

Peta-Scale SoC for Massive AI Serving: REBEL-Quad



Contents

Peta-Scale SoC for Massive AI Serving: REBEL-Quad

Solving Challenges of LLM

- └ Unified Mixed Precision Compute Engine
- └ Predictive DMA & On-Chip Mesh
- └ Holistic Synchronization
- └ Custom Die-to-Die Protocol

Unified Mixed Precision Compute with Chiplet- Scale Scalability

Predictive DMA and High-Bandwidth Memory Access

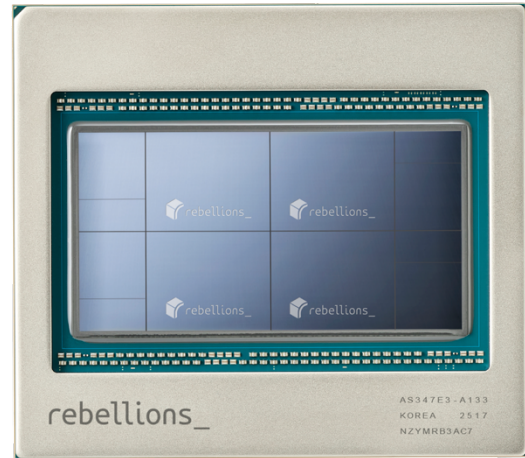
Hierarchical Synchronization and Peer Communication

Scalable Die-to-Die Protocol for Modular Expansion

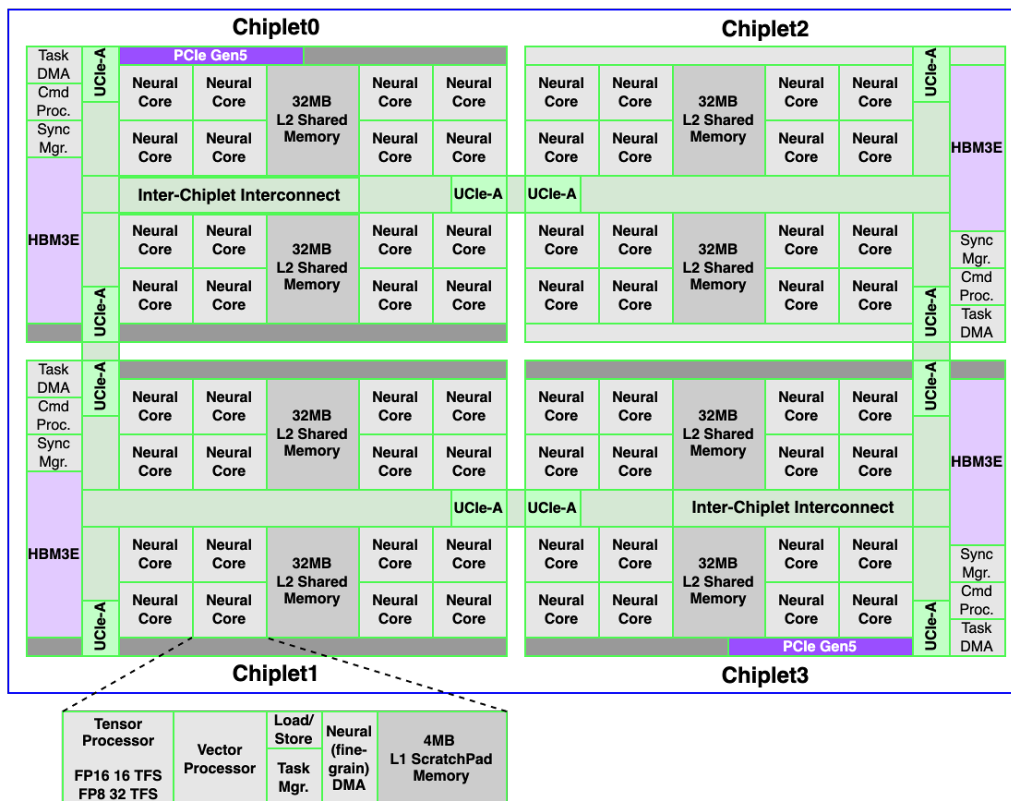
Peta-Scale SoC for Massive AI Serving: REBEL-Quad

REBEL-Quad is an advanced AI SoC built on a UCle-Advanced chiplet architecture, designed to meet the extreme compute and memory demands of serving frontier LLMs at massive scale for hyperscalers, AI data centers, and enterprises.

