# NXPU: A Silicon-Validated NeuroSymbolic Processing Unit for Deterministic AI Reasoning

Zachary Kleckner
*Dyber, Inc.*
April 2026

**Abstract**

We present NXPU, a purpose-built processor for deterministic logical inference that combines Content-Addressable Memory (CAM) for O(1) parallel fact search, a Datalog rule evaluation engine with variable binding and backtracking, spiking neural networks with STDP learning, and autonomous concept formation on a single chip. Silicon-validated on a Xilinx ZCU106 FPGA (xczu7ev-ffvc1156-2-e), NXPU achieves 100% accuracy across 12 benchmarks in 10 domains including pharmacovigilance, cybersecurity, GDPR compliance, and clinical risk assessment, with 74x lower energy consumption than CPUs and 236,000x lower than GPU-based LLM inference. The architecture requires zero training data and produces zero hallucinations by construction, as all conclusions follow from sound deductive logic. We demonstrate NXLang, a complete compiler toolchain that compiles high-level Datalog programs to hardware register sequences, with a Python SDK and interactive REPL. To our knowledge, NXPU is the first silicon-validated commercial reasoning accelerator that guarantees correctness, explainability, and energy efficiency for safety-critical AI applications.

*Keywords: neurosymbolic AI, hardware reasoning, content-addressable memory, Datalog inference, zero hallucination, FPGA, compiler toolchain*

## 1. Introduction

Large Language Models have achieved remarkable capabilities in text generation, translation, and code synthesis. However, they suffer from three fundamental limitations that prevent deployment in safety-critical applications: hallucination, energy consumption, and training cost.

Hallucination remains the primary barrier to enterprise adoption. In 2026 benchmarks across 37 models, hallucination rates ranged from 3.3% to 52%, with advanced reasoning models (GPT-5, Claude Sonnet 4.5, Grok-4) all exceeding 10% [1][2]. In medical contexts, hallucination rates reach 64.1% without mitigation prompts [3]. The fundamental issue is architectural: transformer-based models perform statistical next-token prediction, not logical deduction. They can generate plausible text about reasoning without actually performing it.

Energy consumption presents a sustainability challenge. A single NVIDIA H100 GPU consumes approximately 700W and delivers about 0.39 J per token for Llama3-70B inference at best-case batch sizes [4]. Training GPT-4 consumed an estimated 1,300 MWh [5]. As AI workloads scale, this energy trajectory is economically and environmentally unsustainable.

Training cost creates prohibitive barriers. State-of-the-art models require trillions of tokens, weeks of compute on thousands of GPUs, and millions of dollars per training run. The resulting models are frozen at training time and require expensive fine-tuning to incorporate new knowledge.

NXPU addresses all three limitations through a fundamentally different computational paradigm: purpose-built silicon for deterministic Datalog forward-chaining inference. Rather than learning statistical patterns, NXPU executes sound logical rules with variable binding, backtracking, and deduplication. Conclusions are mathematically guaranteed to follow from premises — hallucination is impossible by construction.

## 2. Architecture

NXPU integrates nine functional subsystems behind a unified AXI4-Lite register interface (42 registers, 8-bit address space). The architecture is implemented in approximately 15 Verilog modules totaling several thousand lines of RTL, synthesized for the Xilinx ZCU106 evaluation board.

### 2.1 Content-Addressable Memory (CAM)

The CAM stores facts as 56-bit entries comprising an 8-bit predicate identifier and three 16-bit argument values. All 256 entries are compared simultaneously against a search pattern in a single clock cycle using parallel XOR-with-mask comparators, providing O(1) query latency of 10 ns at 100 MHz. A popcount module computes the match count, and a priority encoder enables sequential iteration through matches. Don’t-care masking supports wildcard queries. This parallel search architecture is fundamentally different from sequential database scans and is the key enabler of hardware-speed reasoning [6].

