rebellions_

ION™ : AI Compute Core NPU

Compute Granule That Brings Maximum Flexibility and Energy Efficiency

Versatile Compute Engine Amongst a Plethora of NPUs

The Rebellions AI Compute Core ION™ provides the flexible inference capability with low power, a small footprint, and high performance for the edge computing systems. Featuring a customized instruction set architecture (ISA) crafted for >1K multiply-and-accumulate (MAC) units, the ION™ brings high-performance inference acceleration with exceptionally high utilization ratio compared to other AI accelerators (GPU, NPU, etc.). ION™'s versatility, compact size, and low power suffice the demands for edge deployments as in mobile.

Unmatched TOPS/Watt Verified in TSMC 7nm Technology

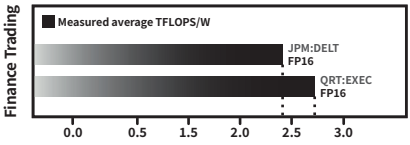
ION™ Compute Core was fabricated in TSMC 7nm technology and successfully measured its functionality and efficiency for the benchmark networks. Supporting mixed-precision (FP16, INT8/4/2) computation with up to 2GHz operating frequency, ION™ brings >2.0 TFLOPS/Watt for FP16-based vision tasks and >10 TOPS/Watt for INT8-based language tasks, respectively.

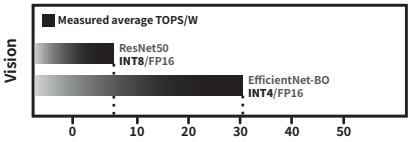
60x End-to-end System Performance on Finance and Edge Computing

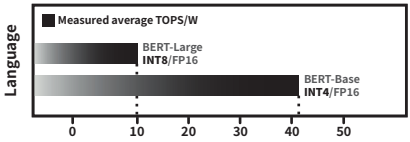
Powered by ION™ HW/SW full stack implementation, Multi-chip integrated FPGA-board is deployed to high frequency trading solution (HFT), LightTrader™, for the Wall Street based investment banks. The LightTrader™ achieves up to 60x more AI-enabled HFT performance, instantly upgrading any HFT solutions to intelligent trading system.

System Specifications

Technology	TSMC 7nm
Package size	8.7mm x 8.7mm
Compute Cores	1
Peak FP16 Perf.	4 TFLOPS
Peak INT8 Perf.	16 TOPS
Peak INT4 Perf.	32 TOPS
Max TDP	2 - 6 Watt (Configurable)
Highlights	Mixed precision Customized ISA Various vision and language models (CNN, LSTM, BERT, etc.)

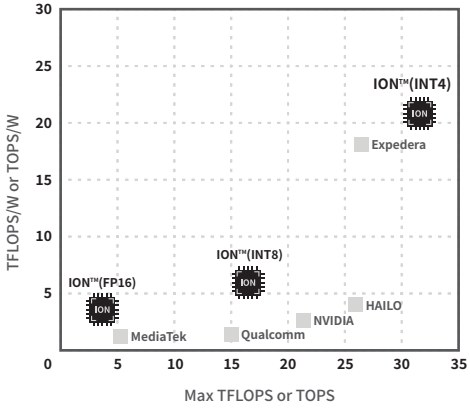




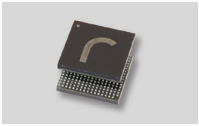


Competitive Energy Efficiency for Intelligent Edge

The ION™ Compute Core offers up to 10x more performance-per-watt figure even with the comparison against the state-of-the -art mobile NPUs. This opens the chances of wide ION™ deployment to different applications, including mobile companion chips, smart cities, robots and retail.



ION™ Brings Breakthrough Trading Solution Architecture Innovations



ION™ Compute Core

The first-generation Compute Core provides FP16-based accurate stock prediction, which utilizes the single batch inference and predicative execution pipelines. ION™ also supports BFloat16 and low precision integer operations such as INT8/4/2.

LightTrader™ : World-first AI-enabled HFT Card

This PCIe card integrates the custom AI accelerators, ION™, and the FPGA-based conventional HFT pipeline for the short-latency-high-throughput trading solutions with a minimized symbol miss rate, which is one-and-only solution for AI-based HFT trading so far.

Ultra-low Latency x High-throughput Finance AI Acceleration Server

Integrating eight LightTrader™ boards into one standard 4U server, the proposed server-level solution extends the capability of our ION™-based Finance Inference up to 0.5 Peta OPS and 3.2 Tbps symbol processing throughput.

Product	# of ION™s	TOPS	Average Power	DL Inference Throughput for DeepLOB*
ION™ Compute Core	1	16	1.6W	12.5K Symbols per Second
LightTrader™	4	64	20W	50K Symbols per Second
Single Server	32	512	300W	400K Symbols per Second

* DeepLOB: Deep Convolutional Neural Networks for Limit Order Books, IEEE Transactions on Signal Processing, 2018

For more information, please visit **rebellions.ai** or email **contact@rebellions.ai.**