rebellions_

ATOM™-Max:

Boosted Performance for Large-Scale Inference

May 02, 2025

As AI workloads grow in complexity and scale, traditional GPU-based systems face mounting challenges in both efficiency and sustainability. Their high power draw and infrastructure demands create bottlenecks for long-term deployment at data center scale. ATOM™-Max addresses this gap with a purpose-built architecture optimized for large-scale inference, offering a more efficient and scalable alternative to general-purpose accelerators. ATOM™-Max ensures high utilization and throughput for the most demanding enterprise and data center AI workloads. Its architectural efficiency not only reduces operational costs but aligns with the increasing need for carbon-aware, performance-per-watt optimized AI infrastructure.

Scalable, High-Efficiency Inference

ATOM™-Max delivers a substantial performance uplift purpose-built for inference at scale, combining exceptional compute throughput with high-bandwidth memory access. With 128 TFLOPS (FP16) and up to 1024 TOPS (INT4), it is engineered to handle the intensive demands of LLMs and AI-driven enterprise workloads. The architecture integrates direct card-to-card communication over PCIe Gen5 x16, enabling fast, low-latency data exchange and efficient horizontal scaling across accelerator nodes. Designed specifically for large-scale inference environments, ATOM™-Max ensures high utilization, predictable latency, and seamless deployment in modern data center infrastructure.

Total Cost of Ownership (TCO) Challenges

Supporting large-scale inference requires balancing throughput, performance-per-watt, and deployment flexibility. Traditional GPU systems incur high CapEx and OpEx, making them difficult to scale sustainably. ATOM™-Max (RBLN-CA25) outperforms its class competitor, L40S, in tokens-per-second per watt (TPS/W), offering superior performance efficiency in real-world scenarios.

With high hardware utilization driven by tightly integrated data and memory management, ATOM™-Max minimizes idle overhead and resource waste. This architectural efficiency translates to substantial TCO savings that scale with deployment.

ATOM™-Max: Scaling Performance-per-Watt

Traditional infrastructures face limitations in power efficiency, memory bottlenecks, and deployment costs. ATOM™-Max addresses these challenges.

Scale More, Save More with High Compute Density

With 128 TFLOPS FP16 and 1024 TOPS INT4 within a 350W envelope, ATOM™-Max delivers industry-leading efficiency, enabling higher throughput per rack. By significantly reducing server requirements for the same AI workload, it lowers infrastructure costs and optimizes system performance. As workloads increase, economies of scale further amplify cost savings, making ATOM™-Max the ideal choice for high-throughput AI infrastructure at scale.

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.

[High-Density AI Server with ATOM™-Max Accelerators]

