

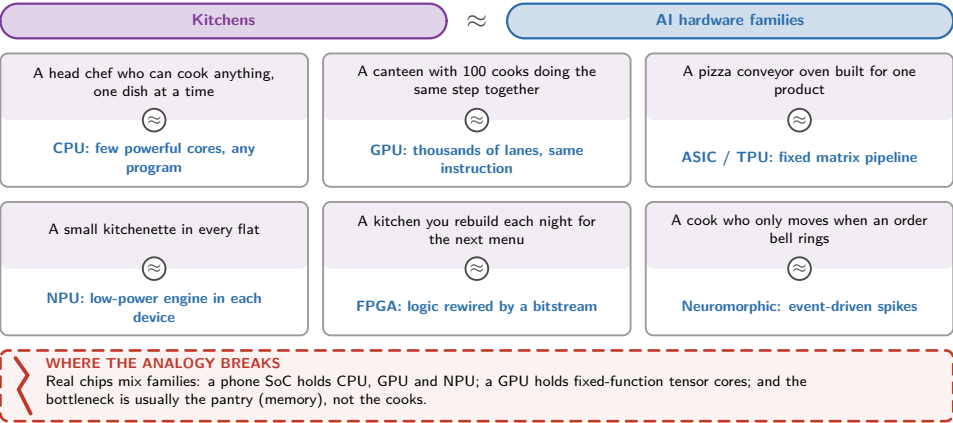


AI Hardware Types Map

AI hardware | CPU, GPU, TPU, NPU, FPGA, ASIC, neuromorphic, analog, photonic and quantum: how each computes, what it is good and bad at, flexibility vs efficiency, derived metrics from published specs, energy per inference

#Hardware #CPU #GPU #NPU #ASIC

Ten families, one job



Kitchens as an analogy for compute families

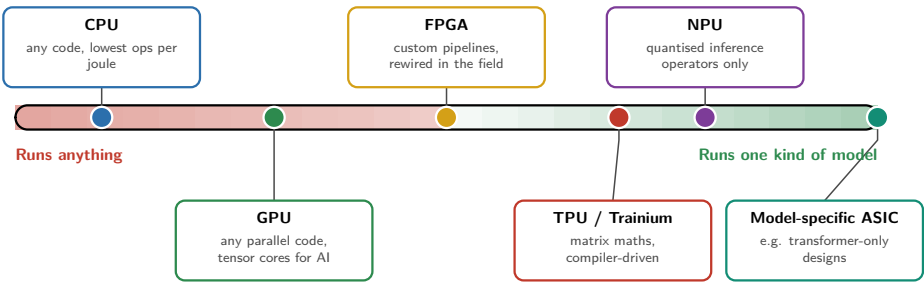
Family	How it computes	Good at	Weak at	Examples
CPU	few wide cores, SIMD vectors, matrix extensions	classic ML, small models, pre/post-processing, low batch	large models: low bandwidth per FLOP	Xeon, EPYC, Graviton, Apple M
GPU	SIMT lanes + tensor cores, HBM or GDDR	training and serving at any scale; software ecosystem	power, price, supply	NVIDIA, AMD Instinct and Radeon
TPU / training ASIC	systolic arrays, HBM, chip-to-chip links	large-scale training and serving in one cloud	portability; buy-only not possible	Google TPU, AWS Trainium
Inference ASIC	fixed dataflow, often large on-chip SRAM	tokens per watt or per second for one model class	new model types, small volumes	Inferentia, Groq LPU, Meta MTIA
NPU	int8/int4 MAC arrays, shared LPDDR	always-on, private, offline inference at 1-10 W	large models; operator coverage	Apple Neural Engine, Qualcomm Hexagon
FPGA	LUTs, DSP slices, block RAM, custom pipelines	microsecond latency, custom precision, low volume	peak throughput per euro, ease of use	AMD Versal, Altera Agilex
Neuromorphic	spiking neurons, event-driven, local memory	sparse sensor streams at mW	standard deep-learning training	Intel Loihi 2, IBM North-Pole (digital)
Analog in-memory	currents through stored conductances	energy per MAC in the-ory	precision, drift, ADC cost	research and early products

Photonic	light interference does matrix maths or carries data	interconnect bandwidth per watt	memory, non-linearities, conversion cost	optical I/O start-ups
Quantum	qubits, interference of amplitudes	no practical ML advantage shown	data loading, error correction	research systems

What each family is good at (state as of 2026)

Flexibility against efficiency

“Where each family sits between a general-purpose and a single-purpose chip”



