

FuriosaAI WARBOY

High performance inference chip for the most advanced vision applications



PRODUCT SPECIFICATIONS

Peak Performance	64 TOPS
On-chip SRAM	32 MB
Host Interface	PCIe Gen4 8-lane
Form Factor	Full-Height Half-Length
	Half-Height Half-Length
Thermal Solution	Passive and Active Fan

TDP	40 - 60W (Configurable)
Operating Temperature	0 - 50 °C
Clock Speed	2.0 GHz
DDR Speed	4266 Mbps
Memory Type	LPDDR4X
Memory Size	16 GB (max. 32 GB)
Peak Memory Bandwidth	66 GB/s

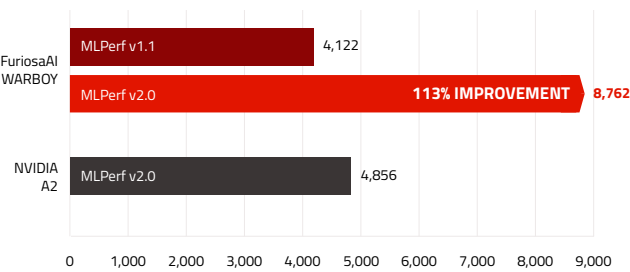
MLPerf™ BENCHMARK RESULTS

FuriosaAI's first silicon chip, WARBOY, marked its successful debut at MLPerf™ 2021. This was only a few weeks after achieving a one-shot silicon success, and having had limited time for optimizations.

In 2022, FuriosaAI remained the only startup to submit results in MLPerf™ Inference, and outperformed NVIDIA A2 in vision area. This time, through purely enhancements in the compiler, our team was able to double the performance on the exact same silicon, WARBOY.

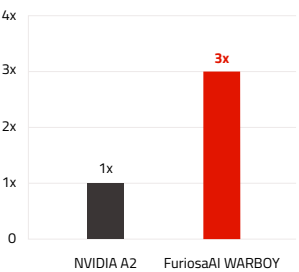
The superior performance across diverse models at MLPerf™ benchmarks, which are carefully designed by the committee, requires single-digit microsecond-level of performance engineering. Not even a single glitch from any of the numerous components comprising the AI chip and system is permitted.

Offline Throughput
Object Detection (SSD-MobileNetV1)



Official MLPerf™ score of v2.0 Inference Edge: Closed. Retrieved from <https://mlcommons.org/en/inference-edge-20/> 12 May 2022, entry 2.0-142. The MLPerf™ name and logo are trademarks of MLCommons Association in the United States and other countries. All rights reserved. Unauthorized use strictly prohibited. See www.mlcommons.org for more information.

Performance / Price
Object Detection (SSD-MobileNetV1)



EFFICIENTNET BECOMES TRULY EFFICIENT

Neural network models are becoming more compact and efficient by adopting operators such as depthwise separable and group convolutions. However, on current GPUs, improvements in models cannot be translated into true efficiency due to architectural limitations.

WARBOY natively exposes algorithms' efficiency improvements, hence resulting in much higher chip utilization and performance compared to existing solutions.

EfficientNetV2-S
Throughput / Watt

