

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](mailto:vollo@myrtle.ai)
- Generate reports on the performance, accuracy, power and efficiency



- Unrivaled machine learning inference latencies
- 99<sup>th</sup> percentile latencies as low as 5.07 $\mu$ s [1]
- Full audit reports at [stacresearch.com](https://stacresearch.com)

[1] Official 99<sup>th</sup> percentile latency in the STAC-ML™ Tacana test suite  
STAC-ML.Markets.Inf.T.LSTM\_A.[1,2].LAT.v1

Evaluation program [vollo@myrtle.ai](mailto:vollo@myrtle.ai)

intel  
AGILEX™

# 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

#### 99th Percentile Latencies

- 24.1  $\mu$ s for LSTM\_A (smallest model tested)
- 64.8  $\mu$ s 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

#### Throughput

- Throughput exceeded 650 K inf/sec for LSTM\_A with 48 NMI

STAC-ML.Markets.Inf.S.LSTM\_A.48.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

#### 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

### Tacana Test Suite

- 5.07  $\mu$ s for LSTM\_A (smallest model tested)
- 6.77  $\mu$ s for LSTM\_B
- 31.0  $\mu$ s 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 exceeded 1.4 M inf/sec for LSTM\_A with 24 NMI

STAC-ML.Markets.Inf.T.LSTM\_A.24.TPUT.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 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



## Myrtle.ai

Full product  
information at  
[myrtle.ai/fintech](https://myrtle.ai/fintech)

"STAC" and all STAC™ related names are trademarks or registered trademarks of the Securities Technology Analysis Center, LLC. Only audited results with a STAC-ML™ code starting STAC-ML.Markets.Inf. are official. Full independent STAC-ML™ audit results are only available to STAC™ members. Other companies STAC™ audit results may only be available to premium STAC™ finance subscribers. Comparisons in this document that are not between STAC-ML.Markets.Inf coded data will vary from system to system. All purchasing decisions should only be made after direct assessment of the product within the official VOLLO Evaluation Program. VOLLO™, VOLLO Accelerator and the VOLLO™ logo are trademarks of Myrtle.ai. PyTorch, the PyTorch logo and any related marks are trademarks of The Linux Foundation. Intel, the Intel logo, and other Intel marks are trademarks of Intel Corporation or its subsidiaries. Other names and brands may be claimed as the property of others. Copyright © Myrtle.ai 2023. Rev 2023.4.27