ATOM™

: Versatile Inference System-on-Chip (SoC) for Server-scale Machine Learning (ML)
Bridging the latency-critical ML task and the hyper-scale ML services

World's Best Inference Performance at Edge and Datacenter Computing

Competing against the most recent AI accelerators such as AI100™ of Qualcomm and Greco™ of Intel Habana Labs, Rebellions' ATOM™ delivers uncompromised inference performance across different types of ML tasks, computer vision, natural language processing and recommendation model. ATOM™ utilizes the silicon-proven neural core, ION™, as a compute granule that scales up with the perfect linearity for large-scale inference operations in edge computers and datacenters.

Equating Latency Optimization with Throughput Maximization

Rebellions' ATOM™ combines ML specialized dataflow architecture and the many-core SoC architecture to bring the best inference performance with SR-IOV based user-level parallelization. In order to minimize the latency overhead for data communication and synchronization in inter-/intra- chip orchestrations, ATOM™ adopts our proprietary multi-level dependency control mechanisms and interrupt protocol that enhance overall system utilization by up to 40% of total end-to-end inference latency. ATOM™ processor can be the key enabler of enterprise/personal server-level AI services, completing the solutions of the customers' AI-as-a-Service (AIaaS).

Chip Design in Samsung 5nm EUV Technology

Samsung 5nm EUV Technology empowers the performance-driven chip architecting and energy-saving chip design methodologies simultaneously, preparing the world-class datacenter inference chip competition. Gearing with PCIe Gen5 and GDDR6 high-speed IO technologies, the ATOM™ processor can approach different technology markets, spanning from edge computer to datacenter inference. Its engineering sample will be available in December 2022 and the mass production is scheduled to the midst of 2023.

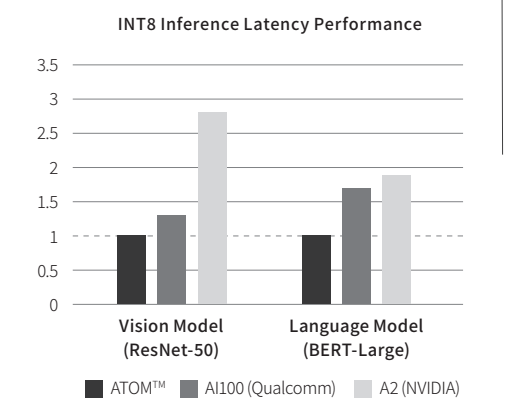
ATOM™ Specifications

	SINGLE CHIP
Process Node	5nm
FP16	32 TFLOPS
INT8 / INT4	128 TOPS / 256 TOPS
On-chip SRAM	64 MB
Max Thermal Design Power (TDP)	60 - 150 Watt (Configurable)
Memory	16 GB, GDDR6 (ECC enabled), 256 GB/s
Host/Chip-to-Chip Interface	PCIe Gen5 , 64 GB/s
Multi-Instance NPU	· HW isolation up to 16 independent tasks

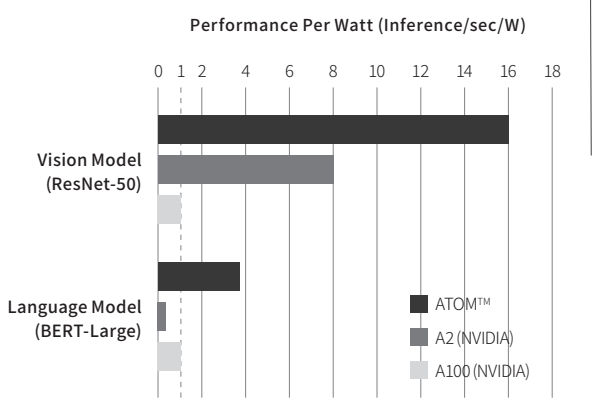
Compute Platform Specifications

Card Types and Power (TDP)	
HHHL	30-60 Watt (Configurable)
FHFL	150 Watt
Dual-chip Combo	300 Watt
Card Types and Performance (Max TOPS and External Bandwidth)	
HHHL	96 TOPS | 256 GB/s
FHFL	128 TOPS | 256 GB/s
Dual-chip Combo	256 TOPS | 512 GB/s
On Card DRAM	Up to 32GB (Combo PCIe card) Up to 16GB (HHHL, FHFL)
Data Types	FP16, INT8, INT4, INT2

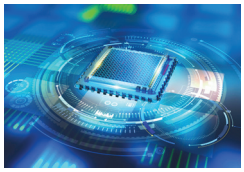
Performance Comparison



Energy Efficiency Comparison



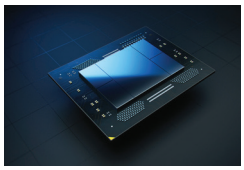
Key Features



✓ ION™ Core Architecture

Silicon-proven Rebellions' AI core can deliver outstanding effective performance via highly utilized dataflow and its efficient hardware implementation. The core can readily handle different networks with different depths and complexities from tiny app to datacenter-level services. The maximized energy efficiency can be obtained with low-power target implementation methodology and tidy core structures.

✓ System-on-Chip Architecture



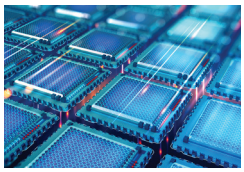
With ION Core tightly orchestrated with the command processor of SoC, ATOM™ can accelerate large-size networks with outstanding efficiency as well as the small networks. The overhead of extra communication can be neatly hidden by the on-chip level control and data dependency handling.



✓ High Speed IOs

GDDR6 and PCIe Gen5 interfaces make the large-size data transaction swift and the host and memory interface powerful. These IPs dramatically widens the applicability of ATOM™ covering hyper-scale model inferences that require large on-chip/off-chip memory bandwidth up to a few hundred GB/s.

✓ Multi-Instance NPU (MIN)



An ATOM™ can be partitioned to support at least 16 jobs simultaneously, isolated at different HW/SW levels. The MIN enhances the competitive edges of datacenter and cloud computing provider in terms of the user capacity and the effective usage of the computing farms.

Supporting Networks

ATOM™ can accelerate different types of neural networks efficiently, including convolutional neural networks (CNNs), long short-term memory (LSTM), bidirectional encoder representations from transformers (BERT) and the recent transformer networks (e.g. T5, GPTs, etc.) as well.

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

© 2022 Rebellions Inc. All rights reserved. Rebellions, the Rebellions logo, ION, and LightTrader are trademarks and/or registered trademarks of Rebellions Incorporated in the U.S. and other countries. All other trademarks and copyrights are the property of their respective owners.