rebellions_

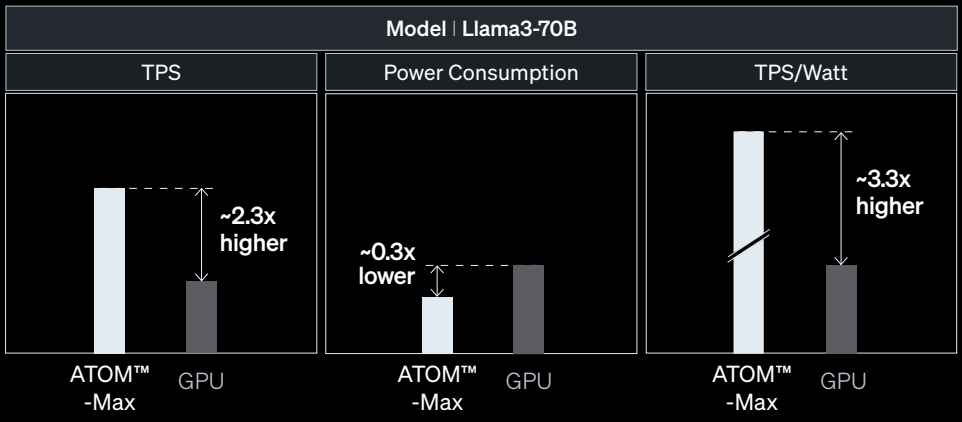
ATOM™-Max

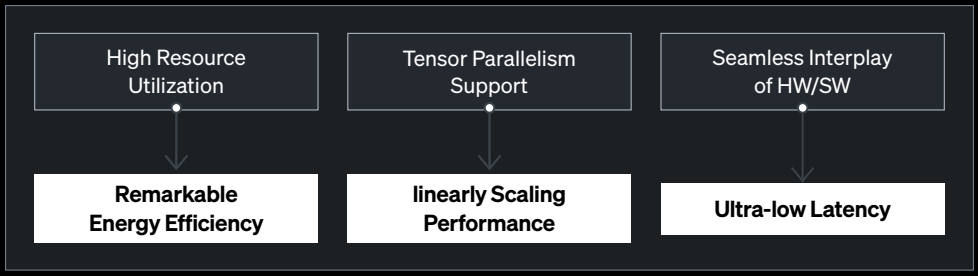
Energy Efficient AI Accelerator
with Flexible Interconnect Topology

Energy Efficiency

ATOM™-Max surpasses its class competitor GPU in TPS per watt. With exceptional hardware utilization, powered by optimized data and memory management on both hardware and software levels, resources are used as efficiently as possible. This efficiency leads to substantial savings in TCO, which multiply as the deployment scales.

ATOM™-Max	
FP16	128 TFLOPS
INT8	512 TOPS
External Memory Capacity (GDDR6)	64 GB
Memory Bandwidth	1 TB/s
Host Interface	PCIe Gen5 x16 (64 GB/s)
Max Power Consumption	350 watts





Scalability

From a single chip to full rack deployments, ATOM™-Max delivers high TPS with linear scalability, all while maintaining excellent performance per watt. Direct data exchange between cards over PCIe Gen5 enhances both efficiency and scalability, allowing ATOM™-Max to handle larger configurations with ease. Additionally, the high bandwidth and capacity of GDDR6 enable fast, efficient data processing, ensuring consistent performance as the system scales.

Hardware-Software Co-Optimization

Leveraging a co-optimized hardware-software stack, ATOM™-Max maximizes memory efficiency and utilization through an advanced synchronization and shared memory (SHM) scheme. Its compiler autonomously transforms any AI model into highly optimized execution instructions, ensuring peak performance on ATOM™-Max's specialized architecture while minimizing latency and computational overhead.

Large-scale Serving Readiness

Rebellions' system is optimized for large-scale LLM serving, with support for vLLM transformers and PyTorch 2.x. The RBLN Software Stack enables seamless integration and scaling, featuring compiler-level optimizations such as tensor parallelism to efficiently handle demanding transformer models.

Discover More
rebellions.ai

rebellions_