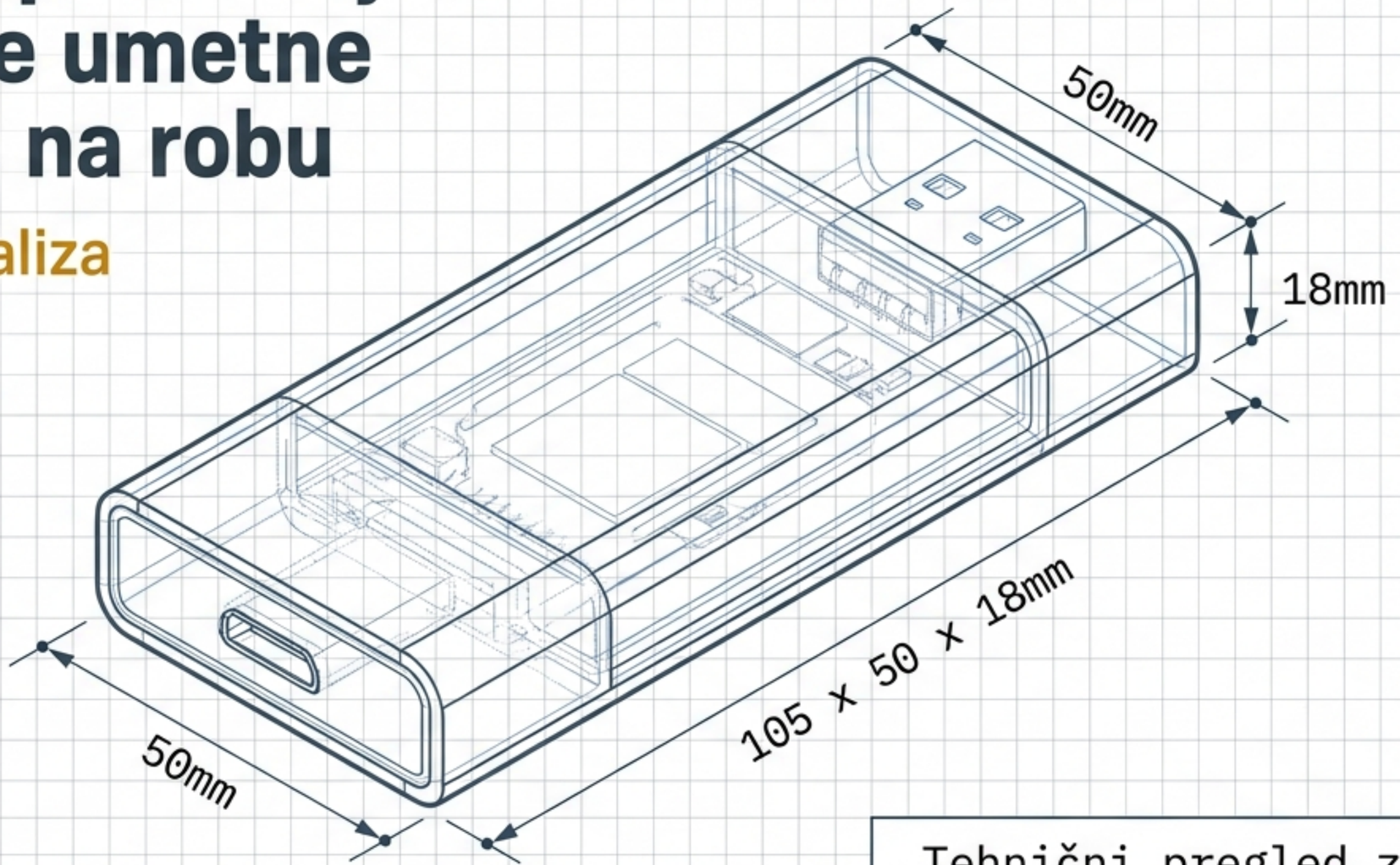
Strojno pospeševanje generativne umetne inteligence na robu