It features a unified hardware-software stack that maximizes utilization and delivers exceptional performance-per-watt across both compute-bound prefill and memory-bound decoding phases. The chiplet-based design enables hyper scalability without compromising latency or coherence.



[Peta-Scale SoC for Massive AI Serving: REBEL-Quad]



[Figure 1. Block diagram of REBEL-Quad with four homogeneous chiplets]

Solving Challenges of LLM

REBEL-Q introduces architectural solutions designed for energy-efficient AI inference at massive scale.

Unified Mixed Precision Compute Engine

Combines FP8/FP16/FP32 arithmetic in a single core, boosting compute density by 2.8× and utilization by 16%.

Predictive DMA & On-Chip Mesh

2.7 TB/s effective bandwidth with software-tuned DMA, backed by a 3.3× faster mesh fabric.

REBEL-Quad's exceptional performance-per-watt, starts from its proprietary IP cores. A high-bandwidth on-chip mesh interconnect ensures seamless communication across all cores.

Holistic Synchronization

Hardware-accelerated peer-to-peer and hierarchical communication enabling seamless distributed execution.

Custom Die-to-Die Protocol

1 TB/s per channel with 11ns inter-chiplet latency, while preserving modularity and enabling various chiplet expansions.

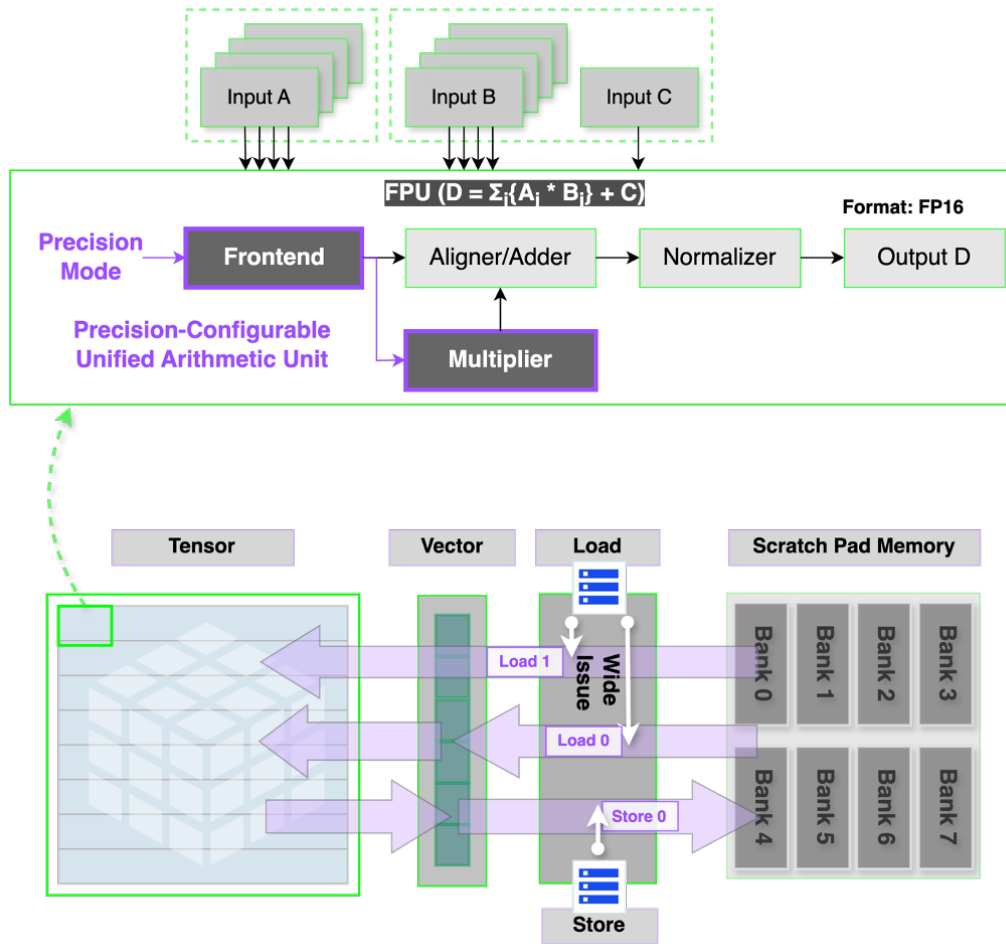
The software stack is tightly coupled to the underlying hardware to enable frictionless scalability to serve frontier models. Backed by a custom die-to-die protocol, REBEL-Quad reliably supports the massive inference demands for hyperscalers and AI data centers.

