The Top 50 Global AI Chips by Speed

By : Khawar Nehal

khawar@atrc.net.pk

Applied Technology Research Center

http://atrc.net.pk

Date : 9 June 2026

Here are the top 50 distinct AI accelerators globally, ranked by their peak 8-bit (FP8/INT8) AI speed.

Rank	Chip Name	Capabilities	Cost (g Ag)	AI Speed (TFLOPS/TOPS)	g Ag per Speed	W per Speed
1	Cerebras WSE-3	Wafer-scale, 4TB SRAM, 125 PFLOPS system	1,500,000	125,000	12.00	0.18
2	NVIDIA B200	Blackwell, 192GB HBM3e, FP4/FP8	40,000	4,500	8.89	0.22
3	NVIDIA B100	Blackwell, 192GB HBM3e, lower bin	35,000	3,500	10.00	0.20
4	Intel Max 1550 (PVC)	Xe-HPC, 128GB HBM2e, 5nm	10,000	3,200	3.12	0.19
5	AMD MI300X	CDNA 3, 192GB HBM3, chiplet	15,000	2,614	5.74	0.29
6	AMD MI325X	CDNA 3, 256GB HBM3	18,000	2,614	6.89	0.29
7	AMD MI300A	CDNA 3 + Zen 4, APU design	12,000	2,614	4.59	0.29
8	AWS Trainium 2	Custom training, 4x Trn1 speed	15,000	2,300	6.52	0.22

Rank	Chip Name	Capabilities	Cost (g Ag)	AI Speed (TFLOPS/TOPS)	g Ag per Speed	W per Speed
9	Cambricon MLU370-X8 (Card)	7nm, 8-chip PCIe card, 48GB HBM/chip	20,000	2,048	9.77	0.24
10	NVIDIA Thor	Automotive, unified architecture, FP8	2,000	2,000	1.00	0.50
11	NVIDIA H200	Hopper, 141GB HBM3e	35,000	1,979	17.69	0.35
12	NVIDIA H100 SXM	Hopper, 80GB HBM3, NVLink	30,000	1,979	15.16	0.35
13	Intel Gaudi 3	128GB HBM2e, 1.8TB/s BW	15,000	1,835	8.17	0.49
14	NVIDIA H100 PCIe	Hopper, 80GB HBM3, PCIe Gen5	20,000	1,500	13.33	0.47
15	NVIDIA RTX 4090	Ada Lovelace, 24GB GDDR6X, Consumer	1,600	1,300	1.23	0.35
16	AMD MI250X	CDNA 2, 128GB HBM2e	10,000	1,250	8.00	0.40
17	AMD MI250	CDNA 2, 128GB HBM2e, dual die	8,000	1,250	6.40	0.40
18	Graphcore Bow IPU	IPU-Mk2, 900MB SRAM, wafer-level	5,000	1,200	4.17	0.33
19	Biren BR100	7nm class chiplet,	15,000	1,024	14.65	0.54

Rank	Chip Name	Capabilities	Cost (g Ag)	AI Speed (TFLOPS/TOPS)	g Ag per Speed	W per Speed
		64GB HBM				
20	Groq LPU v2	Deterministic SRAM, no HBM, low latency	20,000	1,000	20.00	0.30
21	Untether AI Speed AI100	Near-memory compute, 1000 TOPS	5,000	1,000	5.00	0.10
22	d-Matrix Corsair	Photonic interconnect, in-memory compute	10,000	1,000	10.00	0.20
23	SambaNova SN40L	Reconfigurable Dataflow Unit (RDU)	15,000	950	15.79	0.32
24	Google TPU v5p	Liquid cooled, massive scale training	20,000	918	21.79	0.49
25	NVIDIA L40S	Ada, 48GB GDDR6, inference optimized	10,000	916	10.92	0.38
26	NVIDIA L40	Ada, 48GB GDDR6, graphics/inference	9,000	900	10.00	0.33
27	Intel Gaudi 2	96GB HBM2e, 2.4TB/s BW	8,000	900	8.89	0.67
28	Huawei Ascend 910C	7nm class, 64GB HBM, H100 alt	25,000	800	31.25	0.88
29	NVIDIA RTX 4080	Ada, 16GB GDDR6X, Consumer	1,200	700	1.71	0.46
30	Horizon Journey 6P	Automotive, 560 TOPS, transformer	500	560	0.89	0.50

