STAC-ML Update

Bishop Brock
Head of Research, STAC
bishop.brock@STACresearch.com

Peter Nabicht
President, STAC
peter.nabicht@STACresearch.com

STAC-ML Markets (Training) Benchmark: Underway

- Existing ML training benchmarks are not *specific* to Finance:
  - They focus on qualitative problems
  - Finance requires good quantitative models
- We spoke to many both inside and outside of the Working Group
- Came back to the Working Group with several candidate use cases
  - Value to the end user
  - The ability to fairly evaluate the quality of benchmark solutions
- Consensus – Focus on complex derivative modelling
- Now detailing a proposal - **Join us!**

www.STACresearch.com/ML

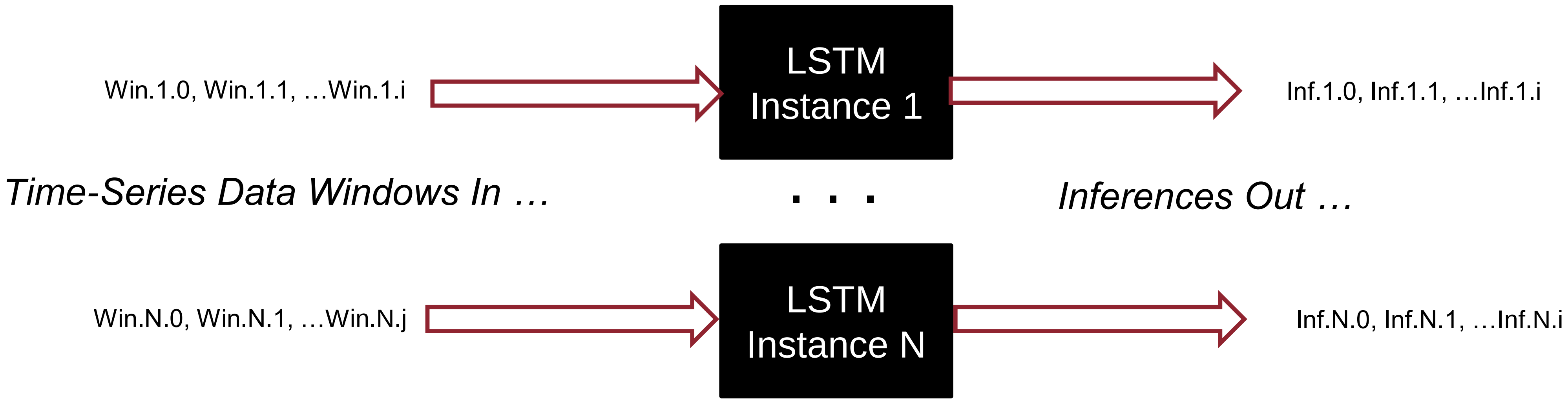
NEWS FLASH

- Benchmark results on three different compute accelerators!
- Will get to those shortly…

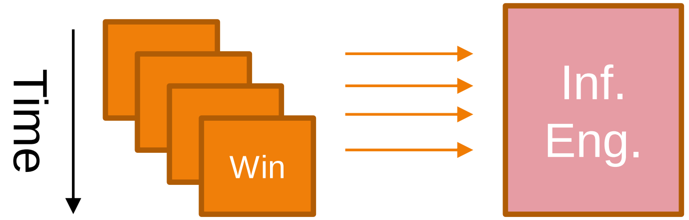
Background - STAC-ML Markets (Inference)

- STAC-ML provides a framework for full-stack evaluation
- Three users of STAC-ML
  - STAC
  - Vendors
  - Financial firms
- I will talk about all three

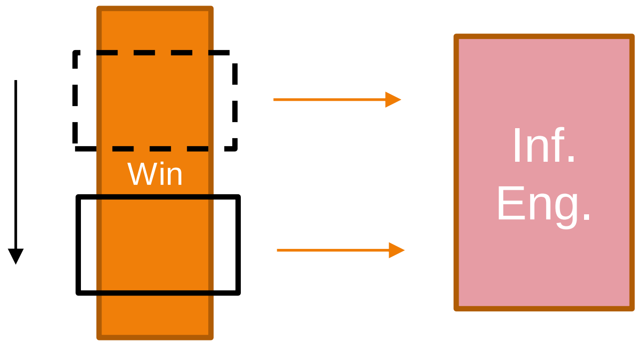
Time-Series Inference using LSTM Models: Perf./Eff./Scalability



- Sumaco – Fixed, Unique Window



- Tacana - Sliding Window (Streaming)



STAC SECURITIES TECHNOLOGY ANALYSIS CENTER

Information for three use cases

1. Latency at any cost
2. Throughput for a given latency
3. Throughput at any cost

Research Available to ML STAC-Track Subscribers

- GCP Cloud SUT
  - Latency- and Throughput-optimized configurations for ONNX inference
- TensorFlow Performance (on CPU)
  - Looked at different ways to configure TensorFlow for inference
- Azure Cloud-SUT Jamboree (Coming up)
- All research available via free trial for remainder of 2022
  - For those responsible for ML research and infrastructure

council@STACresearch.com

STAC® SECURITIES TECHNOLOGY ANALYSIS CENTER

STAC-ML Markets (Inference) Azure Cloud-SUT Jamboree!

