Achieving Low-Latency Market Prediction on Agilex™ 7 FPGAs F-Series

altera™
An Intel Company

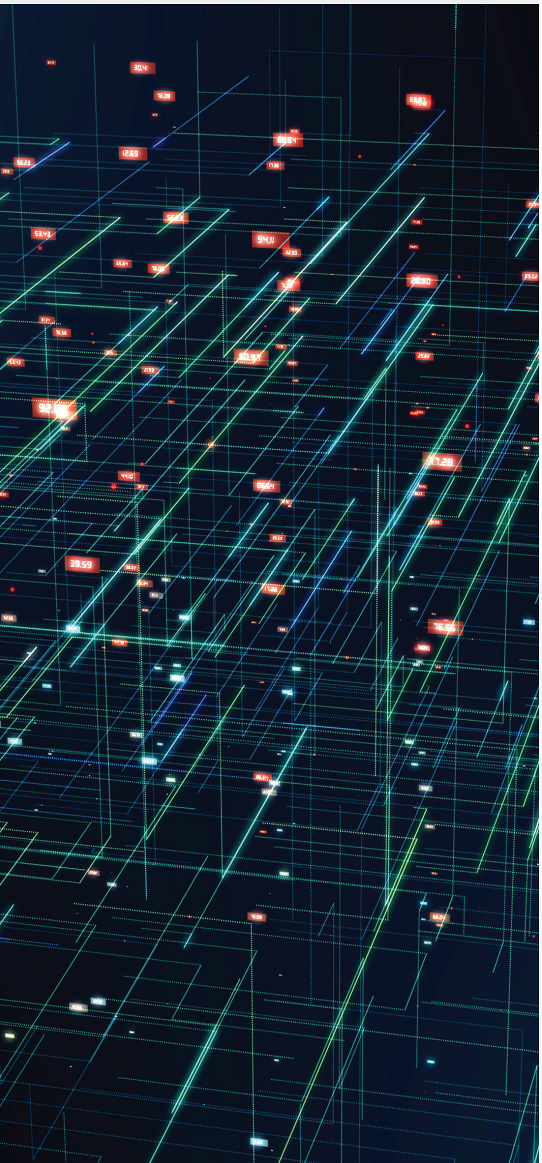
For capital markets trading teams, gaining and maintaining a competitive edge relies on their ability to react quickly to market shifts. Predicting those market movements leads to increased artificial intelligence (AI) model complexity, all while needing to maintain low latency and high throughput. For research teams who influence trading models that are ultimately deployed in live trading scenarios, making intelligent decisions within a given latency window is critical and is impacted by system-level ease of use and computational efficiency.

STAC-ML* Markets (Inference) is a vendor-independent benchmark suite standard for machine learning (ML) inference on real-time financial data from the STAC Benchmark Council*. This benchmark allows trading firms to objectively compare different platforms for latency, throughput, and efficiency in ML inference.

Agilex™ 7 FPGAs F-Series deployed on BittWare's IA-840f accelerator card, paired with the VOLLO* inference library from Myrtle.ai, form a complete, scalable, out-of-the-box solution to help trading teams manage portfolio risk and make real-time sequence predictions on live market data. The key performance metrics of latency, throughput, and rack space efficiency are where the complete solution using Agilex 7 FPGAs F-Series shines, especially in the STAC-ML Markets (Inference) benchmark test for the Tacana Suite.

Of all submissions in the STAC-ML benchmark from various vendors, the system under test (SUT) featuring VOLLO[1] achieved outstanding latency. The SUT achieved a 99th percentile latency as low as 5.1 microseconds in the smallest long short-term memory (LSTM) models (LSTM_A of 16 KB parameters)[2] shown in Figure 1. On the same model, when running four model instances simultaneously, the SUT achieved a throughput of 823,667 inferences per second[3].

As model and parameter sizes increase, so do complexity and pressure on computational efficiency. Using the brain floating-point 16 (bfloat16) format across model sizes ranging from 16 KB to 1 MB, VOLLO's performance consistently scales by retaining solid precision across model and parameter sizes. When the system has a higher supported number of model instances (NMIs), the throughput and space efficiency metrics increase.