The more a chip is specialised, the more operations per joule it delivers, and the fewer workloads it accepts

Item	Trains large models	Serves LLMs	Runs on a battery	Logic rewired after manufacture	Buy or rent today	Research stage	Total
CPU	☐	☒	☒	☐	☒	☐	3
GPU	☒	☒	☐	☐	☒	☐	3
TPU / Trainium	☒	☒	☐	☐	☒	☐	3
Inference ASIC	☐	☒	☐	☐	☒	☐	2
NPU	☐	☒	☒	☐	☒	☐	3
FPGA	☐	☐	☐	☒	☒	☐	2
Neuromorphic	☐	☐	☒	☐	☐	☒	2
Analog in-memory	☐	☐	☐	☐	☐	☒	1
Photonic compute	☐	☐	☐	☐	☐	☒	1
Quantum	☐	☐	☐	☐	☐	☒	1
Total	2	5	3	1	6	4	

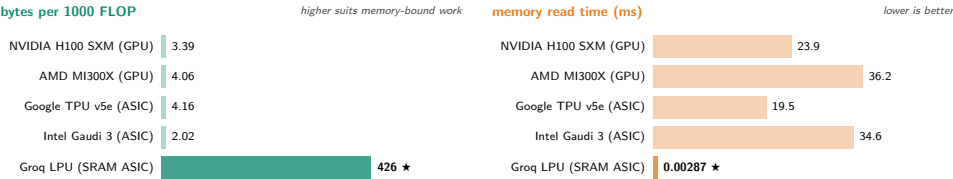
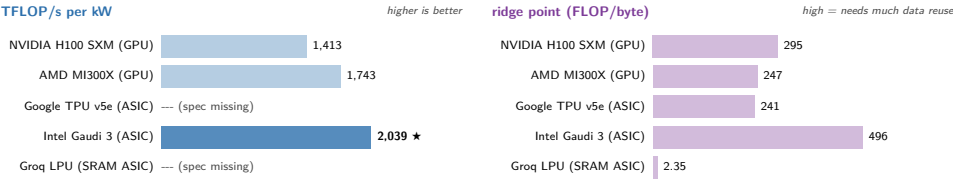
Which family does what in 2026 (a tick means common practice, not a theoretical possibility)

⚠ Reading the comparison

A low ridge point (Groq, about 2 FLOP/byte) means even batch-1 decoding keeps the arithmetic busy, but 230 MB per chip means hundreds of chips per large model. A high ridge point (Gaudi 3, about 500) needs large batches to use its FLOPs. Per-chip numbers ignore interconnect, software maturity and price: they are a first filter, not a verdict (accelerator line-up: compute-infrastructure 02).

Chip (given specs)	TFLOP/s	memory GB	bandwidth TB/s	TDP W
NVIDIA H100 SXM (GPU)	989 bf16	80	3.35	700
AMD MI300X (GPU)	1307 bf16	192	5.3	750
Google TPU v5e (ASIC)	197 bf16	16	0.819	---
Intel Gaudi 3 (ASIC)	1835 bf16	128	3.7	900
Groq LPU (SRAM ASIC)	188 fp16	0.23	80	---

Note: precisions differ (bf16, fp16): peak TFLOP/s are not comparable like for like (fp8 ≈ 2 × bf16 on the same silicon).



TFLOP/s per kW = 1000 × P / TDP

ridge point I* = P / BW

bytes per 1000 FLOP = 1000 × BW / P

memory read time = M / BW

where

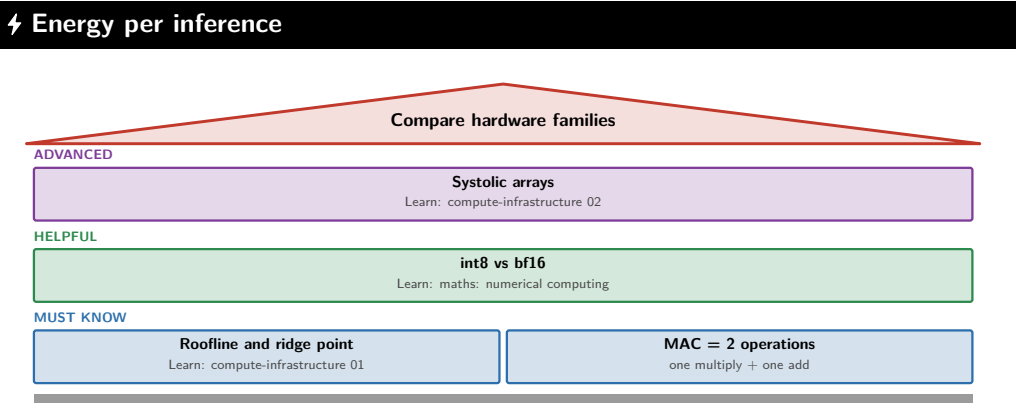
- P peak compute at the stated precision (TFLOP/s = 10¹² FLOP/s)
- BW memory bandwidth (TB/s = 10¹² bytes/s)
- M memory per chip (GB = 10⁹ bytes)

