COMPUTE & AI

Garmo-NPU-1024

AI Compute — 1024 tensor-slice cloud/robotics NPU

IP-2067 · Evaluation package · Evaluation / IP scope

Function and architecture

- 1024 tensor slices × 512-element FP4/INT4 dot-product datapath
- FP4/FP8/BF16/FP16/TF32/INT2/4/8 with FP32 accumulation option
- Hierarchical multicast NoC, 64 MiB distributed SRAM, direct HBM controller attachment

Integration summary

Garmo-NPU-1024 integration role: AI Compute — 1024 tensor-slice cloud/robotics NPU. Documented composition: 1024 tensor slices × 512-element FP4/INT4 dot-product datapath; FP4/FP8/BF16/FP16/TF32/INT2/4/8 with FP32 accumulation option; Hierarchical multicast NoC, 64 MiB distributed SRAM, direct HBM controller attachment.

Technical record

Canonical identity	IP-2067
Responsibility	AI Compute — 1024 tensor-slice cloud/robotics NPU
Documented interface architecture	16x AXI/NoC memory ports, 1024-bit command/data fabric, coherent host port, IRQ/RAS/DFT
Documented building blocks	1024 tensor slices × 512-element FP4/INT4 dot-product datapath; FP4/FP8/BF16/FP16/TF32/INT2/4/8 with FP32 accumulation option; Hierarchical multicast NoC, 64 MiB distributed SRAM, direct HBM controller attachment
Delivery boundary	Evaluation / IP scope
Evidence scope	Evaluation package
Architecture targets · not measurements	2.62 PFLOPS/POPS FP4/INT4 dense @ 2.5 GHz; 5.24 effective with 2:4 sparsity; target <125 W compute core.

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-1024