STAC Update:
Real-time Decisions

Peter Nabicht
President, STAC

peter.nabicht@STACresearch.com

Overview

- FPGA Special Interest Group
- STAC-ML Markets (Inference)
  - Including new results!

FPGA Special Interest Group

Current collaborations: 3 main projects

- RapidWright / RapidStream improvements, including
  - Common requirements, requests, and prioritized bugs
  - Collaborating with developers at AMD at a deeper level
- Language support
  - Jointly contribute to VHDL and SystemVerilog projects that check canonical language feature support in other tools
  - Use to convey of critical features to vendors
- Joint development of open-source Switch and/or NIC reference implementation
  - Exploring currently existing projects as starting points
  - Focus on the primary needs of trading firms

Education

- Previously
  - Financial firms FPGA developers presented different build, test, and deploy pipelines
  - RapidWright project deep dive led by project engineers from AMD
- Upcoming
  - Tutorial for CXL for FPGA to CPU communication and impact on development from Intel

STAC-ML Markets Inference

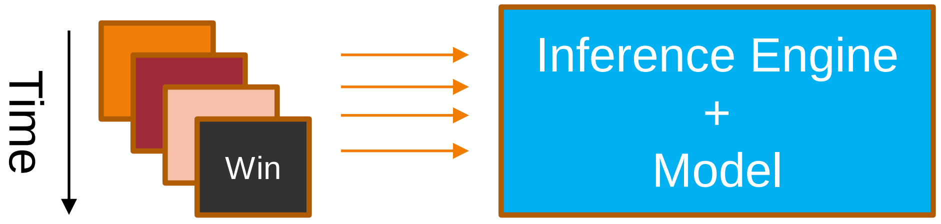
STAC-ML Markets (Inference) : Basics

- LSTM models inferring on simulated market data features
- Goal: isolate inference performance
  - Inference engine software
  - Underlying processors, memory, accelerators, etc.
  - Anything required to optimally use the former with the latter (e.g., data transfer to processor memory)
- Metrics:
  - Latency, throughput, error, power efficiency, space efficiency, cost
- Benchmarks allow any level of precision (including mixed-precision)

Two benchmark suites

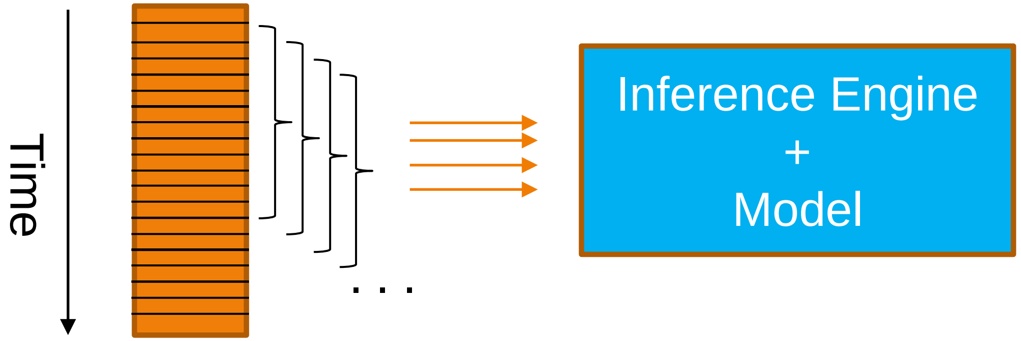
Sumaco

- Operates on fully populated, unique windows of time-series data/features
- Examples:
  - Inference over the recent past in response to an asynchronous event
  - One model may be used to reason about multiple instruments