TDP thermal design power (W): the board's power limit

I* FLOP needed per byte read to keep the chip busy (FLOP/byte)

★ best value of the row (none for the ridge point: a balance, not a score)

Five data-centre chips of three families from their published specs (dense bf16 or fp16; Groq LPU: on-chip SRAM, quoted fp16). Derived bars are computed from these numbers only



What this sheet builds on

PROBLEM

Energy to classify one image with a ResNet-50-class network (about 4.1×10^9 multiply-accumulates) on a chip delivering 2 TOPS/W effective vs one delivering 0.2 TOPS/W effective?

GIVEN

- MACs = 4.1×10^9 ; operations = $2 \times$ MACs (multiply + add)
- η = effective efficiency in TOPS/W = 10^{12} operations per joule
- E = operations $\div$ ($\eta \times 10^{12}$), in joules

1

Count operations

$2 \times 4.1e9 = 8.2e9$ ops

Why: TOPS counts a MAC as two operations

2

Efficient chip (NPU-class)

$8.2e9 / (2 \times 1e12) = 4.1e-3$ J

Why: 4.1 mJ per image

3

Less efficient chip (CPU-class)

$8.2e9 / (0.2 \times 1e12) = 4.1e-2$ J

Why: 41 mJ: ten times more

4

Battery view

10 Wh = 36,000 J; $36,000 / 4.1e-3 \approx 8.8$ million images

Why: only the compute; camera, memory and screen add more

ANSWER

About 4 mJ vs 41 mJ per image: a $10\times$ efficiency gap is $10\times$ battery life for the AI part

CHECK

- Effective efficiency is usually far below the peak TOPS/W of the data sheet, because memory traffic and idle units cost energy too

Same model, two efficiency levels. The efficiencies are assumptions for the exercise, not measured figures for a product

MYTH

✗

GPUs are general-purpose, so they are as flexible as CPUs

✗

An ASIC is always better than a GPU

✗

Neuromorphic, analog and photonic chips will replace GPUs soon

FACT

✓

They excel only at massively parallel, regular work

Branch-heavy, serial code (tokenisers, tree search, many classic ML algorithms) often runs better on a CPU.

✓

Only when the workload is stable and the volume large

A chip takes years to design; if the model architecture changes, a fixed dataflow can become a poor fit.

✓

As of 2026 they serve niches or remain research

Photonics is entering as interconnect first; sheet 08 gives the reality check.

Remember

CPU flexible, ASIC efficient

GPU, FPGA, TPU in between

Ridge point = peak $\div$ bandwidth

low suits batch-1 decoding

Energy = ops $\div$ efficiency

MAC = 2 ops

Most phones run NPUs

most training runs on GPUs and TPU-class ASICs

Exotic compute is niche in 2026

check claims against shipped products

Q1

Why does a GPU beat a CPU at matrix multiplication?

A Thousands of lanes and tensor cores run the same instruction on many data at once, fed by much higher memory bandwidth.

Q2

What is a systolic array?

A A grid of MAC cells passing operands to neighbours in lockstep, reusing each fetched value many times (TPU core).

Q3

Ridge point of a chip with 989 TFLOP/s and 3.35 TB/s?

A About 295 FLOP/byte.

Q4

Why does a Groq LPU need many chips per large model?

A About 230 MB of on-chip SRAM per chip and no HBM.

Q5

Energy for 8.2e9 operations at 2 TOPS/W?

A $8.2e9 \div 2e12 = 4.1$ mJ.

Q6

When is an FPGA preferable to a GPU?

A Microsecond deterministic latency, custom precision or I/O, small volumes, field updates of the hardware logic.

Q7

Which family dominates on-device inference?

A NPUs inside phone and laptop system-on-chips.

Q8

Has quantum computing shown a practical ML advantage by 2026?

A No; it remains research (sheet 08).

Tadata | tadata.fr | contact@tadata.codes