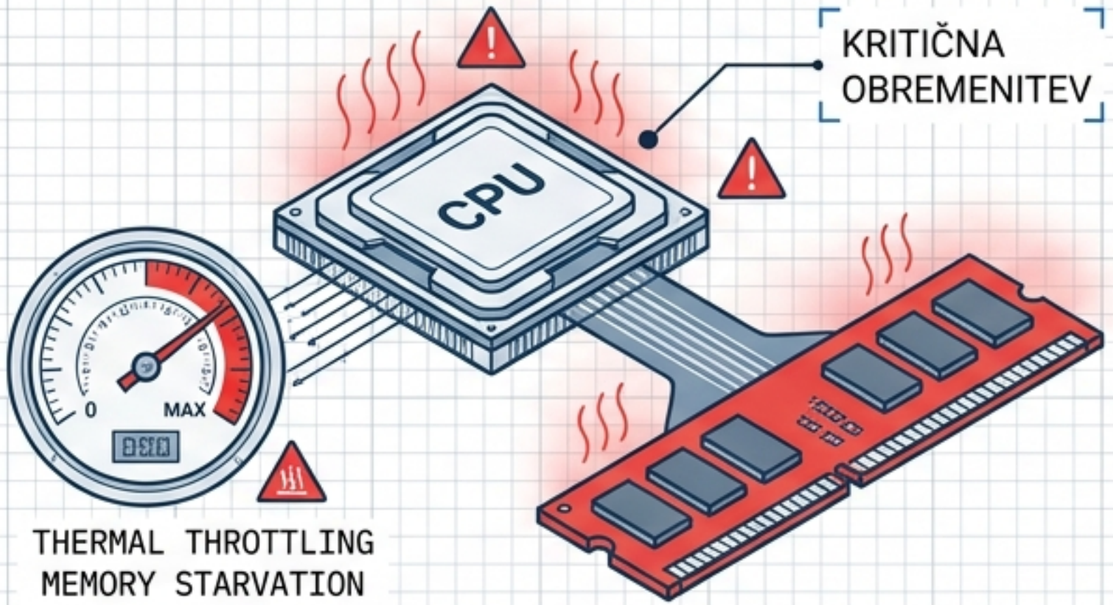
Arhitekturna analiza pospeševalnika ASUS UGen300



