Myrtle.ai

VOLLO
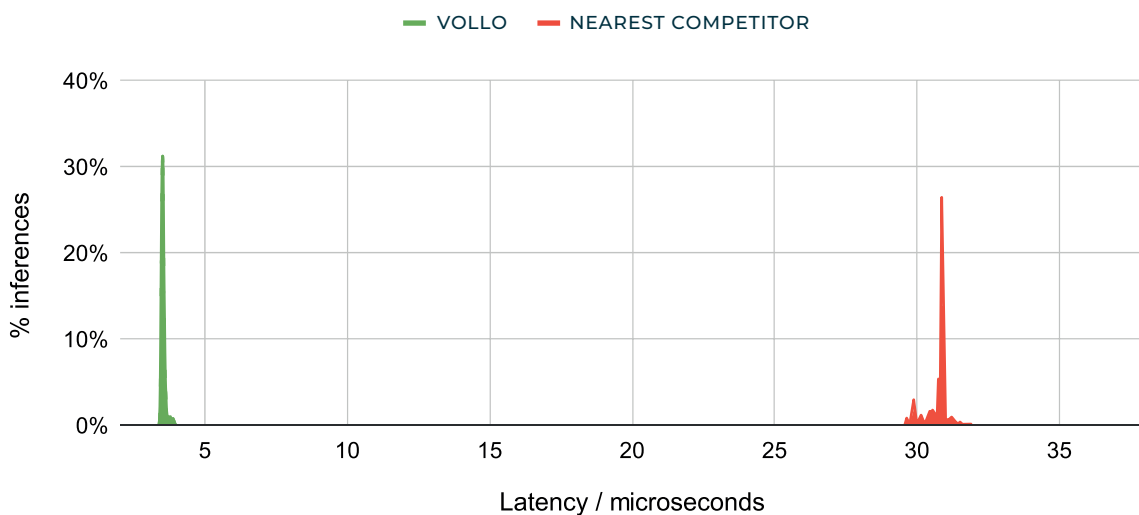
LOWEST LATENCY ML INFERENCE FOR TRADING DECISIONS

RUN YOUR AI MODELS IN UNDER 10µs

- Unrivalled STAC-ML latency[1]
- Trusted by leading firms
- Simple to evaluate and deploy
- High throughput

Evaluate your future performance today using the tools at vollo.myrtle.ai

LATENCY COMPARISON WITH NEAREST COMPETITOR[2]

— VOLLO   — NEAREST COMPETITOR

Development environment:

- Work within your existing framework
- Train in PyTorch or TensorFlow, export in ONNX
- Change models in seconds

PyTorch

TensorFlow

Quickly check how fast your models would run on Vollo
Visit vollo.myrtle.ai | No new hardware required
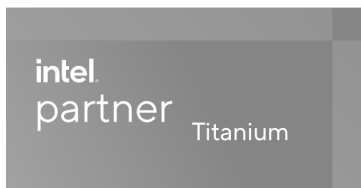
Deployment options:

- Standard FHFL & HHHL PCIe cards
- FPGA netlist for integration into existing systems

STAC-ML™ AUDITED PERFORMANCE

	SUMACO TEST SUITE	TACANA TEST SUITE
99th Percentile Latencies	• 24.1 µs for LSTM_A (smallest model tested) • 64.8 µsec for LSTM_B • 1.35 ms for LSTM_C (largest model tested) STAC-ML.Markets.Inf.S.LSTM_A.[1, 2, 3, 4].LAT.v1 STAC-ML.Markets.Inf.S.LSTM_B.[1, 2, 3, 4].LAT.v1 STAC-ML.Markets.Inf.S.LSTM_C.[1, 2, 3, 4].LAT.v1	• 5.07 µs for LSTM_A (smallest mode tested) • 6.77 µs for LSTM_B • 31.0 µs for LSTM_C (largest mode tested) STAC-ML.Markets.Inf.T.LSTM_A.[1, 2].LAT.v1 STAC-ML.Markets.Inf.T.LSTM_B.2.LAT.v1 STAC-ML.Markets.Inf.T.LSTM_C..LAT.v1
Throughput	Throughput exceeded 650K inf/sec for LSTM_A with 48 NMI STAC-ML.Markets.Inf.S.LSTM_A.48.TPUT.v1	Throughput exceeded 14 M inf/sec for LSTM_A with 24 NMI STAC-ML.Markets.Inf.T.LSTM_A.24.TPUT.v1
Space Efficiency	Space efficiency exceeded 646K inf/sec/cubic foot for LSTM_A with 48 NMI STAC-ML.Markets.Inf.S.LSTM_A.48.SPACE_EFF.v1	Space efficiency exceeded 1.4M inf/sec/cubic foot for LSTM_A with 24 NMI STAC-ML.Markets.Inf.T.LSTM_A.24.SPACE_EFF.v1
Energy Efficiency	Energy efficiency exceeded 1.18M inf/sec/kW for LSTM_A with 48 NMI STAC-ML.Markets.Inf.S.LSTM_A.48.ENERG_EFF.v1	Energy efficiency exceeded 2.32M inf/sec/kW for LSTM_A with 24 NMI STAC-ML.Markets.Inf.T.LSTM_A.24.ENERG_EFF.v1

[1] Audited by the STAC Research Council, the finance industry's own benchmarking organization
[2] Latency model testing by Myrtle.ai on proprietary models

intel. partner Titanium

Full product information at
myrtle.ai/vollo

Myrtle.ai