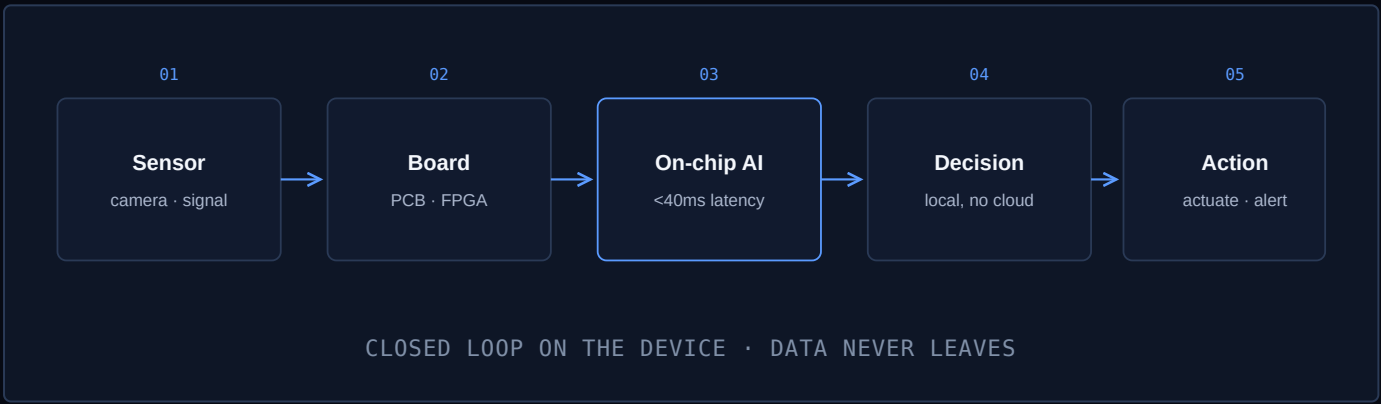
INTELLIGENCE AT THE EDGE, NOT IN THE CLOUD

We Run AI on the Device, Where the Data Is Made.

Cloud inference is expensive, high-latency, and moves your data off the device. We deploy capable models on FPGA and embedded hardware, on-chip, at a fraction of the cost and with nothing leaving the box. **We build the silicon, the board, and the model that runs on it.**



<40ms
first-token latency, by design

18W
target envelope for a capable LLM

20×
modeled cost advantage vs cloud

0
bytes leave the device

THE CLOUD MODEL BREAKS AT THE EDGE

Three Reasons Inference Belongs on the Device

For real-time, regulated, and always-on workloads, sending every frame or token to the cloud fails on cost, latency, or compliance. Often all three.

01

Latency

Below 40ms you cross the human-perception threshold. Real-time voice, vision, and control need on-chip determinism the cloud cannot guarantee under load.

02

Sovereignty

When inference runs on the device, no data leaves it. Compliance, security, and reliability come in one package, without a cloud call.

03

Economics

At a fraction of cloud cost per token, continuous inference becomes affordable. Always-on AI stops being a budget line and becomes a design default.

COST PER MILLION TOKENS, BY WHERE IT RUNS

$0.07	$0.53	$1.40
FPGA on-chip (our stack)	Edge GPU (Jetson class)	Cloud / CPU inference

Illustrative economics on a Kria-class FPGA at commercial power rates, amortised over three years. At 10B tokens/month, that is roughly a $70K/year footprint against a $1.4M/year cloud bill.

SIX EDGE AI CAPABILITY AREAS

From Silicon to Inference

We own every layer of the edge stack. That is why we can put a capable model on an 18-watt board and guarantee it holds in the field.

01

On-chip LLM inference

Capable language models running entirely on FPGA, on-chip weights in BRAM, LoRA adapters swapped at runtime.

02

Edge computer vision

Object detection, defect and anomaly detection, and tracking at line rate on embedded and FPGA hardware.

03

Speech & audio on device

Recognition and speaker diarization tuned for IoT hardware, with overlapped-voice resolution.

04

Predictive maintenance

Vibration, thermal, and acoustic models on-edge that predict failure days ahead, on the machine itself.

05

Hardware security layer

An MCP-based layer that detects a compromised edge device and physically cuts the data lanes. Patent filed.

06

Board & firmware

In-house PCB fabrication, VHDL/Verilog FPGA design, and firmware with a safe field-update path.

WHERE SOFTWARE RUNS OUT OF RUNWAY, HARDWARE TAKES OVER

Four Layers, One Team

Most AI firms stop at the software layer. We go down to the silicon. Each layer below is built in house, which is why the whole system holds together in the field.

01 · SILICON & PCB

In-house fabrication

PCB fabrication at roughly 40 boards per day in our Mysuru facility. VLSI design, full assembly, and rapid prototype-to-validation cycles no pure-software firm can match.

02 · FPGA & EMBEDDED

Custom silicon, microsecond latency

Custom FPGA architecture in VHDL and Verilog. Hardware threat detection, signal processing, and AI inference at the hardware layer. PCIe integration, onboard DDR4.

03 · EDGE AI & SECURITY

Inference and protection on device

AI inference on ASIC and FPGA at a fraction of cloud cost. A hardware security layer that physically disconnects data lanes on breach. Patent filed, shown at RSA.

04 · SOFTWARE & PLATFORMS

Built on the hardware, not floating above it

Model optimization and compression, hybrid edge-cloud orchestration, and regulated deployment. The software layer stands on a foundation we control.