### 2.2 Rule Evaluation Engine

A 12-state finite state machine implements Datalog forward-chaining with full depth-first backtracking. The FSM supports up to 4 body atoms per rule with 8 simultaneous variable bindings. States include SEARCH, WAIT_MATCH, NEXT_MATCH, READ_ENTRY, UNIFY, ADVANCE, DERIVE, INSERT, INSERT_CHECK (deduplication), BACKTRACK, and DONE. When unification fails at any body atom, the FSM backtracks to the previous atom and tries the next CAM match, ensuring complete exploration of the search space [7].

### 2.3 Unifier

The unifier is a purely combinational circuit that matches a pattern containing variable slots against a concrete CAM entry in a single clock cycle. For each argument position: if the variable is unbound, it binds to the entry’s value; if already bound, it checks consistency with the existing binding. This single-cycle pattern matching is the core operation of logical inference [8].

### 2.4 Neural Mesh

Sixteen Leaky Integrate-and-Fire (LIF) spiking neurons with Spike-Timing-Dependent Plasticity (STDP) implement online perceptual learning without backpropagation. The synapse crossbar maintains an 8×16 weight matrix with Q1.7 fixed-point weights. A homeostatic threshold mechanism prevents runaway excitation. Unlike gradient-descent networks, STDP learns from temporal correlations in spike patterns — continuously, locally, and at milliwatt power [9].

### 2.5 Concept Formation

An autonomous learning circuit monitors CAM search results in real-time. When the same argument value appears in three or more failed searches (configurable threshold), the circuit recognizes a recurring unknown pattern, allocates a new predicate identifier starting at ID 200, and writes a new concept fact into the CAM. This is hardware-level inductive learning with zero training data, zero human labeling, and zero intervention. The circuit implements the observe-recognize-name cycle entirely in silicon [10].

### 2.6 Additional Subsystems

Feedback Bus: Three unidirectional FIFO paths enable iterative reasoning: CSE-to-NM reward routing modulates STDP learning rates; NM-to-SLU pattern injection converts perceptual detections to symbolic facts; SLU-to-CSE knowledge updates feed derived facts to the causal graph. Meta-Rule Engine: Monitors per-rule evaluation statistics (success rate, derivation count) and applies self-modification rules — promoting rules with >80% success, demoting those below 20%, and deleting rules with zero success after 50 evaluations. Pattern Comparator: Compares the structural shape of inference chains (hop count, predicate pattern hash, variable sharing structure) across domains, enabling cross-domain analogical reasoning with Q0.8 similarity scoring. NLU Pipeline: Semantic tokenizer (DJB2 hash-based vocabulary with auto-learning), triple extractor (S-V-O pattern recognition), and trace-to-text engine (template-based explanation generation) — all implemented in hardware.

## 3. NXLang Compiler Toolchain

NXPU includes a complete compiler pipeline from high-level source code to hardware register writes, eliminating the need for manual hexadecimal encoding of facts and rules.

### 3.1 Language Design

NXLang v0.3 provides Datalog-style syntax with three primary constructs: facts (ground truth assertions), rules (logical inference specifications with head and body atoms), and queries (predicate-matching requests). Variables are indicated by uppercase identifiers; constants by lowercase. Don’t-care arguments use the underscore wildcard.

### 3.2 Compiler Pipeline

The compiler processes NXLang source through six stages: (1) Lexer with 40+ token types and line/column error reporting; (2) Recursive descent parser producing a typed AST with 15 node types; (3) Type checker with two-pass validation including arity consistency and head-variable-in-body checking; (4) NXIR lowering to an SSA-form intermediate representation with engine annotations; (5) Optimizer performing dead code elimination and parallel region identification; (6) Backend code generation [11].

### 3.3 FPGA Backend

The critical compilation step is the Variable Encoder, which converts human-readable Datalog rules into the packed bit-field format required by the rule_eval FSM. Each variable name is assigned an index (0–7), shared variables across atoms receive the same index (enabling binding consistency), and indices are packed as 3 bits per argument, 9 bits per atom, into a 36-bit field split across two 32-bit registers. The Hardware Encoder then generates the complete sequence of AXI register writes: 3 writes per fact

(DATA_LO, DATA_HI, CMD=ADD_FACT) and 9 writes per rule (8 configuration registers + CMD=DERIVE).

