The accelerator trap

Why specialization is a dead end for the next wave of AI

Learn more on **efficient.computer**

The thesis

Hardware ships behind the workloads it was built for.

ASICs and NPUs are bet on the model architecture of the day. AI moves faster than silicon, so the moment an accelerator ships, it is already chasing a workload it was not designed for. Generality is what keeps a processor relevant across model generations, and it is what unlocks the next wave of AI applications at the edge.

2–3yr
ASIC silicon design cycle from spec to production

6–12mo
Time it takes a new model architecture to redefine the workload

Voices from the field

What the industry is reporting.

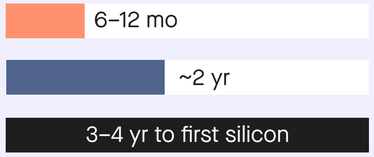
> *We don't want another proprietary SDK. We want an experience that feels like modern software engineering.*
>
> *System architect, industrial automation firm · Embedded World 2025*

> *Hailo, Qualcomm, NVIDIA Jetson each bring incompatible toolchains, creating integration headaches and lock-in.*
>
> *RCR Wireless News · Edge AI accelerators report, 2026*

> *No matter how good the accelerator, customers always need even better software tools.*
>
> *Rehan Hameed, CTO, Deep Vision · via Embedded.com*

Hardware ages in real time

Models reinvent faster than chips ship.



3 to 5 model architectures can emerge during one ASIC design cycle. By the time many AI ASICs reached customers, transformers had already taken over.

Megvii Inc., Arch-Net, arXiv 2111.01135

Right idea, wrong place

Even embedded GPUs miss the edge.

60W
Peak draw of an NVIDIA Jetson AGX Orin, their flagship embedded GPU. Edge AI runs on milliwatts.

NVIDIA Jetson AGX Orin datasheet

~64%
of GPU energy is spent moving data, not computing it.

Intel research, 7nm node

The Fabric architecture uses spatial parallelism to keep data near compute.
No data-center power tax.

The kernel-writing problem

Even AI can't write accelerator code well.

GPU Code
95%+
Frontier LLM correctness writing CUDA

NPU Code
<10%
Frontier LLM correctness writing NPU kernels

Every NPU vendor has its own SDK and instruction set. The Fabric architecture runs **standard C and C++**, so humans and AI assistants stay productive.

Sources: NPUEval (AMD Research), MultiKernel-Bench, 2025

The general-purpose alternative

Three ways to build for AI. Only one keeps up.

Comparing how each architecture handles changing workloads, energy use, and developer effort.

ASIC / NPU — Locked
- × Built for a snapshot of the model
- × Proprietary toolchain, fragmented SDKs
- × Re-tape per model generation

GPU — Parallel
- × Energy-hungry parallelism
- × Low average utilization
- ~ Mature ecosystem, but not built for edge

Fabric architecture — General
- ✓ Spatial dataflow, true energy efficiency
- ✓ Runs existing C and C++
- ✓ One architecture across model generations

What generality unlocks

The next wave of AI is general-purpose at the edge.

- **Real-world workloads** that span signal processing, control, and AI in one chip
- **Decade-long lifecycles** that survive model architecture churn
- **Familiar toolchains** so teams ship faster, with no proprietary kernel rewrites

Compute built for the next edge of possible.