Unified Mixed Precision Compute with Chiplet-Scale Scalability

Traditional NPUs rely on separate arithmetic blocks for different precision formats (e.g., FP8, FP16, BF16), leading to inefficiencies in both area and dataflow scheduling. REBEL-Quad introduces a unified arithmetic engine supporting per-operand configurable precision, eliminating the need for separate functional units.

This design enables:

- **2.8× higher compute density compared to legacy implementations**
- **16% higher average core utilization**
- **Reduced instruction dependency via hardware-managed wide-issue execution**



[Figure 2. Unified multi-/mixed-precision arithmetic core]

The wide-issue mechanism balances memory bandwidth across tensor and vector cores, ensuring simultaneous access to registers and scratchpad memory. These improvements are particularly beneficial in compute-heavy stages of LLM inference, where sustained FP8 throughput is essential. REBEL-Quad achieves 2 PFLOPS (FP8) compute within a single-node four-chiplet package, significantly enhancing performance-per-watt.

Predictive DMA and High-Bandwidth Memory Access

In the decoding phase, token-by-token generation is limited by KV cache memory bandwidth, which scales poorly with longer context windows. To mitigate this, REBEL-Quad implements a predictive, software-configurable DMA engine capable of:

- **2.7 TB/s effective memory bandwidth**
- **Simultaneous local and remote HBM access**
- **Multi-path routing for bandwidth interleaving**

This DMA is tightly integrated with REBEL-Quad's custom on-chip mesh interconnect, providing 3.3× higher per-core bandwidth over previous architectures. The DMA engine also supports per-task QoS, minimizing latency spikes and dependency stalls across long-tail workloads.

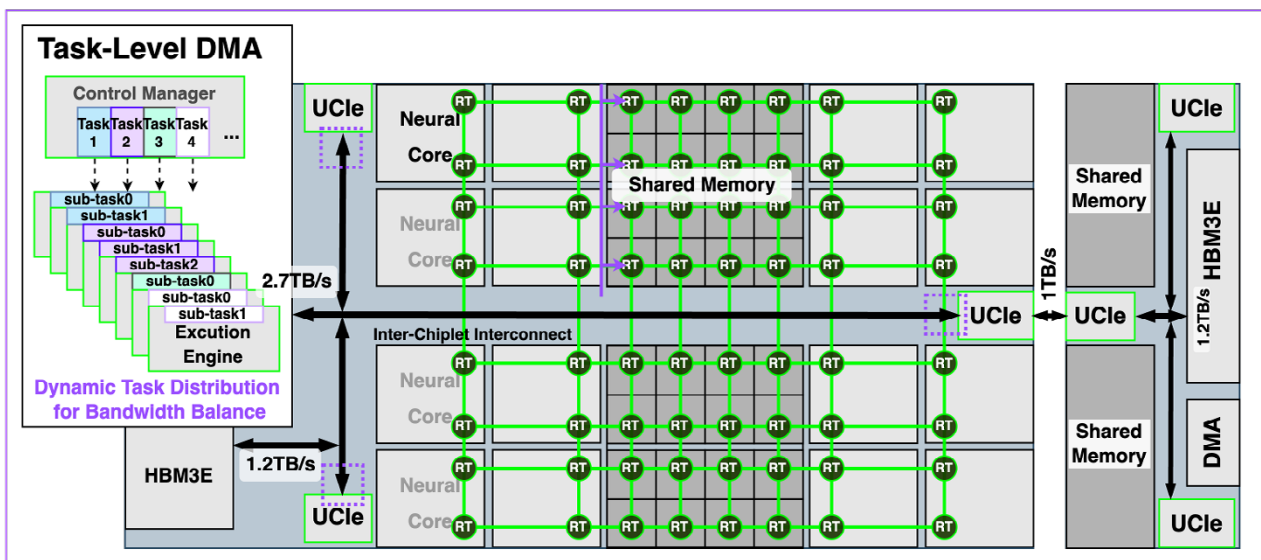
Hierarchical Synchronization and Peer Communication

To sustain performance across distributed workloads, especially in models with complex attention patterns and long-range dependencies, REBEL-Quad employs a full-chip synchronization and communication mechanism.