Tehnični pregled za inženirje in razvijalce

Problem: Računska in pomnilniška lakota pri lokalnem sklepanju

Standardni sistem



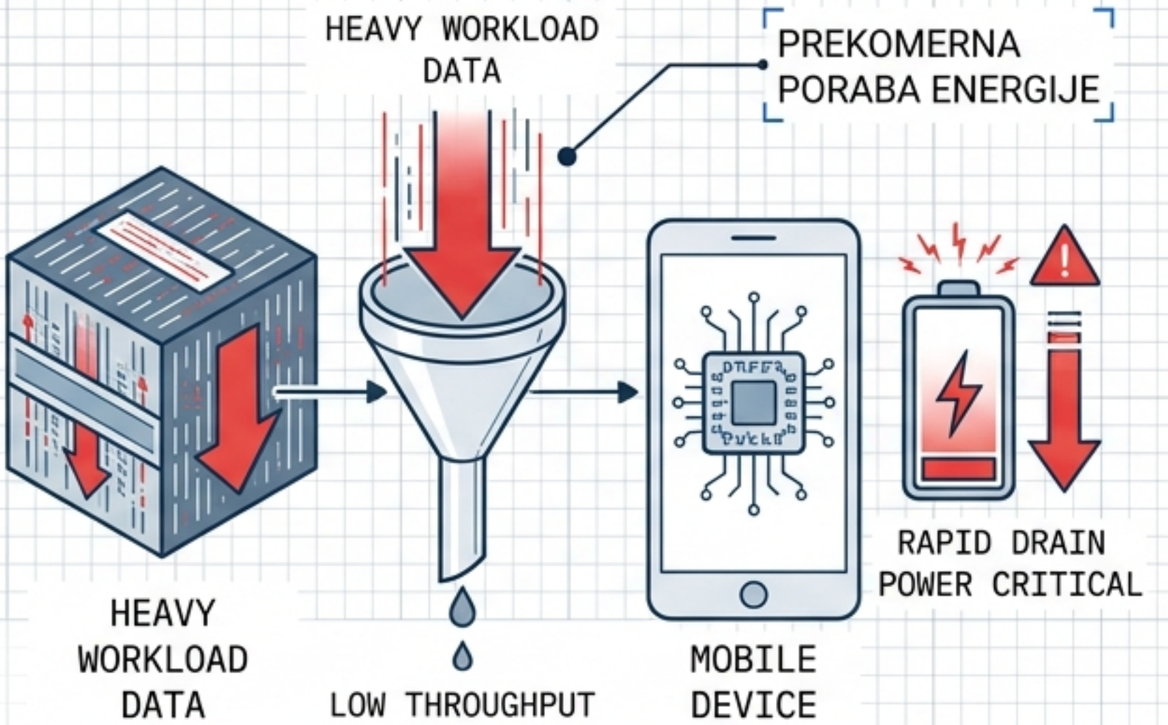
Zasedenost sistemskih virov

Generativni modeli (LLM/VLM) hitro izčrpajo gostiteljev CPE in standardne NPU-je.

Pomnilniško ozko grlo

Skupna uporaba sistemskega RAM-a drastično zmanjša prepustnost podatkov.

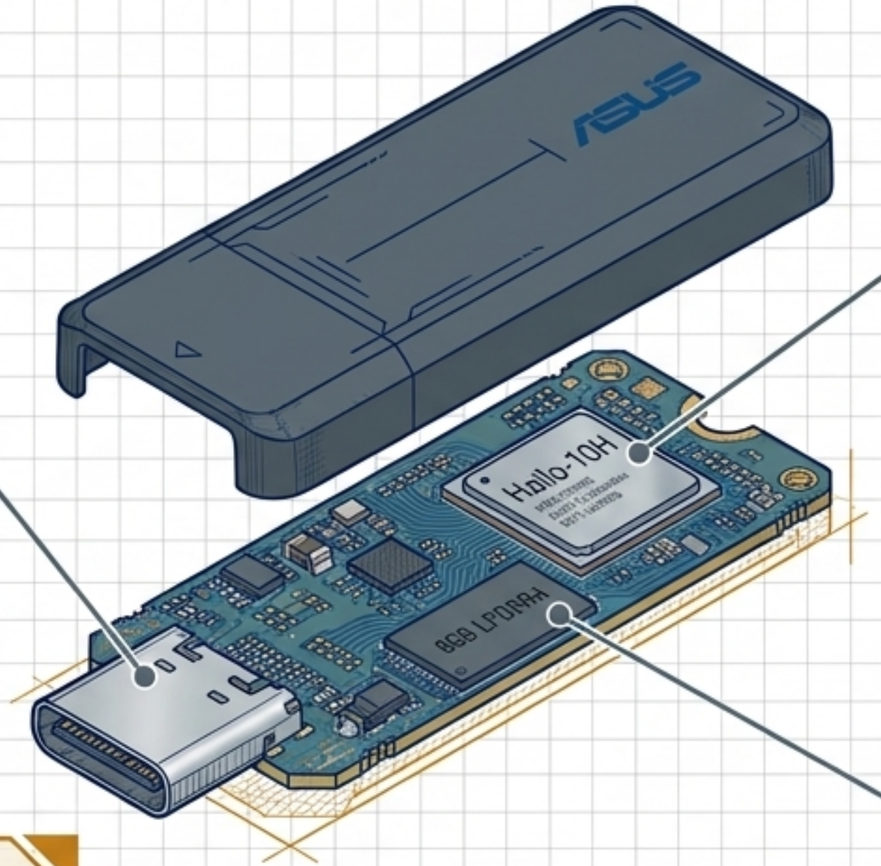
Ozkogrlo na robu



Omejitve pri prenosljivosti

Zahtevni modeli brez namenske strojne opreme povzročajo prekomerno porabo energije na baterijsko napajanih napravah.

