BASIC EDITION

DeepNPU™ Robotic

On-platform VLA inference for robots, drones and AMRs

The whole perception-to-action loop runs on the module. A licensable low-bit GEMM accelerator IP delivered as synthesisable RTL — process-agnostic, driven by a RISC-V or Arm host over a 5-operation coprocessor interface.

40 TOPS
A8B4 peak

16,384
MAC units · 64 × 64 PE array

WHY IT MATTERS ON A MOVING PLATFORM

No host in the inner loop

Three operations on the critical path — set, start, wait. No MMIO polling loop, no driver round trip.

One engine, three heads

Vision, language and action heads execute on the same array — no second accelerator, no partition to manage.

Mature node, edge power

Transformer and CNN workloads that normally demand a leading-edge node run inside an edge power budget.

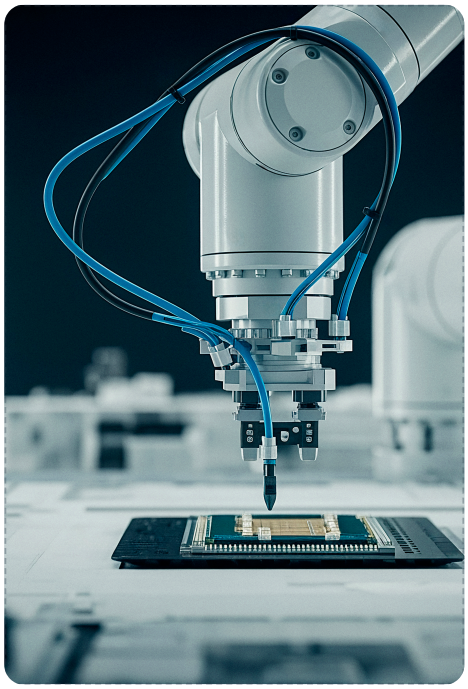
WHERE IT RUNS



Autonomous mobile robots
Navigation, obstacle response and pick instructions from one on-board model.



Inspection drones
Detect, describe and re-plan in flight with no uplink dependency.



Manipulation cells
Language-instructed grasping with action output on the same engine.

Architecture, specifications and delivery

DeepMentor

HOW IT RUNS · SET · START · WAIT

01 Set

One kernel descriptor — shape, addresses, precision, scale, activation — written in a single coprocessor operation.

02 Start

The data mover streams tiles into banked SRAM while the array computes weight-stationary; fetch and compute overlap.

03 Wait

FP32 accumulation and the activation LUT run in place; the next descriptor is already queued.

INSIDE THE BLOCK

Reconfigurable array

CGRA systolic fabric of 8 × 8 PE tiles, 64 × 64 array at 1.2 GHz.

Low-bit GEMM

A8B4 / A8B8 datapaths, FP32 accumulation, in-line activation LUT.

Coprocessor interface

Five ops — set, start, wait, read_csr, abort — from the host core.

Clean SoC boundary

AXI with a dedicated 512-bit NPU port, clear of the CPU cluster.

SPECIFICATIONS · DEEPNPU ROBOTIC

PARAMETER	VALUE	NOTES
Peak throughput	40 TOPS	A8B4 at 1.2 GHz · 20 TOPS A8B8
Compute array	16,384 MAC	8 × 8 tiles of 8 × 8 PE, 64-lane SIMD
Precision	A8B4 / A8B8	INT8 activation, INT4/INT8 weights · FP32 / BF16 / BFP8 out
Memory interface	AXI, 512-bit	Dedicated NPU port, 4 TB address space
Control interface	Coprocessor IF	RISC-V or Arm host; MMIO register map for bring-up
Model support	7B / 13B / 27B	VLA models supported on the same accelerators
Delivery	RTL source	Synthesisable, process-agnostic; no foundry lock-in

Peak figures are architectural at the stated frequency; achieved throughput depends on layer shape and memory bandwidth.

WHAT SHIPS WITH THE LICENCE

Synthesisable RTL

Parameterised SystemVerilog with synthesis and lint constraints.

Verification package

Directed and randomised testbenches; full-scale regression under Verilator.

Bit-accurate golden model

C reference for the full datapath, for RTL co-simulation and bring-up.

Software stack

Compiler flow, runtime and bare-metal drivers with descriptor examples.

Integration support

Interconnect, clocking and memory-path guidance; custom configuration on request.

Model enablement

Quantisation to A8B4 / A8B8 and few-shot retraining without accuracy loss.

DeepMentor

DeepMentor Inc. builds multimodal AI inference for the edge: hardened AI IP, on-device inference, sensor fusion and fast customer-specific training.

Taipei & Hsinchu, Taiwan · Chantilly, USA
+886-2-6604-3279
sales@deepmentor.ai