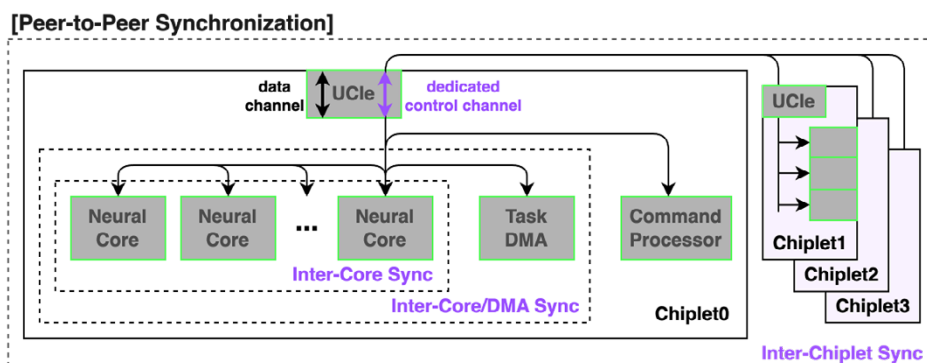
Key mechanisms include:

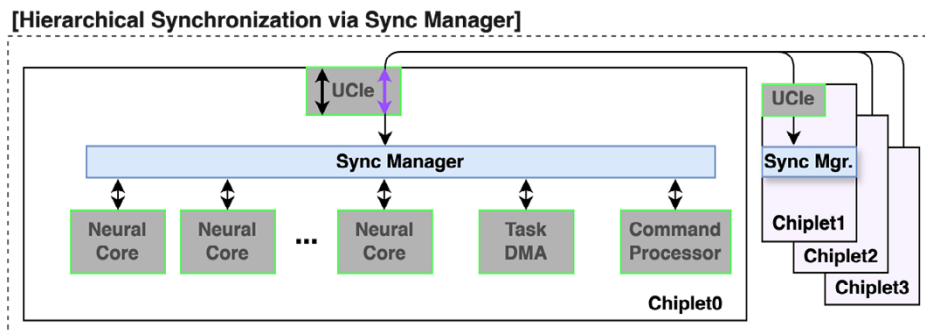
- A dedicated virtual channel for control signals across the mesh network
- A centralized synchronization manager that orchestrates execution flow
- Hardware-accelerated peer-to-peer communication between cores, DMAs, and synchronization units

The hierarchical communication protocol supports both fine-grained intra-chiplet coordination and scalable inter-chiplet dependency resolution, ensuring maximum utilization across all neural cores during concurrent prefill and decoding. This architecture avoids traditional synchronization bottlenecks and minimizes software overhead, enabling compute-dense execution, high utilization rate, and consequently, maximum performance-per-watt.



[Figure 3. Full-chip data transfer utilizing neural cores and DMA engines]





[Figure 4. Full-chip peer-to-peer and hierarchical synchronization scheme using the per-chiplet centralized sync manager]

Scalable Die-to-Die Protocol for Modular Expansion

REBEL-Quad's modular design is enabled by a custom die-to-die protocol based on UClc-Advanced, offering:

- 1TB/s bi-directional per-channel throughput and 11ns full-path inter-chiplet latency
- Load-store memory semantics across dies
- Future-proofed via flexible scale-up and scale-out

This interconnect transforms the multi-chip system into a virtually monolithic unit, while preserving modular scalability for future system expansion. Each chiplet communicates via three UClc channels, with topology-aware die rotation ensuring horizontal mesh continuity. The protocol is paired with a robust switch network and real-time debug mechanisms to support high-reliability, error-free operation, essential for supporting massive AI serving demands of frontier models, especially for hyperscalers, AI data centers, and enterprises. Future extensibility is planned through I/O and memory expander chiplets, enabling even broader system configurations with minimal redesign.

REBEL-Quad delivers the performance, efficiency, and scalability needed to serve next-generation LLMs—without compromise. Its chiplet-based design enables modular upgrades and long-term adaptability, making it an ideal foundation for enterprise-scale AI systems.

Ready to build your next AI platform with REBEL-Quad? Visit rebellions.ai.

Contact Us

rebellions.ai

contact@rebellions.ai