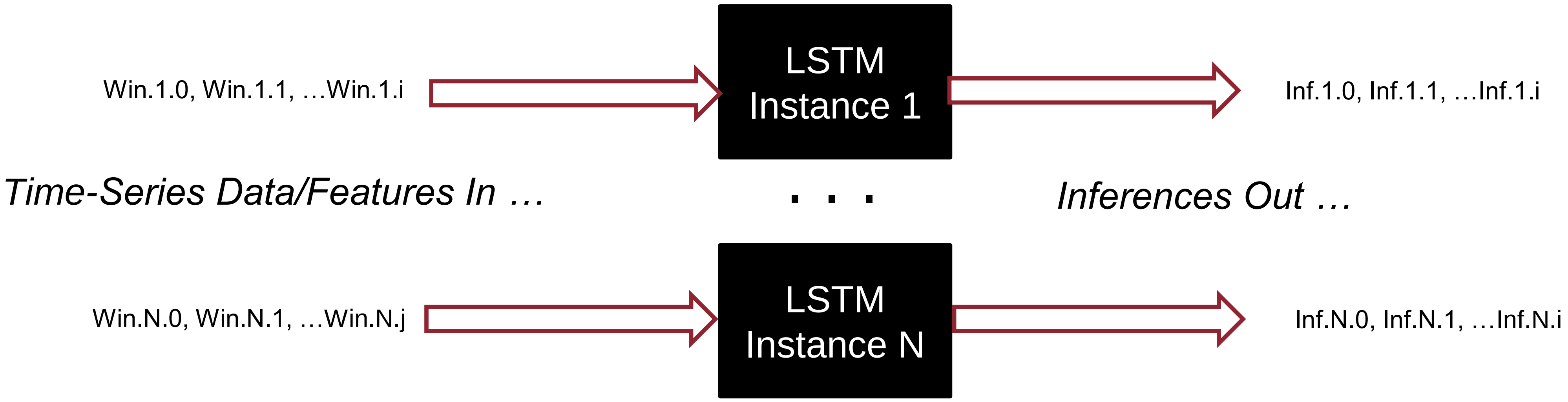
Tacana

- Operates on sliding windows of a single time-series of data/features
- Example:
  - Inference every tick or bar
- May provide lowest possible tick-to-inference latency



STAC® SECURITIES TECHNOLOGY ANALYSIS CENTER

Benchmark Schematic; Scaling Dimensions


Win.1.0, Win.1.1, …Win.1.i
LSTM Instance 1
Inf.1.0, Inf.1.1, …Inf.1.i
Time-Series Data/Features In …
. . .
Inferences Out …
Win.N.0, Win.N.1, …Win.N.j
LSTM Instance N
Inf.N.0, Inf.N.1, …Inf.N.i

- Model size
  - Three are currently specified
  - Input data window scales with model size
- Number of Model Instances running in parallel
  - As specified by the SUT provider
  - Performance / efficiency per model instance is key for co-located inference

Use Cases and Optimizations

- Different Use Cases:
  - Trading – Latency Optimization
  - Backtesting – Throughput Optimization
- Optimization tradeoffs (latency vs throughput vs efficiency vs error) are up to the SUT provider
  - The benchmarks do not assume an inference application
  - The tests collect all metrics every time, no matter the optimization goal
  - Any quantization scheme allowed, if used consistently

STAC-ML Markets (Inference) - Comparability

- The benchmark is agnostic to the architecture of the SUT and inference engine, and the precision of the computation
- Report readers are free to draw their own conclusions
- STAC only allows direct competitive comparisons if all the following are true:
  - Same suite (Tacana to Tacana, or Sumaco to Sumaco)
  - The same LSTM model
  - Error results are comparable
    - SUT A can compare to SUT B if SUT A’s error is strictly less than, or only slightly greater than SUT B’s
  - All performance comparisons must include an efficiency comparison to provide context
  - All latency comparisons must include a throughput comparison for context

Myrtle.ai tested the Tacana Suite with FPGA as accelerator

Last year did STAC-ML Sumaco (MRTL221125) and now Tacana!

- STAC-ML Pack for Myrtle.ai VOLLO™ (Rev B)
- VOLLO SDK 0.2.0
- VOLLO Accelerator 0.2.0
- Ubuntu Linux 20.04.5 LTS
- BittWare TeraBox™ 1402B (1U)
  - 4 x BittWare IA-840f-0001 each with
    - Intel® Agilex™ AGF027 FPGA
    - 4 x 16 GiB DDR4 @ 2666 MHz
  - 1 x Intel® Xeon® Platinum 8351N CPU @ 2.40 GHz
  - 4 x 8 GiB Micron DDR4 @ 2933 MHz (32GiB total)
- Latency-optimized, bfloat16 precision

Myrtle.ai

www.STACresearch.com/MRTL230426

STAC®
SECURITIES TECHNOLOGY ANALYSIS CENTER

Results highlights – Myrtle.ai

- *For LSTM_A (the smallest model) the 99p latency was:*[1]
  - *5.07 µs – 5.08 µs Across 1, 2 & 4 model instances tested (NMI)*
  - *5.97 µs with 8 NMI*
  - *6.96 µs with 24 NMI*

- *For LSTM_B the 99p latency was:*[2]
  - *6.89 µs with 1 NMI*
  - *6.77 µs with 2 NMI*
  - *7.75 µs with 8 NMI*

www.STACresearch.com/MRTL230426

1. STAC-ML.Markets.Inf.S.LSTM_A.[1,2,4,8,24].LAT.v1
2. STAC-ML.Markets.Inf.S.LSTM_B.[1,2,8].LAT.v1

Results highlights – Myrtle.ai

- *For LSTM_C (the largest model) the 99p latency was:*[1]
  - *31.0 µs with 1 NMI*

- *LSTM_A with 24 NMI achieved the following throughput and efficiency:*[2]
  - *1.4M inferences / second*
  - *1.4M inferences / second / cubic foot*
  - *2.3M inferences / second / kW*

Myrtle.ai

www.STACresearch.com/MRTL230426

1. STAC-ML.Markets.Inf.S.LSTM_C.[1].LAT.v1
2. STAC-ML.Markets.Inf.S.LSTM_A.12.[TPUT,SPACE_EFF,ENERG_EFF].v1