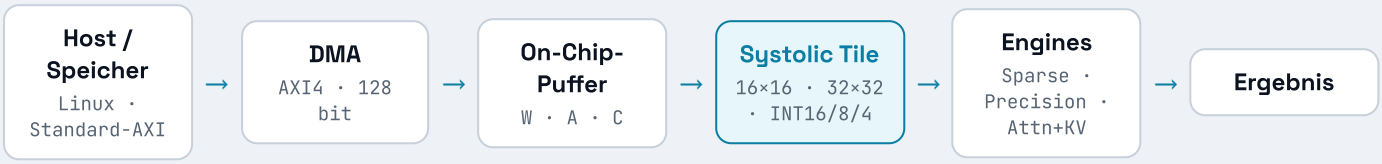
NEUROVEXON · Vexon-1

Edge-AI NPU IP-Core — Product Brief

NPU-IP-KERN · RV32I-SEQUENCER · INT16/8/4 · VALIDIERT AUF 2 FPGA-PLATTFORMEN · 3 GEBRAUCHSMUSTER (DPMA)

Vexon-1 ist ein Edge-AI-NPU-IP-Kern zur Integration in kundeneigene SoCs — auf **zwei FPGA-Plattformen** (AMD Kria XCK26, Zynq UltraScale+ XCZU15EG) bit-exakt verifiziert und für die ASIC-Implementierung ausgelegt. Er ermöglicht lokale KI-Inferenz ohne Cloud-Abhängigkeit, wo **Datenschutz, Latenz und deterministische Verarbeitung** zählen. Entwicklung und Verifikation vollständig in Deutschland.

ARCHITEKTUR



Deskriptor-gesteuert: Der Host schreibt **einen Deskriptor** — die NPU führt die vollständige Matrixmultiplikation über alle Kacheln selbstständig aus (Laden, Rechnen, Zurückschreiben, Weight-Residency). Kein Poll-Verkehr, keine Host-Choreographie.

KERN-MERKMALE

- **IP-Architektur** — vendor-unabhängig und auditierbar, ohne proprietäre NPU-Abhängigkeiten und Vendor-Lock-in.
- **Programmierbare Ablaufsteuerung** — integrierter RV32I-Kern als Sequencer neben der Hardware-FSM: Rechenoperationen und Datenbewegungen werden über Firmware gesteuert; neue Abläufe erfordern keine RTL-Änderung. Der **mitgelieferte Datenfahrplan ist bereits optimiert** (Faktor 1,3 gegenüber einer naiven Kommandofolge, bit-identisches Ergebnis); die Programmierbarkeit erlaubt, ihn je Modell und Lastprofil weiter anzupassen.
- **Multi-Precision** — native Unterstützung für INT16 / INT8 / INT4, wählbar entsprechend Modell und Rechenlast; alle drei Präzisionen hardwareseitig implementiert und bit-exakt verifiziert.
- **Transformer-native** — hardware-beschleunigte Attention und On-Chip-KV-Cache für effiziente Transformer-Inferenz.
- **Standardisierte Integration** — AXI4 für Datenpfade, AXI4-Lite für Control-/Statusregister; direkte Integration in AXI-basierte SoC-Architekturen.
- **Hardware-Validierung** — im Betrieb nachgewiesen auf zwei Systemen: **AMD Kria XCK26** (16×16-Array, 32-bit-Datenpfad) und **Zynq UltraScale+ XCZU15EG** (1 Kachel 32×32 bzw. bis zu 3 Kacheln 16×16, 128-bit-Datenpfad), jeweils 100 MHz mit gehaltenem Timing.
- **Fünf zentrale Architekturmechanismen** — Sparse-Attention-Fusion · Adaptive Precision · HW-KV-Cache · HW-Attention · Multi-Tile. Drei davon sind als **Gebrauchsmuster** beim DPMA eingetragen (Nr. 20 2026 002 048, 20 2026 002 051, 20 2026 002 052).

KONFIGURATION

Parameter	Wertebereich	Beschreibung
Array-Größe je Kachel	16×16 · 32×32	Kantenlänge des systolischen MAC-Arrays einer Kachel — 16×16 = 256, 32×32 = 1.024 MAC-Einheiten. Die Obergrenze setzt das Flächenbudget des Zielbausteins, nicht das RTL.
Anzahl Kacheln	1 – 4	Eine **Kachel** ist eine vollständige Recheninsel: eigenes MAC-Array, eigene On-Chip-Puffer, eigener Adressraum. Kacheln arbeiten unabhängig und parallel. In Hardware sind bis zu 3 Kacheln gebaut und jeweils einzeln bit-exakt nachgewiesen, ohne Frequenzverlust; eine gemeinsame Berechnung über mehrere Kacheln ist noch nicht gemessen.
Datenpfad-Breite	32 · 128 bit	Breite des Datenpfads vom On-Chip-Puffer in das Array und des AXI4-Datenpfads zwischen DMA und Puffer.
Präzision	INT16 / INT8 / INT4	Rechengenauigkeit, je Auftrag wählbar; einschließlich Zero-Skipping bei dünn besetzten Operanden.

Zur Zielplattform: Kein Parameter trennt FPGA und ASIC — das RTL ist technologieneutral; der Unterschied liegt allein in der Abbildung (On-Chip-Puffer auf BRAM bzw. SRAM-Makros, MAC-Einheiten auf DSP-Blöcke bzw. Standardzellen).

