rebellions_



Rebel100™

Peta-Scale MoE Inference. Without the Energy Tax.

- **Built to serve frontier LLMs with high utilization and low power.**
- **Powered by unified mixed-precision cores, responsive DMA scheduling, and UCIe interconnect.**
- **Rack-scale performance. Modular flexibility. Ready for deployment.**

Rebel100 is the world's first UCIe-Advanced AI accelerator card–engineered for efficient, scalable inference and seamless deployment at hyperscale.
At its core is a modular SoC that fuses four identical compute chiplets into a monolithic-acting unit via high-bandwidth UCIe-Advanced links, enabling seamless scaling from single-node to multi-rack clusters.

Accelerator Card Specifications

Architecture	4-homogeneous-chiplet SoC based on UCIe-Advanced
Compute (Dense)	1,024 TFLOPS (FP16) 2,048 TFLOPS (FP8)
External Memory	HBM3E 144GB 4.8TB/s
Chiplet (UCIe-A) Interconnection	16Gbps 1TB/s per channel

Host Connection
2x PCIe Gen5 x16
Power Consumption
Up to 600W
Software
Native-support of PyTorch 2.x, vLLM vLLM PyTorch 2.x

All information is provided "as is" without any warranties or guarantees by Rebellions, Inc.("Rebellions"). Rebellions assumes no liability for use of or reliance upon this information.

SoC Architectural Pillars

Unified Compute Engine: One Core for All Precisions

A mixed-precision core with per-operand flexibility supports FP8 and FP16 in a unified pipeline—eliminating instruction switching and boosting compute density by x2.8 on ATOM™.

Responsive DMA Scheduling: Memory Access Without Waiting

A pre-compiled DMA engine orchestrates KV cache access with command-level bandwidth and QoS control, delivering x3.3 higher per-core bandwidth—keeping 32K-token decoding smooth.

Modular UCIe Interconnect: Multiple Dies as One

Four chiplets are unified via 1TB/s per-channel UCIe-Advanced links into a single system. The same interconnect extends off-chip, enabling rack-scale disaggregation.

Holistic Synchronization: Perfect Orchestration for Peta-Scale Workloads

A hardware-accelerated sync fabric coordinates execution across chiplets, including sparse and MoE models. It enables fine-grained parallelism and eliminates stalls during expert routing.

Rebel100 vs. H200 (Llama 3.3 - 70B)

Throughput (TPS) ~x1.6

Efficiency (TPS/Watt) ~x3.2

Power Consumption (Watt) ~x0.5



Benchmark Condition: Performance measured on Llama 3.3 70B (TP2, FP8) with runtime input/output length 2048/2048.

Rebel100 redefines the economics of AI infrastructure – from silicon to system scale.

In internal evaluations, Rebel100 has demonstrated higher throughput than NVIDIA H200-class GPUs on large-scale LLMs–resolving both compute and memory bottlenecks in a single, modular system. Looking ahead, future REBEL chiplets including I/O dies, are already in development, designed to extend this platform for trillion-parameter models and multi-node exascale deployments.

Disclaimer: All performance data is based on Rebellions' internal estimates and tests under optimized settings. GPU figures are based on public data or third-party benchmarks. Actual results may vary depending on system conditions and software versions. This information is for reference only and subject to change without notice.