COMPUTE & AI

Garmo-NPU-256

AI Compute — 256 tensor-slice edge NPU

IP-2066 · Evaluation package · Evaluation / IP scope

Function and architecture

- 256 tensor slices
- each slice executes a 256-element INT4 dot product per cycle
- INT2/INT4/INT8, FP4 E2M1/E3M0, FP8 E4M3/E5M2, BF16 accumulation
- Native 2:4 and programmable N:M sparsity
- zero-skip before operand fetch

Integration summary

Garmo-NPU-256 integration role: AI Compute — 256 tensor-slice edge NPU. Documented composition: 256 tensor slices; each slice executes a 256-element INT4 dot product per cycle; INT2/INT4/INT8, FP4 E2M1/E3M0, FP8 E4M3/E5M2, BF16 accumulation; Native 2:4 and programmable N:M sparsity; zero-skip before operand fetch.

Technical record

Canonical identity	IP-2066
Responsibility	AI Compute — 256 tensor-slice edge NPU
Documented interface architecture	AXI4 512-bit data master, AXI4-Lite CSR, 512-bit ready/valid tensor stream, IRQ/RAS/DFT
Documented building blocks	256 tensor slices; each slice executes a 256-element INT4 dot product per cycle; INT2/INT4/INT8, FP4 E2M1/E3M0, FP8 E4M3/E5M2, BF16 accumulation; Native 2:4 and programmable N:M sparsity; zero-skip before operand fetch
Delivery boundary	Evaluation / IP scope
Evidence scope	Evaluation package
Architecture targets · not measurements	196.6 TOPS INT4 dense @ 1.5 GHz; 393 TOPS effective with 2:4 sparsity; target <10 W core power on advanced node.

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-npu-256