VERIFIKATION & VALIDIERUNG

Verifikation (Simulation)	Bit-exakte Verifikation gegen ein Python-Referenzmodell (Golden Model): **14 parametrische Testbenches** mit zusammen über **10.000 automatisierten Testfällen**, jede Konfiguration in mehreren Parametrierungen. Code-Coverage über das Kern-RTL: **92,0 %** Punkt-Coverage bzw. **94,3 %** Zeilen-Coverage; Ergebnisse auf zwei Simulatoren identisch.
Validierung Plattform 1	**AMD Kria (XCK26)** — 16×16-Array, 32-bit-Datenpfad, 100 MHz, INT16/8/4. Alle fünf Architekturmechanismen in Hardware nachgewiesen (9/9 Korrektheitstests); Transformer-Inferenz mit realem Sprachmodell (TinyLlama-1.1B) mit der NPU im Datenpfad demonstriert.
Validierung Plattform 2	**Zynq UltraScale+ (XCZU15EG)** — 1 Kachel 32×32 bzw. 1–3 Kacheln 16×16, 128-bit-Datenpfad, 100 MHz, INT16/8/4. Alle fünf Mechanismen über den vollständigen Treiberpfad (eigenes Board-Linux, DMA) validiert, Ergebnisse bit-identisch zum Referenzmodell. Die Sequencer-Firmware prüft beim Start die Hardware-Konfiguration und verweigert den Start bei Fehlkonfiguration.
Messpfad	Der Messpfad liegt vollständig auf dem Zielsystem: Host-Anwendung (C++) auf dem ARM-Kern des SoC → Deskriptor an die NPU → eigener DMA holt Gewichte und Aktivierungen in die On-Chip-Puffer → Array rechnet → Ergebnis über DMA zurück in den Speicher → Vergleich gegen das Referenzergebnis auf demselben Board. Kein Host-PC im Datenpfad: gemessen wird der Pfad, den der Kunde später integriert.

Einordnung. Der FPGA-Prototyp ist Korrektheits- und Architekturnachweis bei 100 MHz; Durchsatz und Energieeffizienz skalieren mit dem Silizium (höherer Takt, breiteres Speicher-Subsystem). **Systemkontext** (kein IP-zu-IP-Vergleich): Dieselbe End-to-End-Matrixmultiplikation läuft auf dem XCZU15EG mit der NPU 4,6-mal schneller als auf allen vier Cortex-A53-Kernen desselben SoC. Fläche, Leistungsaufnahme und PPA-Kennzahlen stehen im **Data Sheet** (auf Anfrage unter NDA).

LIEFERUMFANG (IP-PAKET)

`RTL`	Synthetisierbarer Verilog-Kern (NPU + Deskriptor-Sequencer als FSM **und** RV32I-Kern + DMA mit Telemetrie), parametrierbar (Kacheln, Array-Größe, Datenpfad-Breite, Präzision).
`Verifikation`	Parametrische Testbenches + Regressionssuite (bit-exakt, jede Konfiguration in mehreren Parametrierungen), Coverage- und Testfall-Dokumentation, Test-Report.
`Software`	C++-Treiber (Device · DMA · NPU · Buffer), Offline-Gewichts-Vorbereitung (Compiler-Vorstufe), Deskriptor-API, Benchmark-Suite.
`Firmware-SW`	Sequencer-Firmware für den RV32I-Kern inklusive Toolchain und optimiertem Datenfahrplan; dazu ein lauffähiges Board-Linux auf **PetaLinux**-Basis — Treiber und Tests bauen nativ auf dem Zielsystem.
`Dokumentation`	Data Sheet, Design-Spezifikation, Integrations-Guide, vollständige Mess-Doku.

ANWENDUNGEN

Industrie & Produktion
Visual Inspection, Qualitätskontrolle, Predictive Maintenance (Vibration/Akustik).

KRITIS & öffentl. Sektor
Lokale Detektion, Smart Grid, Netz-Anomalien — kein Datenabfluss.

Verteidigung & Sicherheit
On-Device-KI ohne Cloud-Abhängigkeit; auditierbare IP.

Medizintechnik
Edge-Pre-Screening, Vital-Sign-/EKG-Analyse — lokal, datenschutzkonform.

Automotive & Mobilität
Lokale Inferenz mit niedriger Latenz, Sensor-Fusion am Edge.

Landwirtschaft & Logistik
Bild-/Sensoranalyse direkt im Feld bzw. an der Anlage.

INTEGRATION & LIZENZIERUNG

Vexon-1 ist ein vollständig eigenentwickeltes, geschütztes Chip-Design, entwickelt in Deutschland, als NPU-IP-Block zur SoC-Integration ausgelegt (Standard-AXI-Schnittstellen). Modelle: **IP-Lizenz + Royalty pro Chip**; alternativ OEM-Modul. Konditionen vertraulich.
ASIC-Pfad (gemeinsam mit Design-/Tape-out-Partner): Multi-Tile-Skalierung, breites Speicher-Subsystem (HBM/LPDDR5), höherer Takt. Nächste Schritte: Proof of Concept (konkreter Use-Case) → Partnerschaft / Co-Entwicklung → ASIC.