LAB L4 | BETA | EDGE AI & SENSING

Ambient Intelligence Fabric

Edge NPU - TinyML - Multimodal - Privacy-by-default

Always-on, low-power inference at the edge - vision, audio and signal where the data is born.

Thesis: Sovereignty, latency and unit economics all point to inference at the edge for ambient workloads.

EDGE INFERENCE P95	BANDWIDTH SAVED	COST PER INFERENCE
9 ms	96%	89%

MANIFESTO

Why this lab exists

Round-tripping every frame to the cloud is dead. Ambient Intelligence Fabric runs vision, audio and signal models on edge NPUs with sub-10ms latency, sovereign by default - only the inference, never the raw stream, leaves the device. Privacy and physics, finally aligned.

KPIS

Outcomes we measure

- Edge inference p95: 9 ms
- Bandwidth saved: 96%
- Cost per inference: 89%

ACTIVE EXPERIMENTS

What the lab is testing now

> Quantised VLMs on consumer NPUs
INT4 / INT8 vision-language models on Snapdragon, Apple Neural Engine and Hailo.

> Federated learning loops
Models improve from edge feedback without raw data ever leaving the device.

> Multimodal sensor fusion
Vision + audio + IMU on a single NPU pipeline for safety and ops use cases.

> WebGPU as edge runtime
Browser-side inference for laptops, kiosks and POS - zero install.

SHIPPABLE ARTEFACTS

Everything that ships

> Edge runtime
ONNX + WebGPU runtime with hot-swap models, signed deploys, observability.

> Quantisation toolkit
INT4 / INT8 / mixed-precision pipelines with eval-preserving guardrails.

> Federated trainer
Aggregated gradients only; differential privacy budget per round.

> Privacy receipts
Per-inference cryptographic receipt: what ran, where, on what data class - never the data itself.

> Reference deploys
Retail loss-prevention, factory safety, contact-centre ambient and fleet POS patterns.

LAB TEAM

Who you'll work with

- Edge AI Principal
- Embedded / NPU Engineer
- Quantisation Researcher
- Privacy Engineer

ENGAGEMENT TIMELINE

Weeks 1-10 - first edge node live by week 3

1. **Weeks 1-3 - Edge node baseline**
   Hardware bring-up, signed deploy, telemetry, first quantised model live.
2. **Weeks 3-6 - Multimodal fusion**
   Vision + audio + signal fused; privacy receipts on every inference.
3. **Weeks 6-10 - Federated loop live**
   Models improve from edge feedback under DP budget; uplift measured.

FLAGSHIP PODS

Squads that productionise this lab

- Retail Loss-Prevention Pod
- Factory Safety Pod
- Contact Centre Ambient Pod
- Fleet & POS Edge Pod

PARTNERS

Who we collaborate with

NVIDIA Jetson - Hailo - Qualcomm AI Hub - Apple Core ML - ONNX Runtime - WebGPU W3C

PUBLICATIONS

Receipts

Sub-10ms multimodal inference on consumer NPUs
MLSys Workshop - 2025

Federated edge fleets without raw data: a 6-site field study
AXP Internal Whitepaper - 2026

FAQS

What partners actually ask

Q. What hardware do we need?
A. We support Hailo, Apple Neural Engine, Snapdragon, NVIDIA Jetson and WebGPU laptops. Choice is policy, not lock-in.

Q. How is this private?
A. Raw streams never leave the device. Only signed inference receipts and aggregated metrics uplink - by default.

Q. Federated learning safety?
A. Differential privacy budget per round, gradient clipping, server-side aggregation only - no raw data, ever.

Q. What about model drift?
A. Edge eval harness runs nightly; drift triggers re-deploy or shadow mode automatically.

Partner with Ambient Intelligence Fabric

Outcome-priced. Sovereign by default. Refund-backed if the contracted KPI isn't hit.

Apply: alfaxprienz.com/labs/ambient-intelligence-fabric#partner