### 3.4 Output Formats and Runtime

The compiler produces four output formats: Vivado Tcl scripts for direct FPGA execution; NXB binary (16-byte header + register writes + symbol table + CRC32) for portable deployment; JSON for debugging and visualization; and C headers for embedded integration. The Python SDK provides a Session API with connect(), add_fact(), add_rule(), derive(), and query() methods, abstracted over JTAG, simulation, and future PCIe/USB transports. An interactive REPL (nxc repl) enables exploratory reasoning sessions.

## 4. Silicon Validation

All results were obtained on a Xilinx ZCU106 evaluation board (xczu7ev-ffvc1156-2-e, Zynq UltraScale+ MPSoC) at 100 MHz. Timing analysis reports WNS = +1.651 ns (timing met with margin). Design rule checks report 0 errors. Measurements use the hardware cycle counter and JTAG-to-AXI debug interface. These are real silicon measurements, not simulation results.

**Table 1: Silicon Benchmark Results (ZCU106 FPGA, April 2026)**

| Benchmark | Domain | Facts | Rules | Derived | Cycles |
|---|---|---|---|---|---|
| Security threats | Cybersecurity | 12 | 3 (3-body) | 5 | 165 |
| Drug interactions | Pharmacovigilance | 15 | 2 (4-body) | 4 | 164 |
| Math reasoning | Education (7 levels) | 24 | 10 | 12 | 214 |
| Novel domains | 4 unseen domains | 27 | 4 | 15 | 317 |
| Dep. vulnerability | DevSecOps | 8 | 3 | 4 | 184 |
| RBAC access control | IAM / Zero Trust | 9 | 2 | 5 | 107 |
| GDPR compliance | Legal / Privacy | 12 | 2 | 3 | 141 |
| Clinical risk | Healthcare | 11 | 3 (diamond) | 3 | 275 |
| Supply chain | Economics | 8 | 2 | 4 | 116 |
| Ecology | Biology | 7 | 1 | 4 | 72 |
| 100-rule stress | Scale test | 50 | 100 | — | 4,400 |
| Code synthesis | Programming | 10 | 3 | 9 | ~60 |

Accuracy is 100% across all benchmarks with zero false positives. The pharmacovigilance benchmark detected a real warfarin-fluconazole drug interaction (a documented cause of bleeding events and patient deaths [12]) through 4-body-atom reasoning — the most complex rule type — with zero false positives and zero false negatives. The novel domain test proved that NXPU reasons correctly about biology, ecology, supply chains, and geography with zero domain-specific training, using only general logical rules (transitivity, inheritance, causal chaining). The GDPR benchmark correctly identified a data processing violation (health records processed without explicit consent) while confirming two compliant operations.

# 5. Performance Comparison

**Table 2: NXPU vs. CPU vs. GPU for Logical Reasoning Tasks**

| Metric | NXPU (FPGA) | Python (x86 CPU) | H100 (LLM) |
|---|---|---|---|
| Query latency | 10 ns (1 cycle) | 370 ns | ~500 ms |
| Derivation (3 rules, 5 facts) | 1.65 µs | 6.10 µs | N/A |
| Energy per derivation | 1.65 µJ | 122 µJ | ~390,000 µJ/token |
| Accuracy (logical reasoning) | 100% | 100% | 80–90% [1][2] |
| Training energy | 0 Wh | 0 Wh | ~1,300 MWh [5] |
| Hallucination rate | 0% (by construction) | 0% | 10–64% [1][3] |
| Native query throughput | 100M/sec | — | — |
| Native derive throughput | 2.3M/sec | — | — |
| Power consumption | ~1 W (PL logic) | ~20 W | ~700 W [4] |

NXPU achieves 74x energy efficiency over CPUs and approximately 236,000x over GPU-based LLM inference for logical reasoning tasks. The Python benchmark (10,000 iterations, median timing) runs the identical 12-fact, 3-rule security benchmark on the same machine. Native throughput at 100 MHz is 100 million queries per second and 2.3 million derivations per second. ASIC projections at 500 MHz–1 GHz target 0.05 µJ per derivation and 1 billion queries per second [13].

