe Gen4
Clock Frequency	Variable, depending on throughput requirements
Memory	800MB on-chip SRAM

Thermal Specification

Parameter	tsunAI mi accelerator card
Total board power	TDP 300W
Cooling	Passive heatsink, bidirectional airflow
runAI200 maximum operating temperature	85°C Junction
Slowdown temperature (board sensor)	60°C
Shutdown temperature (board sensor)	65°C

Environmental

Parameter	tsunAI mi accelerator card
Operating temperature	0°C to 45°C
Storage temperature	-40°C to 75°C
Operating humidity	5% to 90% relative humidity
Storage humidity	5% to 95% relative humidity

Power Connector

8-pin CPU power connector, capable of supplying 225W

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

© 2020 Untether AI Corporation. All rights reserved.

UNTETHER AI