intel partner Titanium

Myrtle.ai

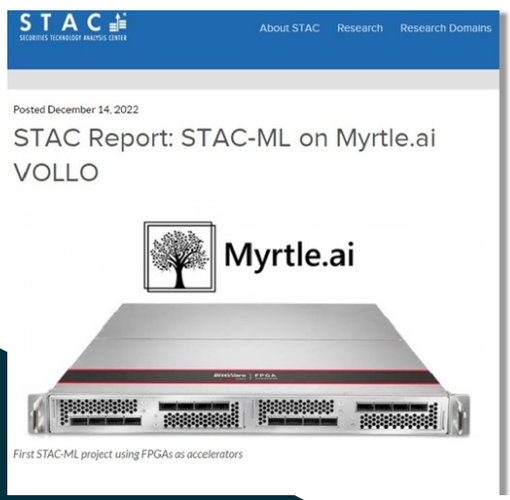
YOUR BEST MACHINE LEARNING RUNNING WITHIN THE TIGHTEST LATENCY BOUNDS

- Achieve the lowest latency inference
- Rely on deterministic short tail latency
- Run large models within strict bounds

BENCHMARKED USE CASES

- Risk Analysis
- Anomaly Detection
- Portfolio Optimization
- Market Predictions

- Evaluate today vollo@myrtle.ai
- Generate reports on the performance, accuracy, power and efficiency



- Unrivaled machine learning inference latencies
- 99th percentile latencies as low as 5.07µs [1]
- Full audit reports at stacresearch.com

[1] Official 99th percentile latency in the STAC-ML™ Tacana test suite STAC-ML.Markets.Inf.T.LSTM_A.[1,2].LAT.v1

intel AGILEX

Evaluation program vollo@myrtle.ai

VOLLO

Smart Inference on Market Data

Smarter

Your best machine learning models running in the latency budget you have

Trusted

Intel systems that are space efficient for co-located deployments

Straightforward

Works seamlessly with the PyTorch deep learning framework

Faster

Unsurpassed 99th percentile latencies

STAC-ML™ Audited Performance

System Under Test

- VOLLO SDK 0.2
- VOLLO Accelerator 0.2
- Ubuntu 20.04.1 LTS
- BittWare TeraBox™ 1402B
- 4 x BittWare IA-840f-0001, each with one Intel® Agilex™ AGF027 FPGA and 4 x 16 GB DDR4 @ 2666 MHz – 64GB total
- One Intel® Xeon® Platinum 8351N Processor @ 2.40 GHz
- 4 x 8 GB DDR4 (Micron 2933 MHz) – 32 GB Total
- Rust 1.16.0

	Sumaco Test Suite	Tacana Test Suite
99th Percentile Latencies	• 24.1 µs for LSTM_A (smallest model tested) • 64.8 µs 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 model tested) • 6.77 µs for LSTM_B • 31.0 µs for LSTM_C (largest model 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.1.LAT.v1
Throughput	• Throughput exceeded 650 K inf/sec for LSTM_A with 48 NMI STAC-ML.Markets.Inf.S.LSTM_A.48.TPUT.v1	• Throughput exceeded 1.4 M inf/sec for LSTM_A with 24 NMI STAC-ML.Markets.Inf.T.LSTM_A.24.TPUT.v1
Space Efficiency	• Space efficiency exceeded 646 K inf/sec/cubic foot for LSTM_A with 48 NMI STAC-ML.Markets.Inf.S.LSTM_A.48.SPACE_EFF.v1	• Space efficiency exceeded 1.4 M 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 1.18 M inf/sec/kW for LSTM_A with 48 NMI STAC-ML.Markets.Inf.S.LSTM_A.48.ENERG_EFF.v1	• Energy efficiency exceeded 2.32 M inf/sec/kW for LSTM_A with 24 NMI STAC-ML.Markets.Inf.T.LSTM_A.24.ENERG_EFF.v1

VOLLO Software & SDK

- Optimized C API for low latency streaming applications
- API runs on host server & PCIe card
- Supports Ubuntu 20.04.1 LTS or later versions
- VOLLO tool suite imports pre-trained models from PyTorch or TensorFlow
- LSTM model library for a range of solutions using PyTorch
- Pre-built bitstreams for Intel Agilex FPGA. No FPGA knowledge required
- Supports model architectures up to 67 M parameters
- Up to 12 models can be loaded into one PCIe card

Full product information at myrtle.ai/fintech

Myrtle.ai