[1] SUT ID: MRTL230426, https://stacresearch.com/news/MRTL230426

[2] STAC-ML.Markets.Inf.T.LSTM_A.[1,2,4].LAT.v1

[3] STAC-ML.Markets.Inf.T.LSTM_A.4.TPUT.v1

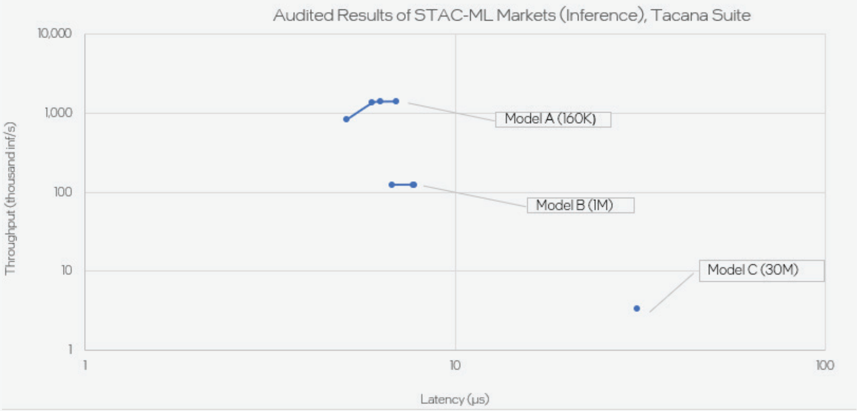


Figure 1. Audited results for VOLLO SUT in the STAC-ML Markets (Inference), Tacana Suite.

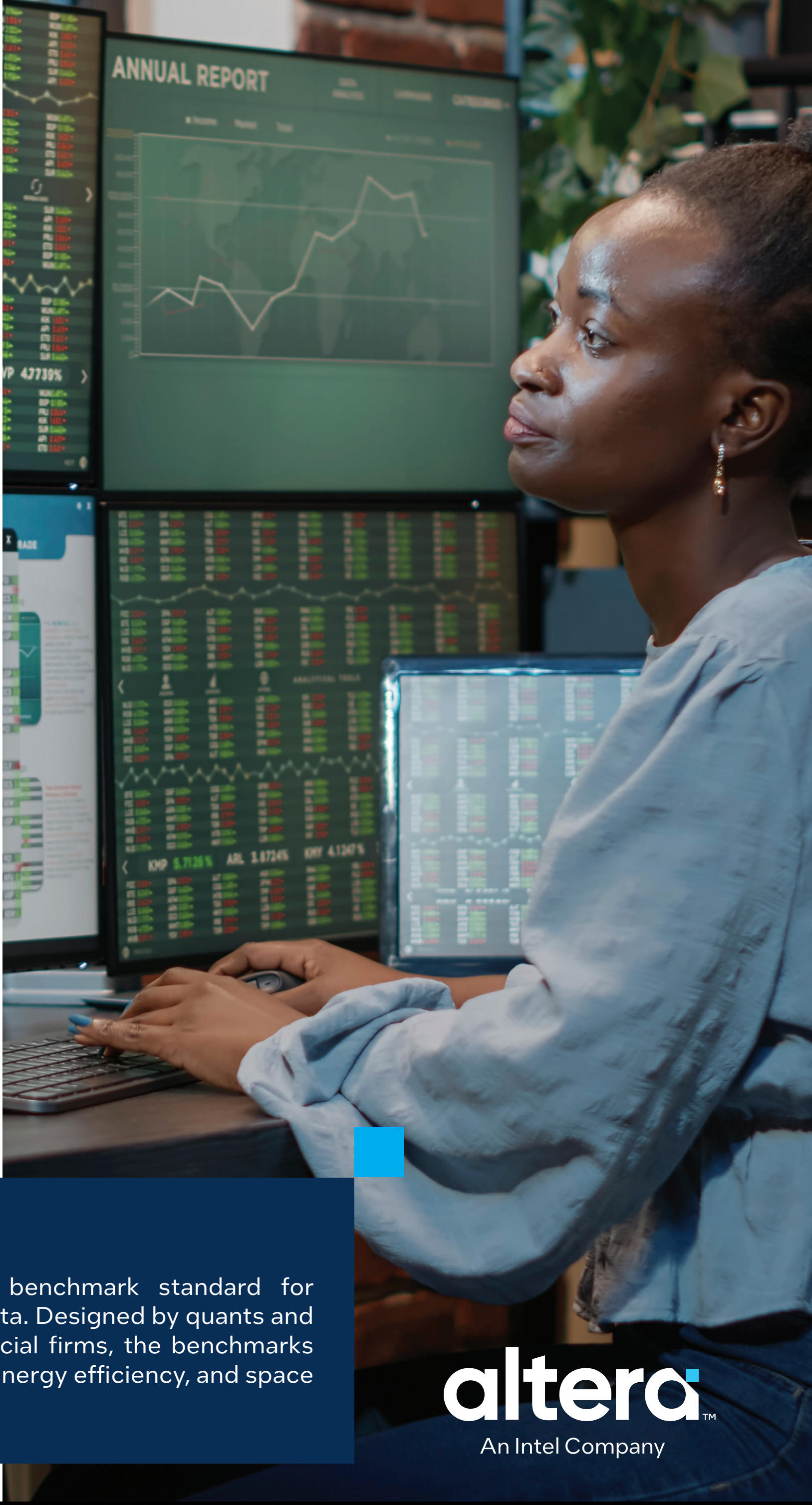
See backup for workloads and configurations. Results may vary.

