UNTETHER AI

Untether AI sets records in MLPerf® 4.1 Benchmarks

Accelerating the Future of AI
Industry-leading performance and energy efficiency
in Image Classification benchmark - ResNet-50 v1.5

Datacenter Closed

Highest performance PCIe submission

Performance

speedAI 240 Preview 150W TDP	70,097
H100-NVL 400W TDP	62,519
speedAI 240 Slim 75W TDP	51,625

Normalized Single Card Server Queries/s *(higher is better)*

>4X improvement in energy efficiency

Energy Efficiency

speedAI 240 Preview	467
H100-NVL	156
speedAI 240 Slim	688

Normalized Single Card Server Queries/s/W *(higher is better)*

Untether AI and competitor performance information based on MLPerf 4.1 ResNet-50 verified best performance for PCIe Cards, 4.1-0076, 4.1-0040, 4.1-0066. Normalized to single PCIe-card and TDP is not an official MLPerf benchmark.

Datacenter Closed - Power

20% of power │ 1/4 of size

Performance & Power

H200 SXM 8U	480,131 Queries/s
	5,014 Watts
speedAI 240 Slim 2U	309,752 Queries/s
	986 Watts

>3X improvement in energy efficiency

Energy Efficiency

96 Queries/s/W   314 Queries/s/W

Server Queries/s/W *(higher is better)*

Untether AI and competitor performance and power information based on MLPerf 4.1 verified best performance, 4.1-0066, 4.1-0049 Server queries/s/W is not an official MLPerf benchmark.

Edge Closed - Latency

Best in class latency

Latency (ms)

Single Stream

speedAI 240 Slim 75W DTP	0.12
L40S 350W TDP	0.33

Multi Stream

speedAI 240 Slim 75W DTP	0.17
L40S 350W TDP	0.53

Latency (ms) *(lower is better)*

Untether AI and competitor performance information based on MLPerf 4.1 Edge verified best performance ResNet-50, 4.1-0069, 4.1-0041.

Edge Closed - Performance

Highest performance submission

Performance

speedAI 240 Preview 150W TDP	70,348
L40S 350W TDP	43,152
speedAI 240 Slim 75W TDP	56,277

Normalized Single Card Offline Samples/S *(higher is better)*

6X improvement in energy efficiency

Energy Efficiency

speedAI 240 Preview	469
L40S	123
speedAI 240 Slim	750

Normalized Single Card Offline Samples/s/W *(higher is better)*

Untether AI and competitor performance information based on MLPerf 4.1 Edge verified best performance ResNet-50, 4.1-0077, 4.1-0041, 4.1-0067. Normalized to single PCIe-card and TDP is not an official MLPerf benchmark.
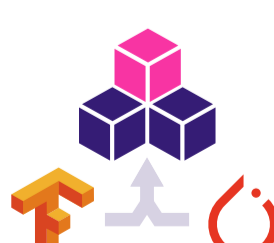
---

speedAI Product Family

speedAI Device
At-Memory architecture realized

speedAI Cards
Industry leading performance

imAIgine
software development kit (SDK)

Explore Untether AI products and how we can
accelerate your projects, products, and services

Learn More