Strojna arhitektura ASUS UGen300



Fizični vmesnik:
USB-C® 3.1 Gen 2
(Združljivo s standardom plug-and-play; opcijsko na voljo tudi kot M.2 modul).

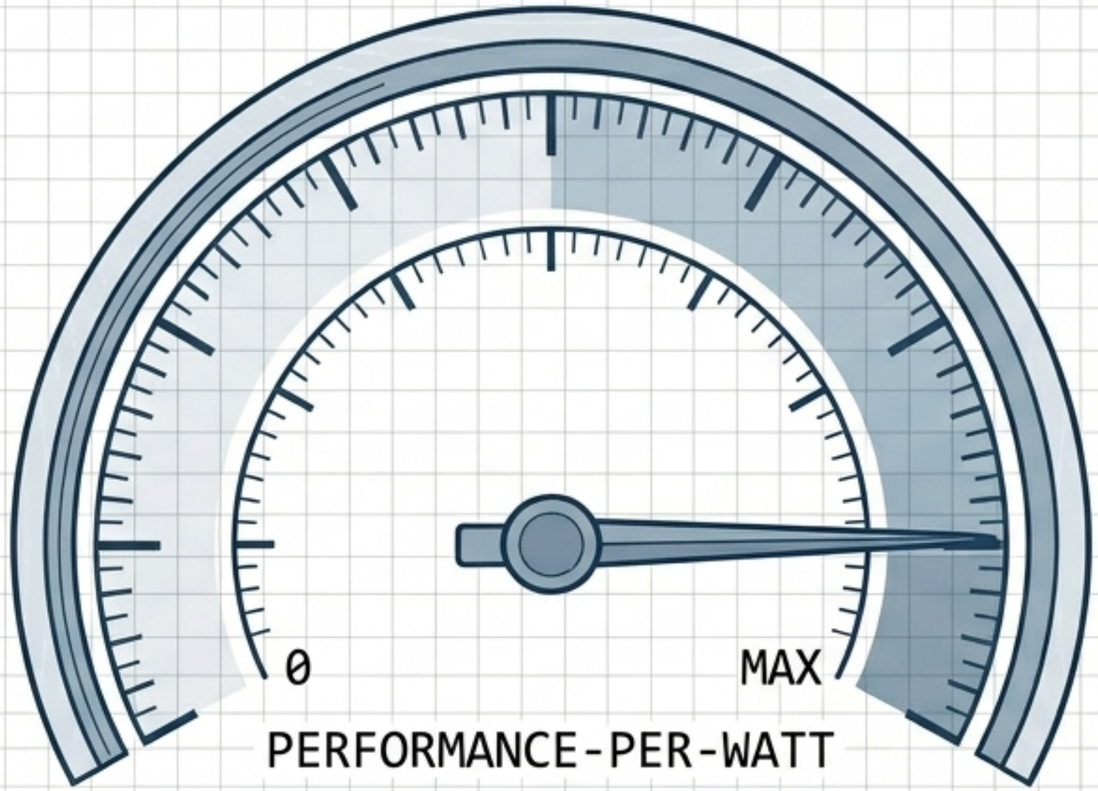
Računska enota:
Procesor Hailo-10H
(Namensko optimiziran za generativne obremenitve).

Pomnilniški podsistem:
8GB LPDDR4
(Integriran, namenski pomnilnik).