The native support for hardened bfloat16 within Agilex 7 FPGAs F-Series and the 3rd Gen Intel® Xeon® Scalable processors meets a growing industry trend towards lower precision arithmetic for efficiency. The hardened bfloat16 support and 259 Mb of M20K on-chip memory blocks within Agilex 7 FPGAs F-Series combine for optimal acceleration of the recurrent networks in the STAC-ML benchmark – allowing any LSTM model to be implemented in FPGA logic with deterministic delay and ultimate power efficiency. By deploying Myrtle.ai's VOLLO inference library on Intel® technology, trading teams can expect to retain the required precision while achieving outstanding inference performance. Additional advantages of the VOLLO inference library include:

- ✓ Simple to program
  - No specialist FPGA knowledge is required
  - Use existing ML development environment
  - Train in PyTorch* or TensorFlow*, and export in ONNX*
- ✓ Flexible and scalable
  - Run up to 48 parallel models in a 1U server
  - Deploy multiple ML applications on each accelerator card
  - Supports a wide range of LSTM-based models

The audit results show that capital markets trading teams can use this solution to make intelligent decisions quickly while using minimal rack space.

Accelerate time to insight today by deploying the full solution to achieve maximum throughput and low latency – all without the need to maintain specialist FPGA teams to sharpen competitive edge.



What is a STAC-ML benchmark?

STAC-ML Markets (Inference) is the technology benchmark standard for solutions that can run inference on real-time market data. Designed by quants and technologists from some of the world's leading financial firms, the benchmarks test a technology stack's latency, throughput, quality, energy efficiency, and space efficiency across three model sizes and different NMIs.

altera™
An Intel Company

System Configuration

SUT ID: MRTL221125: STAC-ML™ Pack for Myrtle.ai VOLLO™ with 4 x BittWare IA-840f (Intel Agilex® AGF027 FPGA) Cards in a BittWare TeraBox™ 1402B Server

SUT ID: MRTL230426: STAC-ML™ Pack for Myrtle.ai VOLLO™ (Rev B) with 4 x BittWare IA-840f (Intel Agilex® AGF027 FPGA) Cards in a BittWare TeraBox™ 1402B Server

Notices & Disclaimers

Performance varies by use, configuration and other factors. Learn more at www.Intel.com/PerformanceIndex.

Performance results are based on testing as of dates shown in configurations and may not reflect all publicly available updates. See backup for configuration details. No product or component can be absolutely secure.

Your costs and results may vary.

Intel does not control or audit third-party data. You should consult other sources to evaluate accuracy.

Intel® technologies may require enabled hardware, software, or service activation.

* Other names and brands may be claimed as the property of others.

Details related to the STAC-ML benchmark can be found at https://stacresearch.com/ml

"STAC" and all STAC names are trademarks or registered trademarks of the Securities Technology Analysis Center, LLC.

Myrtle.ai and VOLLO are registered trademarks of Myrtle Software Ltd.