The Next Frontier in AI Hardware

PROCESSOR	ARCHITECTURE	LATENCY	THROUGHPUT	POWER	MATURITY	USAGE
AWS Inferentia2	Custom ASIC					Cloud inference (LLM, vision)
Blaize	Edge AI SoC					Edge inference (vision, robotics)
BrainChip Akida	Neuromorphic SoC					Ultra-low-power, event-based AI
Cerebras WSE-3	Wafer-scale ASIC (SoW)					Large-scale training, inference
Google TPU v4	Matrix ASIC					LLM training, inference
Graphcore (IPU)	Fine-grained AI accelerator ASIC					High control inference, training
Groq	Tensor streaming ASIC					Ultra-low latency LLM inference
HP ReRam	In-memory AI compute					Experimental memory-centric AI
IBM Analog Compute	Analog AI ASIC					Low-power analog neural inference
Intel Loihi 2	Spiking neuromorphic ASIC					Spiking sensor for edge AI
NVIDIA A100	GPU (Ampere)					General-purpose AI compute
NVIDIA H100	GPU (Hopper)					Transformer-optimized LLM inference
Tachyum Prodigy	Unified CPU/GPU/ NPU Hybrid					Hyperscale AI & HPC (claimed)
Tesla Dojo D1	SoC ASIC					Video processing for Full Self Driving and AI
Quantum Processors	Qubits (various technologies)					Quantum research (non-AI)

Disclaimer: The data in this table is for relative informational purposes only. All performance data are based on publicly available sources and may vary across use cases and configurations.

KEY

Latency

1 = Very High Latency

3 = Moderate Latency

5 = Ultra-low Latency

Throughput

1 = Low throughput

3 = Moderate

5 = Very High

Maturity

1 = Lab/Internal Only

3 = Limited Commercially

5 = Fully Commercial

Power

1 = Ultra Low Power

2 = Low-power Edge/Neuromorphic

3 = Moderate Inference Chips

4 = Data Center Accelerators

5 = Wafer-scale/GPU-class