FPGA LANGUAGE MODEL ARCHITECTURE WE DELIVER

A Capable LLM on a $300 Board, No Cloud

We design on-chip inference for FPGA. Base weights sit in on-chip memory; compact LoRA adapters swap at runtime to serve multiple fine-tuned variants from one bitstream. The figures below are the design targets this architecture is built to hit on a Kria-class board.

~187 t/s
target throughput, fully on-chip

18W
target power envelope

$0.07
modeled cost per million tokens

HOW IT WORKS

- Base model weights fixed in on-chip BRAM
- Compact LoRA adapters in configurable SRAM
- Adapter delta applied per transformer layer
- One bitstream serves many fine-tuned variants

WHY IT MATTERS

This is a commercially deployable, multi-tenant inference architecture on $300 hardware. Swap the adapter, change the behaviour. No reflashing, no cloud, no data egress. We scope the target model with you, then build and validate to these design points.

HARDWARE: KRIA-CLASS FPGA · ON-CHIP

MCP-BASED HARDWARE SECURITY LAYER PATENT FILED

When the Edge Device Is Breached, the Data Never Leaves

An AI device in the field is a target. We built a hardware layer that watches the device itself. On breach, it physically disconnects the SSD data lanes before a single byte is exposed.

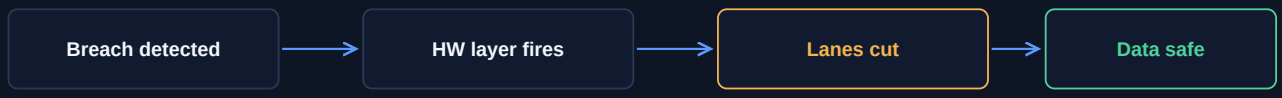
WHAT IT DOES

- Detects compromise at the hardware layer, not in software
- Physically severs SSD data lanes on breach
- Protects data before exfiltration can begin
- Sits underneath the AI workload it protects

THE STANDING OF THE WORK

Described as the first MCP-based hardware security layer of its kind. The patent is filed with our engineers named on the submission, and the system was demonstrated at the RSA Conference. This is the layer no pure-software security firm can build.

STATUS: PATENT FILED · SHOWN AT RSA



BREACH TO PHYSICAL DISCONNECT · FASTER THAN EXFILTRATION

CLINICAL Q&A, ZERO DATA EGRESS ARCHITECTURE WE DELIVER

A Medical Answer System That Keeps PHI on Site

For a hospital staff question-answer system with zero tolerance for hallucination and no patient data leaving the building, the answer is retrieval-augmented generation at the edge, on FPGA-accelerated local inference. The figures below are the targets this design is built to meet.

99%+
target accuracy on medical queries

<100ms
target response time, on device

Zero
PHI data egress by design

THE DESIGN

- Edge-deployed RAG over a validated knowledge base
- FPGA acceleration with a local LLM
- Encrypted data, nothing leaves the premises
- Built for continuous uptime in critical-care settings

WHY IT MATTERS

Grounding holds the hallucination rate near zero, response stays under 100ms on the device, and no PHI ever crosses the network boundary, at a fraction of a cloud deployment's cost. We scope the knowledge base and hardware target with you, then build and validate to these points.

DOMAIN: HEALTHCARE · ON-PREM EDGE

WE GRADE EDGE AI BECAUSE WE SHIP EDGE AI

A Track Record on Real Hardware

These are edge systems we have designed, built, and deployed, across retail, industrial, automotive, and security.

01

Automotive gateway

A dual-variant connected-vehicle gateway. FPGA, PCB layout, embedded firmware, and edge inference, fabricated in house.

02

Industrial predictive maintenance

AI on the edge predicting equipment failure from vibration and thermal signals on the machine itself.

03

Smart retail edge

On-device inventory intelligence and contextual upsell, running at the shelf without a cloud round trip.

04

Edge fraud detection

Real-time fraud detection on edge-based POS terminals, deciding locally in milliseconds.

05

Smart surveillance on edge

AI vision on the camera, detecting and tracking without streaming raw footage off the device.

06

Speech on IoT hardware

Advanced speech recognition tuned to run on constrained IoT devices, on-device and offline.

MODEL TO METAL, IN ONE ENGAGEMENT

Scope, Optimize, Deploy, Sustain

01 Scope › 02 Optimize › 03 Deploy › 04 Sustain

01 · SCOPE
Target and constraints

We fix the workload, the latency budget, the power envelope, and the hardware target before a line of RTL is written.

02 · OPTIMIZE
Compress and map

Model compression, quantization, and mapping to the FPGA dataflow. We make the model fit the metal, not the other way round.

03 · DEPLOY
Board and bitstream

We fabricate the board, load the bitstream, and validate on the bench and in the field against the rated operating range.

04 · SUSTAIN
Update path and fleet

A safe firmware update path and a plan for the fleet, so the device can be patched without a truck roll.

> **Who asks for this.** Teams with a latency, cost, or data-residency wall that the cloud cannot clear. Industrial, healthcare, automotive, retail, and defence.

BRING THE INTELLIGENCE TO THE DEVICE

If Your AI Has a Latency, Cost, or Data Wall, the Edge Is the Answer.

> Send us the workload and the constraint. We will tell you whether it runs on-chip, what it costs, and what board it takes. Then we build it, from the silicon up.

Tech4Biz Solutions

Bengaluru, with an in-house PCB fabrication facility in Mysuru. 80+ engineers across silicon, FPGA, embedded, and AI. We build the hardware that runs edge AI, and the AI that runs on it.

contact@tech4biz.io · tech4bizsolutions.com