Inženirski vpogled

Integracija procesorja in pomnilnika na eni sami tiskanini znotraj formata USB omogoča dodajanje zmogljivosti umetne inteligence sistemom, ki prvotno niso bili zasnovani za strojno učenje.

Kvantifikacija zmogljivosti: Hailo-10H NPU



40 TOPS

Namenska računska moč za umetno inteligenco (Optimizirano za sklepanje pri velikih jezikovnih in vizualnih modelih - LLM in VLM).

2.5W

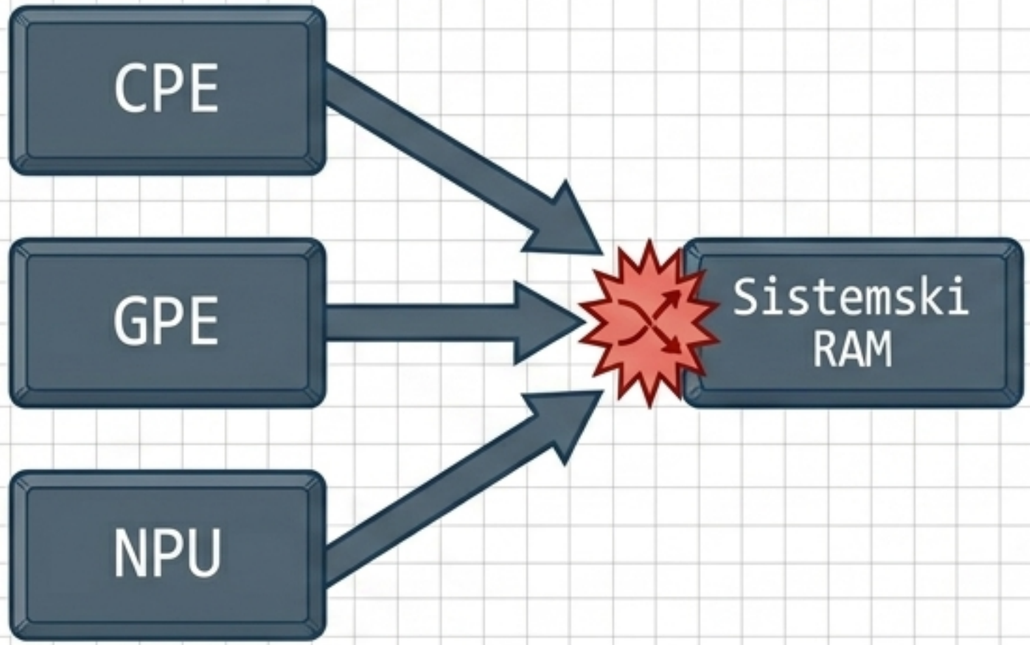
Poraba energije pri tipični obremenitvi.

