COMPUTE & AI

Garmo-Transformer-E

AI Compute — dedicated transformer attention/MoE engine

IP-2069 · Evaluation package · Evaluation / IP scope

Function and architecture

- Fused QKV projection, RoPE, RMSNorm/LayerNorm, softmax, attention and MLP epilogues
- Paged KV cache, FP8/INT8/INT4 KV compression, speculative-decode assist
- Block-sparse and sliding-window attention, sequence parallel dispatch

Integration summary

Garmo-Transformer-E integration role: AI Compute — dedicated transformer attention/MoE engine. Documented composition: Fused QKV projection, RoPE, RMSNorm/LayerNorm, softmax, attention and MLP epilogues; Paged KV cache, FP8/INT8/INT4 KV compression, speculative-decode assist; Block-sparse and sliding-window attention, sequence parallel dispatch.

Technical record

Canonical identity	IP-2069
Responsibility	AI Compute — dedicated transformer attention/MoE engine
Documented interface architecture	AXI4/HBM ports, coherent KV-cache port, command queue, token stream in/out
Documented building blocks	Fused QKV projection, RoPE, RMSNorm/LayerNorm, softmax, attention and MLP epilogues; Paged KV cache, FP8/INT8/INT4 KV compression, speculative-decode assist; Block-sparse and sliding-window attention, sequence parallel dispatch
Delivery boundary	Evaluation / IP scope
Evidence scope	Evaluation package
Architecture targets · not measurements	64K-token native attention window; >90% array occupancy target; 32K FP8 MAC/cycle; token scheduling without CPU intervention.

Delivery and release binding

- Interface version and parameter set
- Clock/reset domains and integration constraints
- Companion memory, PHY and software requirements
- Verification collateral and acceptance criteria

Confirm for the selected release: Node/library binding, Clock/reset limits, Physical and protocol dependencies, Artifact hashes, Acceptance criteria.

Evidence and qualification

Canonical technology and its architecture/integration role are recorded. The offered delivery is confirmed against the exact release before licensing.

No node-port, silicon-performance, standards certification or analog qualification is inherited from a related product.

Canonical status: P1 / REPRESENTATIVE ARCHITECTURE RTL — concrete integration specification and synthesizable RTL shell/representative datapath only; all frequency, throughput, power, area, jitter and standards-superiority figures remain architecture targets until the required evidence hierarchy closes.

Evaluation and licensing

Evaluation: Evaluate an individual IP configuration. Production: Named-design production license. Custom work: Node binding or derivative integration. Rights, pricing, support and acceptance criteria are agreed for the exact scope.

Contact: patrick@garmolabs.com | Product path: /ip/garmo-transformer-e