rebellions_

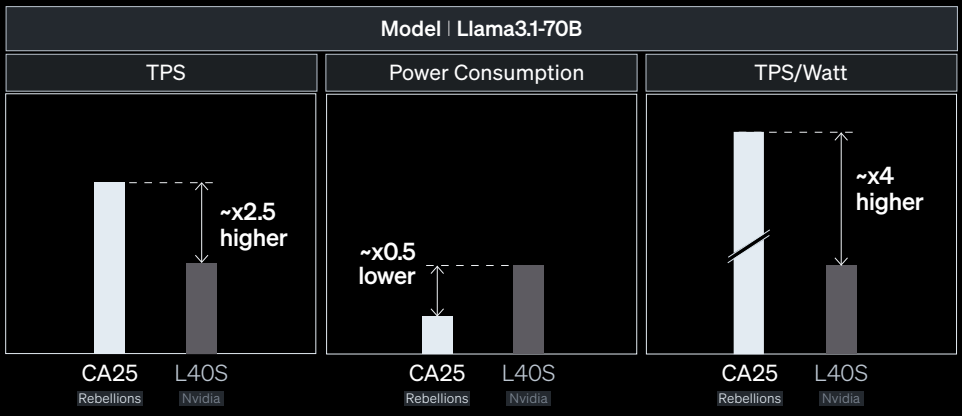
ATOM™-Max (RBLN-CA25)

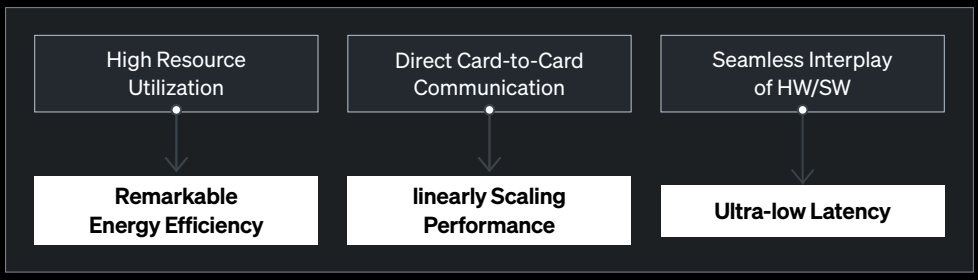
Energy Efficient AI Accelerator with Flexible Interconnect Topology

Energy Efficiency

ATOM™-Max (RBLN-CA25) surpasses its class competitor L40S 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.

RBLN-CA25	
FP16	128 TFLOPS
INT8	512 TOPS
External Memory Capacity (GDDR6)	64 GB
Memory Bandwidth	1,024 GB/s
Host Interface	PCIe Gen5 x16 (64 GB/s)
Intercard Interface (MCIO)	
TDP	350 watts





Flexibility

To reduce communication overhead by bypassing the CPU, each card is equipped with two dedicated connectors. These connectors allow a flexible configuration of up to eight **RBLN-CA25** cards, adjusting seamlessly to compute- or memory-intensive operations. This adaptability enables a range of topologies, including torus, ring, and tree, for optimized performance across applications.

Scalability

From a single chip to full rack deployments, **RBLN-CA25** 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 **RBLN-CA25** 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.

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_