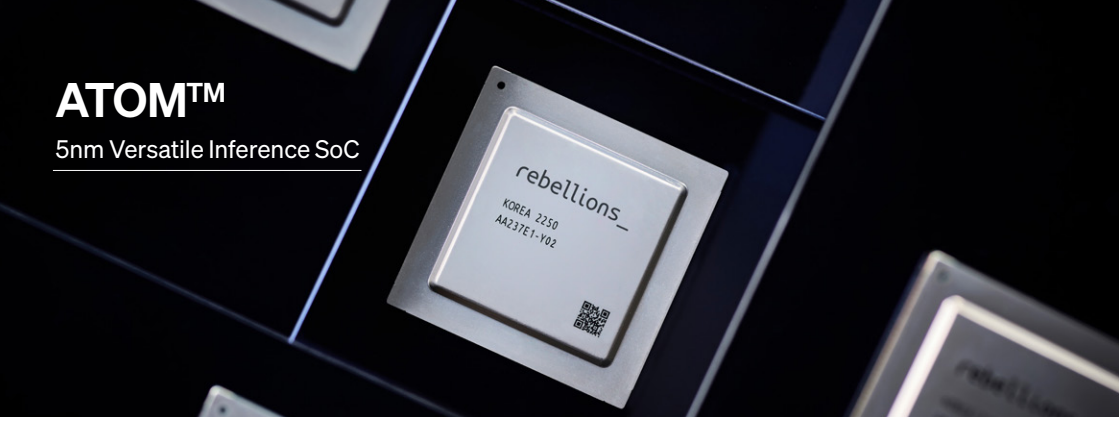


ATOM™

5nm Versatile Inference SoC

World's Best Inference Performance for Edge and Cloud Computing

Competing against state-of-the-art AI accelerators, ATOM™ delivers uncompromised inference performance across different types of ML tasks, computer vision, natural language processing and recommendation models. ATOM™ utilizes the silicon-proven neural core, ION™, as a compute granule that scales up with perfect linearity for large-scale inference operations required in edge computers and datacenters.

Equating Latency Optimization with Throughput Maximization

ATOM™ combines ML specialized dataflow architecture and 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 is a key enabler for enterprise/personal server-level AI services, perfecting the customers' AI-as-a-Service (AIaaS) stack.

Advanced Manufacturing

ATOM™ is built with Samsung Electronics' most advanced EUV process node, delivering the best performance and energy efficiency. Geared with PCIe Gen5 and GDDR6 high-speed IO technologies, the ATOM™ processor can serve different markets, spanning from edge computer to datacenters.

ATOM™ Specifications

Single Chip	
FP16	32 TFLOPS
INT8	128 TOPS
On-chip SRAM	64 MB
Max Thermal Design Power (TDP)	30 - 130W (Configurable)
Memory	16 GB, GDDR6 (ECC enabled), 256 GB/s
Host / Chip-to-Chip Interface	PCIe Gen5 16 lane 64GB/s
Multi-Instance NPU	HW isolation up to 16 independent tasks

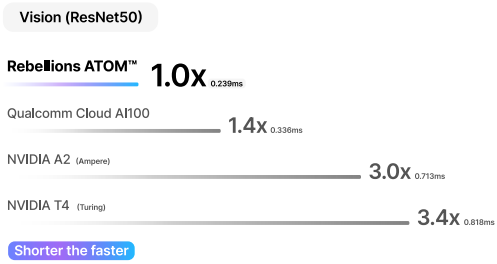
Compute Platform Specifications

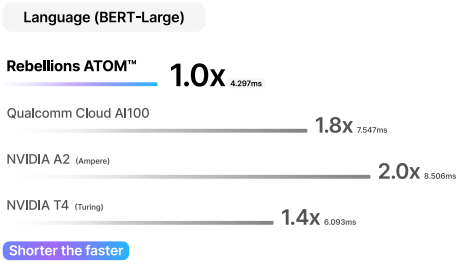
Card Types and Power	
HHHL(Single Slot)	30-60W(Configurable)
FHFL (Single Slot)	130W
Dual Slot	250W
Card Types and Performance (Max TOPS and External Bandwidth)	
Single Slot	128 TOPS I 256 GB/s
Dual Slot	256 TOPS I 512 GB/s
On-card DRAM	32GB (Dual Slot) 16GB (HHHL, FHFL)

MLperf™

Inference v3.0 (Latest), Single Stream Latency

ATOM™ readily supports the most advanced AI networks, including MLperf benchmark models and state-of-the-art mid-size Large Language Models (LLMs). Built with the most advanced silicon technology, ATOM™ delivers the best-in-class performance and energy efficiency.





Key Features

✓ ION™ Core Architecture

Rebellions' silicon-proven AI core delivers outstanding performance owing to its highly utilized dataflow and efficient hardware implementation. The AI core can readily handle networks with different depths and complexities from tiny applications to datacenter-level services. Energy efficiency is maximized by implementing a low-power design methodology and using tidy core structures.

✓ System-on-Chip Architecture

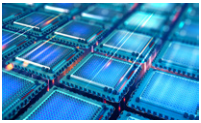
With ION™ Core tightly orchestrated with an internal command processor, ATOM™ accelerates both large and small networks with outstanding efficiency. The overhead of extra communication is neatly hidden by the on-chip level control and data dependency handling.

✓ 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 widen the applicability of ATOM™, enabling hyper-scale model inferences that require large on-chip/off-chip memory bandwidth up to a few hundred GB/s.

✓ Multi-Instance NPU (MIN)

ATOM™ can be partitioned to support up to 16 jobs simultaneously, isolated at different HW/SW levels. This feature provides a clear competitive advantage to service providers by boosting the user capacity and the effective usage of 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 recent transformer networks (e.g. T5, GPTs, etc.).

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