# 6. Target Markets

NXPU targets applications where correctness, explainability, and energy efficiency are more important than creative generation. Seven market segments have been identified with demonstrated silicon benchmarks:

**Table 3: Target Market Segments**

| Market | Problem | NXPU Solution | TAM |
|---|---|---|---|
| Network Security | Real-time threat inference at edge | CAM+rules in firewall silicon | $22B |
| Healthcare/Pharma | Drug interaction, clinical decision | 100% accurate, FDA-friendly | $14B |
| IAM / Zero Trust | Hardware policy evaluation | 10 ns per policy check | $5B |
| DevSecOps | Transitive dependency scanning | 3-hop vulnerability chains | $500M |
| Legal / Compliance | GDPR/HIPAA violation detection | Deterministic, auditable | $10B |
| Industrial Safety | Deterministic safety interlocks | Hardware-level correctness | $8B |
| Edge IoT | On-device reasoning | Milliwatt power (ASIC) | $50B+ |

## 7. Competitive Landscape

To our knowledge, no commercially available chip combines all of the following capabilities on the same silicon: O(1) parallel fact search, Datalog rule evaluation with variable binding and backtracking, spiking neural network with STDP learning, autonomous concept formation, cross-domain analogical reasoning, and a complete compiler toolchain. The nearest comparison points are:

GPUs and TPUs (NVIDIA, Google, AMD) are optimized for dense matrix multiplication — the core operation of transformer inference and neural network training. Running Datalog on a GPU requires implementing the inference engine in software on hardware designed for a fundamentally different workload [14]. Neuromorphic chips (Intel Loihi 2 [15], IBM NorthPole, BrainChip Akida, Syntiant) implement spiking neural networks with excellent energy efficiency for perceptual tasks, but lack symbolic reasoning, variable binding, and rule chaining capabilities. TCAM chips (Synopsys, Broadcom) provide O(1) parallel matching — functionally similar to NXPU's CAM — but perform no inference beyond pattern matching [6]. The CoCoSys center (Georgia Tech, DARPA JUMP 2.0) taped out a neuro-symbolic chip in 2025 targeting IMO-level mathematical reasoning [16], but as an academic/military research project with no published commercial benchmarks or zero-hallucination guarantees.

## 8. Scaling Roadmap

The current prototype uses a 256-entry register-based CAM. Three scaling paths are in development: (1) The scalable_cam module provides hash-partitioned BRAM-backed storage for 4K–64K entries, using predicate-based banking for O(1) latency within each bank. At 64K entries, BRAM consumption is approximately 3.6 Mbit (32% of ZU7EV capacity). (2) DDR4 integration via the Zynq PS S_AXI_HP slave port enables million-fact knowledge bases with bulk loading at ~17 GB/s per HP port. (3) A multi-rule sequencer wrapping the existing rule_eval FSM enables fixed-point evaluation of 100+ rules without host intervention.

ASIC projections at 10 nm target 500 MHz–1 GHz clock frequency, approximately 100 mW power consumption, 1–2 mm² die area, and 1 billion queries per second. The form factor roadmap progresses from IP licensing (Verilog RTL, available now) to M.2 AI accelerator module (2027, similar to Hailo-8 [17] and Axelera Metis [18]) to standalone ASIC (2028).

## 9. Conclusion

We have presented NXPU, a silicon-validated neurosymbolic processing unit that achieves 100% accuracy on logical reasoning tasks across 12 benchmarks in 10 domains, with 74x energy efficiency over CPUs and zero training requirements. The architecture's zero-hallucination guarantee is not a benchmark achievement but a mathematical property: deductive logic is sound by construction — if premises are true, conclusions necessarily follow.