Inženirski vpogled - Razmerje moč/poraba

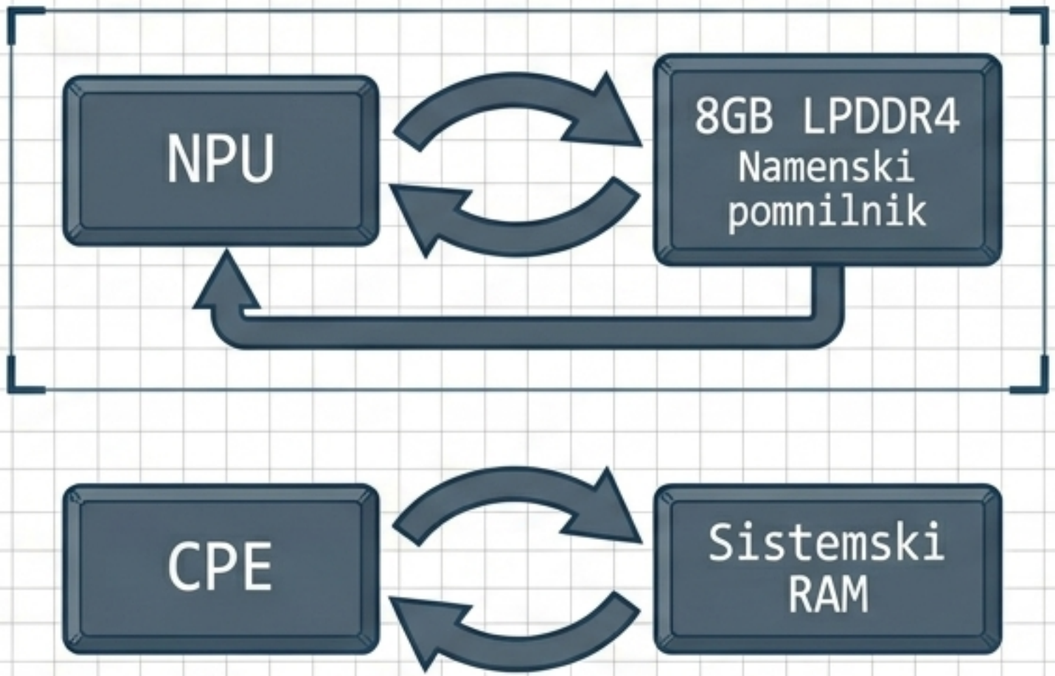
Ovojnica s porabo samo 2.5W je kritična prednost. Omogoča visoko zmogljivo izvajanje nevronskih mrež neposredno na mobilnih, industrijskih ali baterijsko napajanih Edge IoT napravah brez potrebe po masivnem aktivnem hlajenju.

Arhitekturna prednost: Odprava ozkih grl z namenskim pomnilnikom

Konvencionalni NPU



UGen300 Arhitektura



Za razliko od tipičnih NPU-jev, ki si delijo sistemski pomnilnik,

Za razliko od tipičnih NPU-jev, ki si delijo sistemski pomnilnik, namenska arhitektura RAM-a pri UGen300 preprečuje ozka grla in zagotavlja dosledno prepustnost pri kompleksnih cevovodih strojnega učenja. Gostiteljski sistem ostane prost za večopravilnost.

Strojno agnostična povezljivost

Združljivost s potrošniškimi in industrijskimi napravami, idealno za razvijalce, pedagoge in ponudnike vgrajenih ki potrebujejo prenosljivo računsko moč.

USB 3.1 Gen 2 Type-C

Windows

Podpora za gonilnike (na voljo sredi maja 2026).

Linux

Nativna podpora za industrijske in vgrajene sisteme.

Android

Trenutno na voljo za B2B stranke (širša dostopnost v prihodnjih izdajah).

Diagnostična matrika: Robno v primerjavi z oblačnim računanjem

Dimenzija		Oblak (Cloud AI)	Rob (ASUS UGen300)
	Zakasnitev	Visoka/Odvisna od omrežja	Ničelna (Lokalno izvajanje)
	Zasebnost podatkov	Podatki zapustijo napravo	Popolna (Neprekosljiva zanesljivost in varnost na napravi)
	Stroškovni model	Mesečne naročnine	Enkratna investicija v strojno opremo
	Zanesljivost	Zahteva internetno povezavo	100% zmožnost delovanja brez povezave

