

Rebellions Scalable Design

Rebellions Scalable Design

Rebellions' engineering philosophy is rooted in a scalable and modular architecture, which is exemplified by the Rebellions Scalable Design (RSD). RSD serves as the cornerstone for all of our current and future products, providing a robust foundation that enables seamless scaling across various deployment scenarios.

RSD delivers linear scalability and ensures that your system performance continues to grow without compromise. This enables Rebellions to offer a high-performance solution that is scalable and energy efficient, making it the optimal choice for inference tasks of all sizes, including those required by hyperscalers.

RSD also features comprehensive support for large language models, backed by a dedicated software stack that maximizes performance and compatibility. Whether handling convolutional neural networks or the more demanding transformer-based workloads, RSD is engineered to deliver consistent and powerful performance.



Core Technologies

Scaling model size significantly expands the potential for diverse AI applications, but behind the scenes, maintaining seamless communication between AI processors to manage such large workloads demands deep technical expertise. RSD leverages robust **tensor parallelism** to efficiently distribute computation, ensuring that even the largest models run smoothly across multiple processors. The RBLN Compiler also plays an important role in **optimizing models** of models to the finest detail. The integration of **PCIe Gen5** further supports high-speed input/output as well as direct card-to-card communication, drastically reducing latency and maximizing data throughput across the entire system. This combination of advanced parallelism, optimizations and cutting-edge interconnect technology ensures that RSD can handle the most demanding AI tasks with precision and speed.

Tensor Parallelism

Running inference for LLMs on AI processors presents numerous challenges, with the prefill phase, which is highly computation-intensive, and the decoding phase, which places significant demands on memory resources. Here tensor parallelism offers a solution by distributing the computational workload across multiple devices, effectively reducing the per-device memory footprint as well as computation load. Implementing tensor parallelism correctly can mitigate the memory bandwidth constraints associated with KV caching and the large weights of LLMs. It also ensures that the hardware can efficiently manage the complexities of LLM inference while maintaining high throughput and low latency.

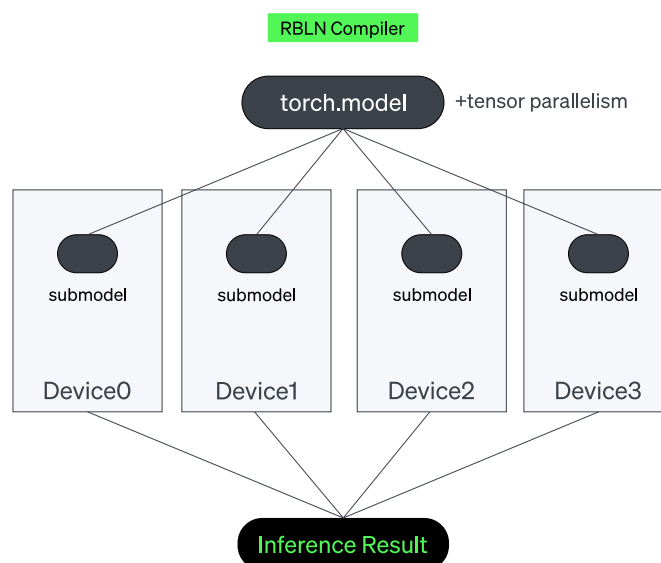
The RBLN Compiler is optimized to manage tensor parallelism in a way that maximizes performance and utilization rate. At compile time, the Compiler meticulously divides the model into tensors across several devices, allowing each chip to handle only a portion of the computation. The resulting Command Stream, a set of instructions for the Command Processor, also includes the inter-device data movement information required between chips during inference.

Compiler-level Optimizations

RBLN Compiler is designed to handle the complexities of scaling AI workloads. By efficiently supporting tensor parallelism, it enables seamless distribution of models across multiple devices, ensuring optimal resource utilization and fast execution. The Compiler's advanced optimizations for multi-device communication, automatic splitting, and layer pipelining further enhance scalability, making it an essential tool for high-performance AI inference in large-scale deployments.

1. Automatic Multi-Device Splitting

RBLN Compiler automatically handles both the splitting and reconnecting of operations, simplifying tensor parallelism from the developer's perspective.

