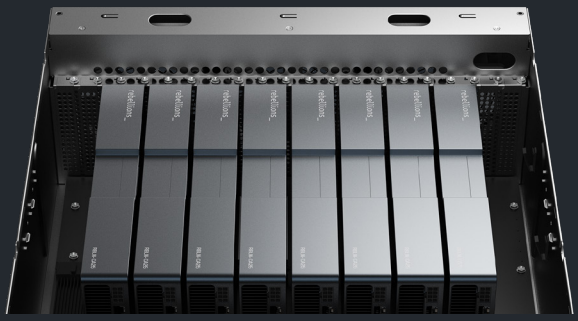
rebellions_



ATOM™-Max Server

High Performing Server for Large-Scale AI Inference

ATOM™-Max Server is a power-efficient, single-server solution built for large-scale AI inference. It supports up to 8 ATOM™-Max PCIe cards, enabling deployment of hundreds of AI models spanning Vision AI, LLMs, Multimodal AI, and even Physical AI workloads. Fully compatible with leading inference frameworks like vLLM, Triton, and Kubernetes, you can seamlessly transition from GPU workflows with familiar tools and guided tutorials.

Key Features

Performance at Any Scale

Even under heavy demand, the ATOM™-Max delivers stable, high-throughput performance—generating thousands of tokens and processing image frames per second, all from a single system.

Sustainable AI Infrastructure

ATOM™-Max, with its exceptional power efficiency, significantly lowers total cost of ownership (TCO) and enables a more sustainable AI infrastructure.

Variety of Models Applications

From LLMs and Vision AI to Multimodal AI and Physical AI, build tailored services with hundreds of AI models.

Develop As You Always Have

Start right away with familiar workflows (PyTorch, TensorFlow, etc.) and step-by-step tutorials.

Full-Stack Software Support

Compatible with popular open-source ecosystems, supporting efficient serving, flexible resource management, and monitoring through tools like vLLM, Triton Inference Server, Kubernetes, and Prometheus— you can build full end-to-end services with ease.

Spec

NPU	ATOM™-Max NPU Card *8
NPU Memory	512GB GDDR6, 8TB/s
Performance	1,024 TFLOPS (FP16) 4,096 TOPS (INT8)
Form Factor	4U
CPU	5th Gen. AMD EPYC Processor *2
Memory	1.5~2.3TB
Storage	1.92TB SSD * 2
Network	10G 2Port * 2 400G 1port (Optional)
Max Power Consumption	Typical 3.4kW (Max ~4.3kW)
PCIe Slots	13x PCIe gen5 x16 [FHFL slots] [8x ATOM™-Max + 1x 400G 1-port NIC + 1x 10G 2-port NIC]
Compatible Software	- **OS**: Ubuntu, RHEL, AlmaLinux, RockyLinux - **Frameworks & Tools**: Hugging Face, PyTorch, TensorFlow, Triton - **Inference Serving**: vLLM, Triton Inference Server, TorchServe, Ray Serve - **Orchestration**: Docker, OpenStack, Kubernetes

RBLN SDK

We deliver a full-stack inference platform that combines the familiar usability of GPUs with architecture built for next-generation AI workloads. From PyTorch development to LLM serving and deployment, every stage is designed for enterprise environments.

Driver SDK

Core system software and tools for running NPUs

- Firmware
- Kernel Driver
- User Mode Driver
- System Management Tool

NPU SDK

Development toolkit for models and services

- Compiler, Runtime, Profiler
- Hugging Face Integration
- Major Inference Servers Supported (vLLM, TorchServe, Triton Inference Server etc.)

Model Zoo

300+ ready-to-run PyTorch and TensorFlow models on Rebellions NPUs

- Natural Language Processing
- Generative AI
- Speech Processing
- Computer Vision
- Physical AI