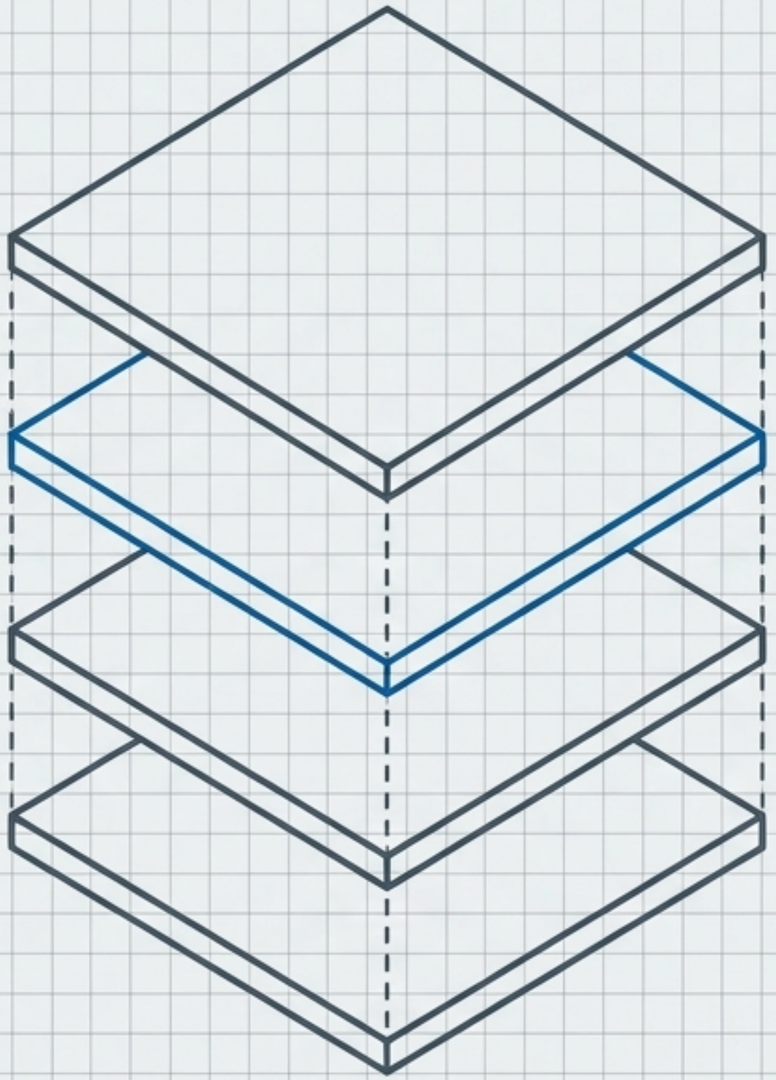
Programski sklad: Od strojne opreme do aplikacije

Inženirski vpogled

Hailo Developer Community ponuja vadnice, referenčne zasnove in deljene vpoglede za hitrejšo uvedbo aplikacij.

Sloj 4 (Razvojna orodja)

UGen Utility (omogoča hitro validacijo).



Sloj 3 (Podprta ogrodja AI)

Nativna podpora za TensorFlow, PyTorch, ONNX (Deluje takoj po priključitvi - out of the box).

Sloj 2 (Operacijski sistem)

Windows / Linux / Android.

Sloj 1 (Strojna oprema)

Hailo-10H + 8GB LPDDR4 prek USB-C.

Ciljne inženirske obremenitve

Generiranje besedila (LLM)

Lokalno izvajanje velikih jezikovnih modelov brez zamikov.

Povzemanje videa

Analiza in stiskanje vizualnih podatkov na robu.

Strojni vid v realnem času

Zaznavanje in klasifikacija objektov pri nizki latenci.

Sprožanje dogodkov

Avtomatizacija lokalnih sistemov na podlagi senzorskih vhodov.

Glas v dejanje (Voice-to-action)

Zasebna in takojšnja obdelava naravnega jezika.

Sinteza: Celoten cevovod uvedbe na robu

1. Izbira modela

Dostop do 100+ predhodno naučenih modelov prek prihajajočega orodja **UGen Utility**.

2. Integracija ogrodja

Uporaba standardov (ONNX, PyTorch, TensorFlow) znotraj želenega OS (Linux/Windows/ Android).

3. Strojno pospeševanje

Takojšen prenos računske obremenitve prek priključka **USB-C** na namensko arhitekturo **UGen300**.

4. Izvajanje na robu

Doseganje **40 TOPS** lokalne zmogljivosti pri porabi samo **2.5W** – sproščanje sistemskih virov gostitelja za optimalno ravnovesje.