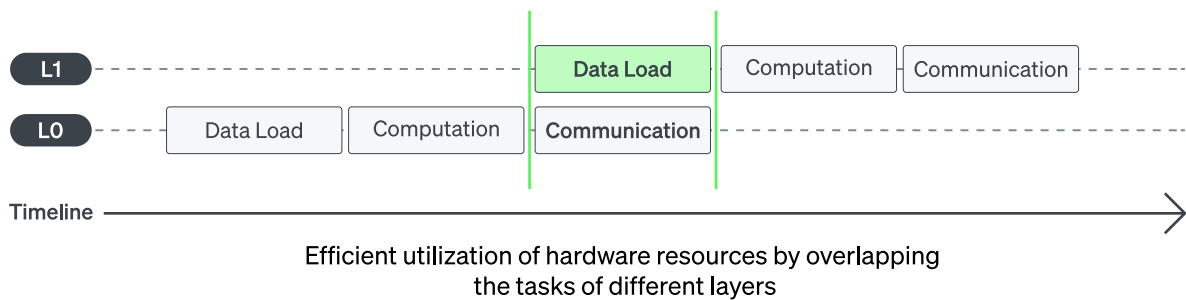


2. Optimization of Inter-Device Communication

RBLN Compiler optimizes inter-device communication during the execution of LLMs, with efficient processing of collective communication patterns such as broadcast, reduce and partial sums. It minimizes the communication overhead while reducing the memory footprint, enabling more efficient processing of distributed workloads.

3. Efficient Layer Pipelining for Intra-Device Communication

With layer pipelining within each device, RBLN Compiler enables seamless inter-device communication. This technique ensures that all operations are processed in parallel in a way to reduce idle time, maximizing hardware utilization while minimizing communication overhead.



PCIe Gen5

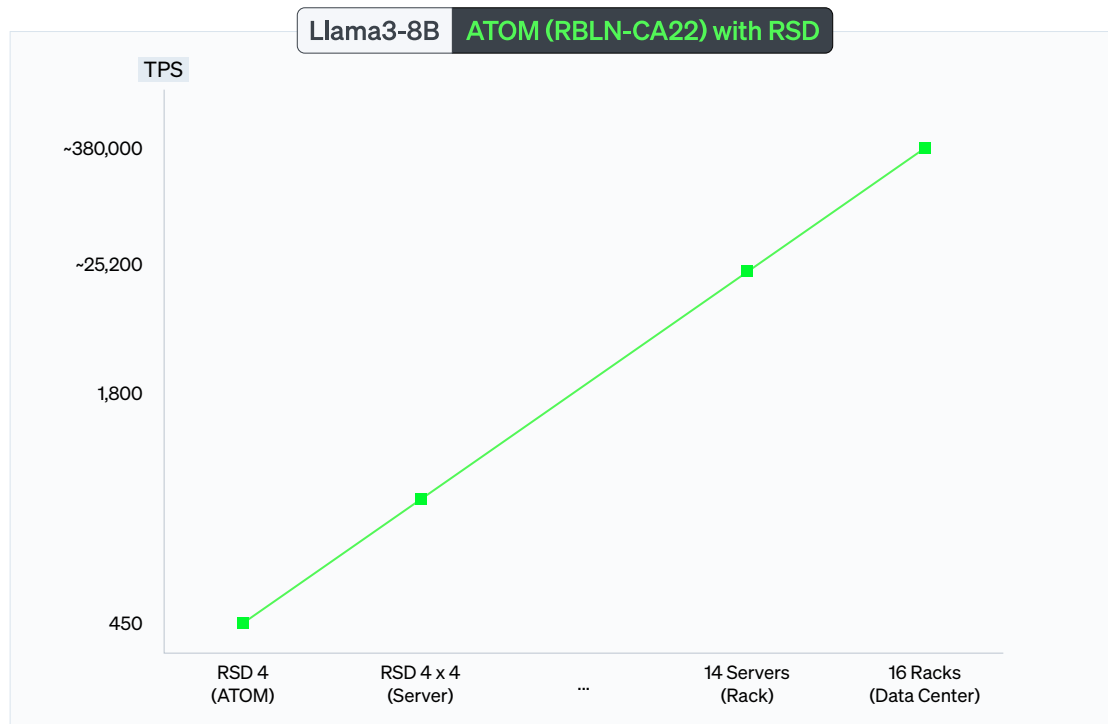
Rebellions' RSD employs the advanced PCIe Gen5 x16 interface, offering full-duplex bidirectional 64GB/s bandwidth for both host connectivity and direct inter-card communication. Efficient card-to-card communication is especially important for tensor parallelism, ensuring high throughput and low latencies in communication across all Neural Engines and delivering superior inference performance with scalable output speeds. To fully harness PCIe's capabilities, Rebellions' firmware is specially optimized, while the PCIe switch and CPU have been adapted for maximum interoperability, further enhancing system-wide efficiency.