The complete NXLang compiler toolchain enables deployment from high-level Datalog programs to FPGA silicon, with a Python SDK for integration. The pharmacovigilance benchmark demonstrates clinical relevance: detecting a life-threatening warfarin-fluconazole drug interaction through 4-body-atom logical reasoning in 164 clock cycles, with zero false positives — a task where LLMs hallucinate up to 64% of the time in medical contexts.

NXPU occupies a previously empty market position: the first commercial reasoning accelerator that guarantees correctness, explainability, and energy efficiency for safety-critical AI applications. For domains where AI must be right — not probably right — purpose-built reasoning silicon provides an answer that no GPU, no LLM, and no software system can match.

## References

[1] Vectara, "Introducing the Next Generation of Vectara's Hallucination Leaderboard," 2026. [Online]. Available: https://www.vectara.com/blog/introducing-the-next-generation-of-vectaras-hallucination-leaderboard

[2] SQ Magazine, "LLM Hallucination Statistics 2026: Hidden Risks Now," 2026. [Online]. Available: https://sqmagazine.co.uk/llm-hallucination-statistics/

[3] M. Brinsa, "Hallucination Rates in 2025 — Accuracy, Refusal, and Liability," Frontiers in Artificial Intelligence, 2025.

[4] TokenPowerBench, "Benchmarking the Power Consumption of LLM Inference," arXiv:2512.03024, 2025.

[5] P. Luccioni et al., "Estimating the Carbon Footprint of BLOOM, a 176B Parameter Language Model," Journal of Machine Learning Research, 2023.

[6] K. Pagiamtzis and A. Sheikholeslami, "Content-Addressable Memory (CAM) Circuits and Architectures: A Tutorial and Survey," IEEE J. Solid-State Circuits, vol. 41, no. 3, pp. 712–727, 2006.

[7] S. Ceri, G. Gottlob, and L. Tanca, "What You Always Wanted to Know About Datalog (And Never Dared to Ask)," IEEE Trans. Knowl. Data Eng., vol. 1, no. 1, pp. 146–166, 1989.

[8] J. A. Robinson, "A Machine-Oriented Logic Based on the Resolution Principle," J. ACM, vol. 12, no. 1, pp. 23–41, 1965.

[9] G. Bi and M. Poo, "Synaptic Modifications in Cultured Hippocampal Neurons: Dependence on Spike Timing, Synaptic Strength, and Postsynaptic Cell Type," J. Neuroscience, vol. 18, no. 24, pp. 10464–10472, 1998.

[10] D. H. Fisher, "Knowledge Acquisition via Incremental Conceptual Clustering," Machine Learning, vol. 2, no. 2, pp. 139–172, 1987.

[11] A. V. Aho, M. S. Lam, R. Sethi, and J. D. Ullman, Compilers: Principles, Techniques, and Tools, 2nd ed. Addison-Wesley, 2006.

[12] T. Baber et al., "Fluconazole-warfarin interaction: A review of the evidence," J. Clin. Pharmacy and Therapeutics, vol. 45, no. 6, pp. 1271–1280, 2020.

[13] N. P. Jouppi et al., "In-Datacenter Performance Analysis of a Tensor Processing Unit," Proc. ISCA, pp. 1–12, 2017.

[14] NVIDIA, "NVIDIA H100 Tensor Core GPU Architecture," Technical Brief, 2022.

[15] M. Davies et al., "Loihi 2: A Neuromorphic Processor with Quantized Sparsity," IEEE Micro, vol. 42, no. 5, pp. 14–22, 2022.

[16] CoCoSys, "CoCoSys Develops Groundbreaking Neuro-Symbolic AI Chip," Georgia Tech ECE News, May 2025.

[17] Hailo Technologies, "Hailo-8 M.2 AI Acceleration Module," Product Datasheet, 2024.

[18] Axelera AI, "Metis M.2 AI Inference Acceleration Card," Product Brief, 2025.