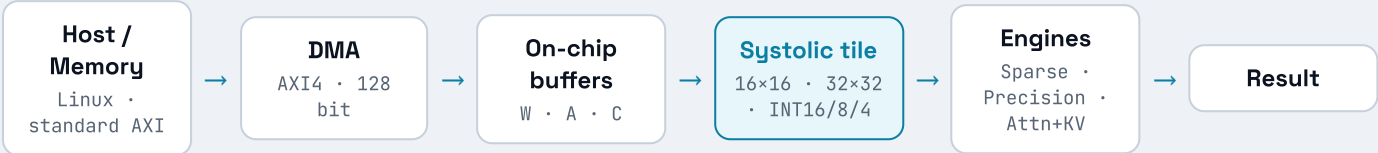


Vexon-1 is an edge-AI NPU IP core for integration into customer SoCs — verified bit-exact on **two FPGA platforms** (AMD Kria XCK26, Zynq UltraScale+ XCZU15EG) and designed for ASIC implementation. It enables local AI inference without cloud dependency wherever **data protection, latency and deterministic processing** matter. Developed and verified entirely in Germany.

ARCHITECTURE



Descriptor-driven: The host writes **one descriptor** — the NPU executes the complete matrix multiplication across all tiles on its own (load, compute, write back, weight residency). No polling traffic, no host choreography.

KEY FEATURES

- **IP architecture** — vendor-independent and auditable, with no proprietary NPU dependencies and no vendor lock-in.
- **Programmable sequencing** — an integrated RV32I core acts as sequencer alongside the hardware FSM: compute operations and data movement are controlled by firmware; new operations require no RTL change. The **data schedule shipped with the core is already optimised** (factor 1.3 over a naive command sequence, bit-identical result); programmability allows it to be tuned further per model and load profile.
- **Multi-precision** — native INT16 / INT8 / INT4, selectable to match model and workload; all three precisions implemented in hardware and verified bit-exact.
- **Transformer-native** — hardware-accelerated attention and on-chip KV cache for efficient transformer inference.
- **Standard integration** — AXI4 for data paths, AXI4-Lite for control/status registers; drops directly into AXI-based SoC architectures.
- **Hardware validation** — proven in operation on two systems: **AMD Kria XCK26** (16×16 array, 32-bit data path) and **Zynq UltraScale+ XCZU15EG** (one 32×32 tile or up to three 16×16 tiles, 128-bit data path), each at 100 MHz with timing closed.
- **Five core architecture mechanisms** — sparse attention fusion · adaptive precision · HW KV cache · HW attention · multi-tile. Three of them are registered as **utility models** with the German Patent and Trade Mark Office, DPMA (No. 20 2026 002 048, 20 2026 002 051, 20 2026 002 052).

CONFIGURATION

Parameter	Range	Description
Array size per tile	16×16 · 32×32	Edge length of a tile's systolic MAC array — 16×16 = 256, 32×32 = 1,024 MAC units. The upper limit is set by the area budget of the target device, not by the RTL.
Number of tiles	1 – 4	A tile is a complete compute island: its own MAC array, its own on-chip buffers and its own address space. Tiles work independently and in parallel. Up to three tiles have been built in hardware and each proven bit-exact individually, with no frequency loss; a joint computation across several tiles has not been measured yet.
Data path width	32 · 128 bit	Width of the path from the on-chip buffers into the array and of the AXI4 data path between DMA and buffers.
Precision	INT16 / INT8 / INT4	Arithmetic precision, selectable per job; includes zero skipping for sparse operands.

Target platform: No parameter separates FPGA and ASIC — the RTL is technology-neutral; the only difference is the mapping (on-chip buffers to BRAM or SRAM macros, MAC units to DSP blocks or standard cells).

VERIFICATION & VALIDATION

Verification (simulation)	Bit-exact verification against a Python reference model (golden model): 14 parametric testbenches with more than 10,000 automated test cases in total, every configuration in several parameter sets. Code coverage over the core RTL: 92.0 % point coverage and 94.3 % line coverage; results identical on two simulators.
Validation platform 1	AMD Kria (XCK26) — 16×16 array, 32-bit data path, 100 MHz, INT16/8/4. All five architecture mechanisms proven in hardware (9/9 correctness tests); transformer inference with a real language model (TinyLlama-1.1B) demonstrated with the NPU in the data path.
Validation platform 2	Zynq UltraScale+ (XCZU15EG) — one 32×32 tile or 1–3 16×16 tiles, 128-bit data path, 100 MHz, INT16/8/4. All five mechanisms validated through the complete driver path (own board Linux, DMA), results bit-identical to the reference model. At start-up the sequencer firmware checks the hardware configuration and refuses to start on a mismatch.
Measurement path	The measurement path runs entirely on the target system: host application (C++) on the SoC's ARM core → descriptor to the NPU → own DMA loads weights and activations into the on-chip buffers → array computes → result returned to memory via DMA → comparison against the reference result on the same board. No host PC in the data path: what we measure is the path the customer will integrate.

Context. The FPGA prototype is a proof of correctness and architecture at 100 MHz; throughput and energy efficiency scale with silicon (higher clock, wider memory subsystem). **System context** (not an IP-to-IP comparison): the same end-to-end matrix multiplication runs 4.6× faster on the XCZU15EG with the NPU than on all four Cortex-A53 cores of the same SoC. Area, power and PPA figures are provided in the **Data Sheet** (on request, under NDA).

DELIVERABLES (IP PACKAGE)

RTL	Synthesisable Verilog core (NPU + descriptor sequencer as FSM and RV32I core + DMA with telemetry), parameterisable (tiles, array size, data path width, precision).
Verification	Parametric testbenches + regression suite (bit-exact, every configuration in several parameter sets), coverage and test case documentation, test report.
Software	C++ driver (device · DMA · NPU · buffer), offline weight preparation (compiler front stage), descriptor API, benchmark suite.
Firmware SW	Sequencer firmware for the RV32I core including toolchain and optimised data schedule; plus a ready-to-run board Linux based on PetaLinux — driver and tests build natively on the target system.
Documentation	Data Sheet, design specification, integration guide, full measurement documentation.

APPLICATIONS

Industry & manufacturing

Visual inspection, quality control, predictive maintenance (vibration/acoustics).

Critical infrastructure & public sector

Local detection, smart grid, network anomalies — no data leaves the site.

Defence & security

On-device AI without cloud dependency; auditable IP.

Medical technology

Edge pre-screening, vital sign / ECG analysis — local and privacy-compliant.

Automotive & mobility

Local low-latency inference, sensor fusion at the edge.

Agriculture & logistics

Image and sensor analysis directly in the field or at the machine.

INTEGRATION & LICENSING

Vexon-1 is a fully in-house, protected chip design developed in Germany, built as an NPU IP block for SoC integration (standard AXI interfaces). Models: **IP licence + royalty per chip**; alternatively an OEM module. Terms are confidential. **ASIC path** (together with a design/tape-out partner): multi-tile scaling, wide memory subsystem (HBM/LPDDR5), higher clock. Next steps: proof of concept (specific use case) → partnership / co-development → ASIC.