System Solution

From lightweight use cases like workstations to demanding AI workloads such as SLMs and LLMs, **RSD** stands out as a powerful, cost-effective solution delivering exceptional **TPS/Watt**. This efficiency is the result of our system's true linear scalability, enhanced by rack-level performance optimizations that seamlessly integrate **vLLM** with **LiteLLM**. By combining these technologies, RSD not only scales effortlessly with increasing computational demands but also maximizes throughput per unit of energy and cost, making it an ideal choice for AI deployments of any scale.

Linear Scalability

Linear scalability is a critical factor in AI workloads because it allows performance to increase proportionally with the addition of more computational resources, ensuring that investments in hardware directly translate to gains in throughput. Such capability is vital for managing the escalating requirements of extensive AI models and data-heavy applications without compromising on efficiency or speed.



* Projections based on internal testings.

RSD, with its wide range of deployment options, including cards, servers, and rack systems, successfully delivers linear scalability. Achieving throughput that scales directly with the increasing number of devices while maintaining consistently low latency is a technically complex challenge. It requires precise optimization of data synchronization across multiple nodes, minimizing inter-node communication overhead, and efficiently handling memory bandwidth to prevent bottlenecks. Additionally, maintaining low-latency performance demands advanced techniques in load balancing and dynamic workload distribution to ensure that all devices operate at peak efficiency without delays. By tackling these complexities head-on through cutting-edge hardware architecture and intelligent software frameworks, RSD not only meets these challenges but excels at enabling scalable, high-speed processing as computational demands grow.

Rack-level Optimizations

For rack-level AI inference solutions, seamless communication between multiple servers is dependent on effective routing protocols for workload distribution.

RSD integrates a router server with **vLLM** to deliver optimal performance and scalability at the rack-level. The router server serves as a comprehensive framework that bundles multiple **vLLM instances**, allowing them to operate as a unified system. It also enhances the entire rack's capability by intelligently distributing workloads among multiple servers. By bundling several **vLLM instances** into a cohesive system, this approach ensures that computational tasks are optimally balanced across all available resources, preventing server overload and maximizing throughput. This dynamic load balancing not only boosts the overall efficiency of our rack solution but also maintains consistent low-latency performance, even under heavy AI workloads.

From the user's perspective, the LLMs become accessible through API endpoints, prioritizing usability while delivering robust, scalable AI inference. By implementing a router server alongside vLLM, we create a powerful synergy that transforms individual servers into a highly coordinated, high-performance AI system.

Llama3-8B

Llama3-8B results emphasize the power efficiency of RSD compared to competitors on the market. With a single ATOM server delivering 1800 TPS at 2,500 watts, the throughput scales to 25,200 at the rack-level with 224 ATOM cards. The gains become even more impressive at a larger scale, with RSD achieving over 8x gains in both energy and cost, as measured by TPS/Watt and TPS/\$, respectively.

Server			Rack (assumes rack power = 35 kW)		
Target model: Llama3-8B, FP16, Input/output = 2K/2K	rebellions_ ATOM™	Peer A	rebellions_ ATOM™	Peer A	
			At rack power of 35 kW		
			224 cards	72 cards**	
TPS (max TPS at best batch)	1.8 k	constraints a single node cannot run Llama3-8B	25.2 k	3 k**	
TDP (watt)	2.5 k		Rack power fixed to 35 kW		
TPS/Watt	0.72		energy 8X ↑ vs. Peer A	0.72	0.09**
TPS/\$	0.012		cost 8X ↑ vs. Peer A	0.008	0.001**


*Highly likely to be subjected to semiconductor export regulations; strongly advised to check regulatory risks before proceeding;

**Estimated based on the public information researched

Conclusion

RSD embodies our commitment to delivering cutting-edge AI infrastructure that can seamlessly scale to meet the demands of increasingly complex and resource-intensive AI workloads. Built on a modular and scalable architecture, RSD ensures linear scalability across a wide range of deployment scenarios, from single workstations to full rack systems. This capability, combined with advanced tensor parallelism, PCIe Gen5 integration, and the powerful optimizations enabled by RBLN Compiler, guarantees efficient, high-performance AI inference at any scale.

Contact Us

 rebellions.ai

 contact@rebellions.ai