Built for Tomorrow's AI

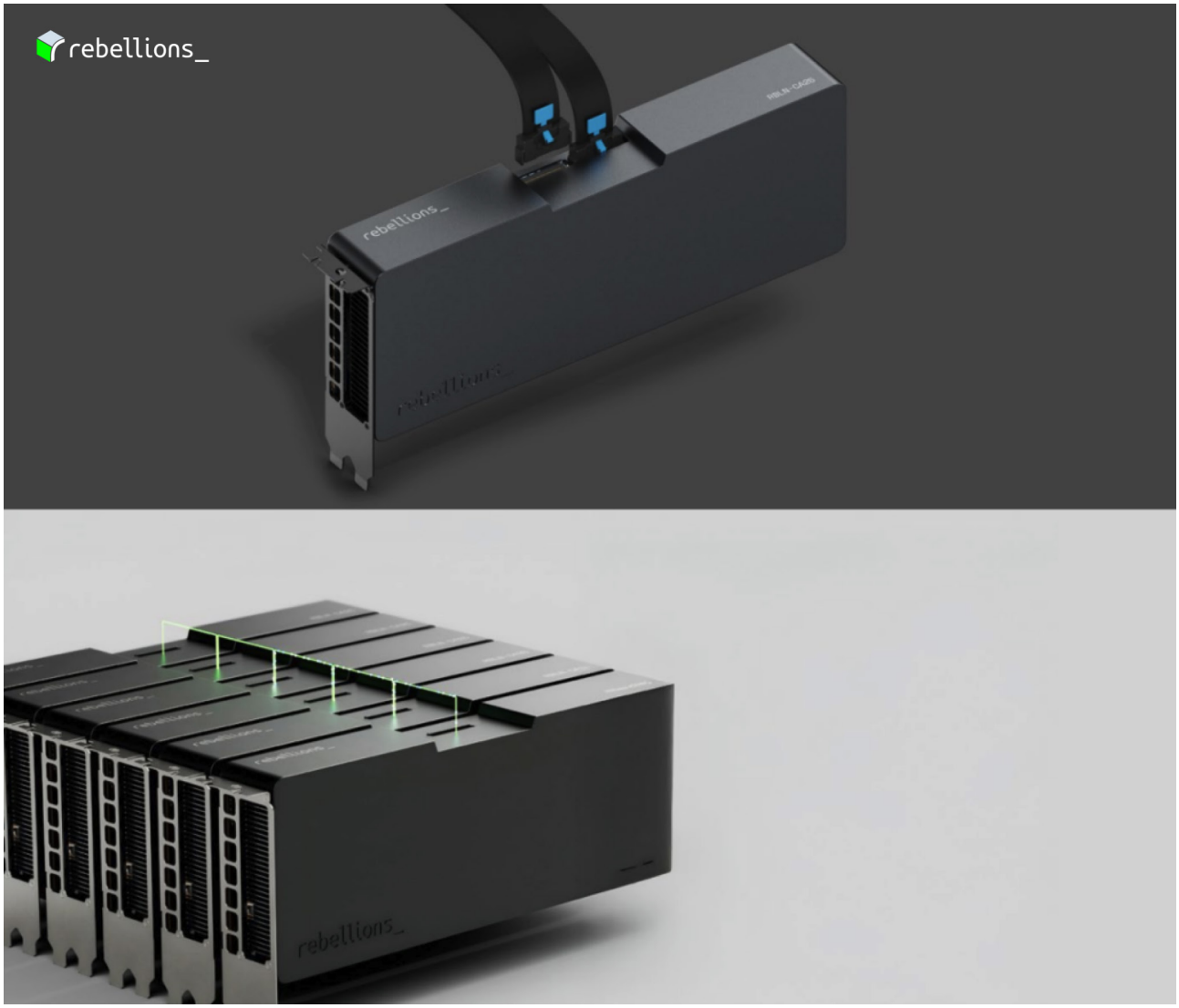
High Bandwidth

With 64 GB of high-bandwidth memory and 1024 GB/s bandwidth, ATOM™-Max delivers the speed and capacity needed to power demanding AI and LLM workloads. Its optimized

architecture ensures efficient data flow, maximizing throughput and eliminating bottlenecks in large-scale deployments.

Seamless Scaling

Beyond its standalone performance, ATOM™-Max is designed for seamless multi-card scalability. Dedicated intercard cables enable direct high-speed connections between up to eight ATOM™-Max cards, significantly reducing communication overhead and unlocking even greater AI acceleration at scale.



[High-Speed Connector for Efficient Multi-Card Scaling]

Conclusion

As AI inference demand outgrows training, traditional GPU-based architectures are becoming increasingly inefficient, driving up energy costs, infrastructure overhead, and

operational constraints. ATOM™-Max offers a purpose-built alternative, delivering:

- **Higher compute density for more efficient large-scale inference.**
- **Lower power consumption, reducing both TCO and environmental impact.**
- **Seamless scalability, ensuring enterprises can deploy future AI models effortlessly.**

Power-efficient, high-density AI accelerators like ATOM™-Max will be key to sustainable, cost-effective, and scalable AI deployment.