Rank	Chip Name	Capabilities	Cost (g Ag)	AI Speed (TFLOPS/TOPS)	g Ag per Speed	W per Speed
		support				
31	Google TPU v4	32GB HBM2e, liquid cooled pod	12,000	550	21.82	0.76
32	Horizon Journey 5	Automotive, 512 TOPS, 7nm	400	512	0.78	0.59
33	Biren BR104	Single die variant of BR100	8,000	512	15.62	0.59
34	NVIDIA RTX 4070 Ti	Ada, 12GB GDDR6X, Consumer	800	500	1.60	0.56
35	AMD MI210	CDNA 2, 64GB HBM2e	5,000	450	11.11	0.67
36	AMD MI100	CDNA 1, 32GB HBM2	4,000	450	8.89	0.67
37	Google TPU v5e	Air cooled, inference/training balance	5,000	400	12.50	0.50
38	Qualcomm Cloud AI 100 Ultra	High-end edge/cloud inference	3,000	400	7.50	0.19
39	AWS Trainium 1	Custom training, 16nm	5,000	400	12.50	1.00
40	Google TPU v3	16GB HBM, liquid cooled pod	8,000	400	20.00	1.05
41	Huawei Ascend 910B	7nm class, 64GB HBM	15,000	320	46.88	1.25
42	NVIDIA A100 80GB	Ampere, 80GB HBM2e	15,000	312	48.08	1.28
43	Tesla FSD HW4	Automotive, 16nm, vision-only	1,000	300	3.33	0.23

Rank	Chip Name	Capabilities	Cost (g Ag)	AI Speed (TFLOPS/TOPS)	g Ag per Speed	W per Speed
44	Apple M3 Ultra NPU	Neural Engine, 32-core, unified memory	500	300	1.67	0.20
45	Huawei Ascend 910A	7nm, 32GB HBM2	10,000	256	39.06	1.17
46	NVIDIA Orin	Automotive, 16GB LPDDR5, 7nm	500	254	1.97	0.59
47	Qualcomm Cloud AI 100	Standard edge/cloud inference	1,500	200	7.50	0.35
48	Mobileye EyeQ Ultra	Automotive, 5nm, ADAS	300	186	1.61	0.70
49	Google TPU v2	8GB HBM, pod training	5,000	180	27.78	1.67
50	Tenstorrent Wormhole	RISC-V + Tensix cores, 144MB SRAM	2,000	150	13.33	0.80

Key Insights from the "Silver Standard" Data

1. **The Consumer GPU Anomaly (The Lowest Silver Cost):**
    - The **NVIDIA RTX 4090** is the most "silver-efficient" high-end chip on the list at just **1.23 grams of silver per TFLOPS**. Because it is mass-produced for the gaming market, its cost-to-performance ratio is vastly superior to enterprise chips like the H100 (15.16 g/TFLOPS). If you can fit a workload into 24GB of VRAM, consumer hardware offers a 12x better "silver return" on investment than datacenter hardware.

2. **The "Sanctions Tax" on Chinese Silicon:**
    - Chips like the **Huawei Ascend 910B (46.88 g/TFLOPS)** and **910A (39.06 g/TFLOPS)** require nearly 3x to 4x more silver (purchasing power) per TFLOPS than an NVIDIA H100. This metric perfectly quantifies the "sanctions tax"—the massive inefficiencies, lower yields, and older manufacturing processes (DUV instead of EUV) required to build advanced chips in China today.

3. **The Scale Advantage of Cerebras:**
    - Despite costing a staggering 1.5 million grams of silver (1.5 million USD), the **Cerebras WSE-3** achieves 125,000 TFLOPS. This brings its silver cost per speed down to **12.00 g/TFLOPS**. This makes it cheaper per compute than the H100, H200, and even the B200, proving that extreme physical scale can overcome high absolute costs.

4. **Automotive & Edge Efficiency:**
    - Automotive chips like the **Horizon Journey 5 (0.78 g/TFLOPS)** and **Journey 6P (0.89 g/TFLOPS)** are incredibly cheap in terms of silver cost per compute. They are highly optimized, low-margin mass-market silicon designed to run specific neural networks efficiently, rather than general-purpose training.