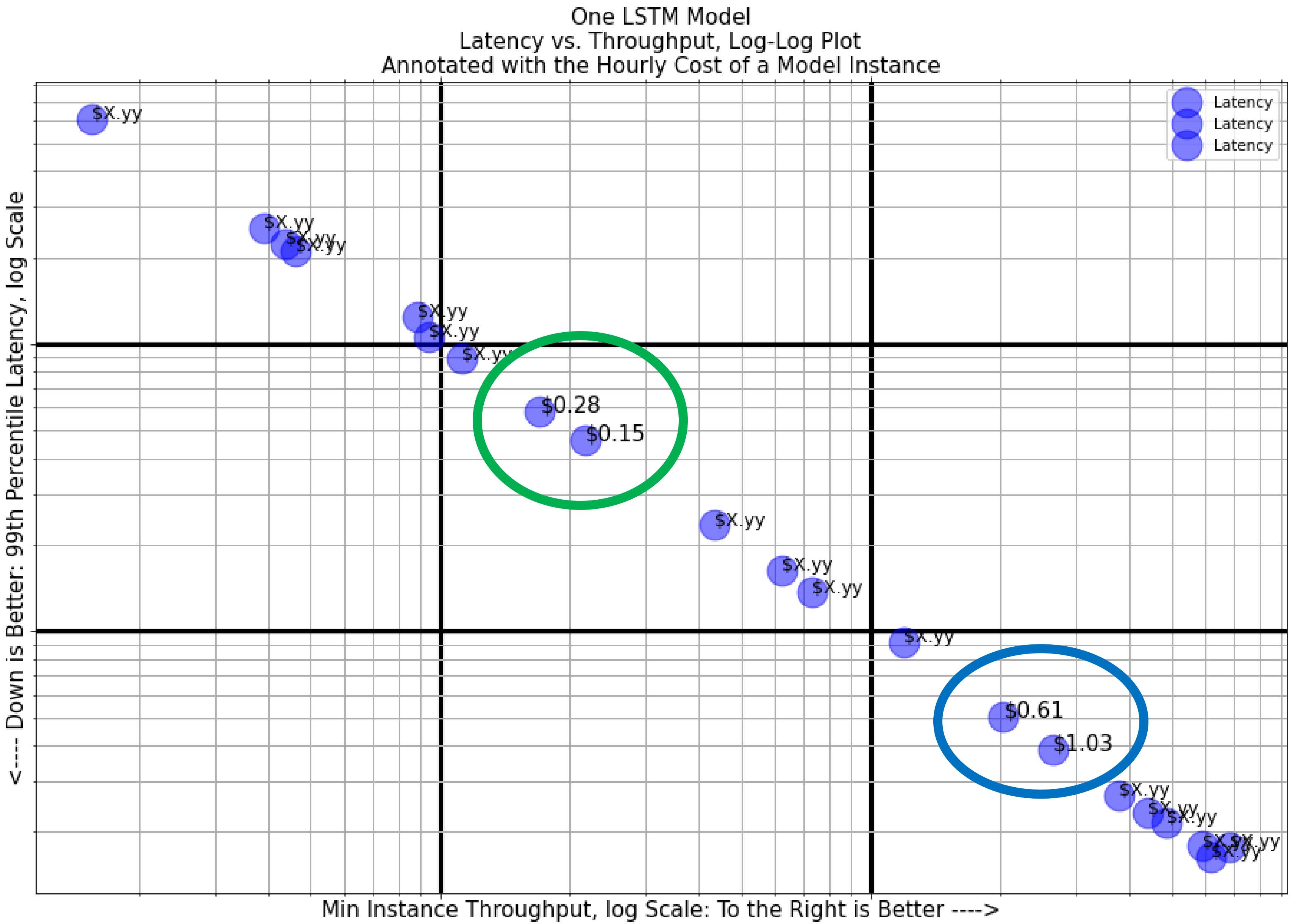
- Goal: compare 3 CPU architectures for inference
  - Intel, AMD, Ampere (ARM)
- Used the STAC “Naive” Python implementation with ONNX
- Tested on Microsoft Azure

Thanks to Microsoft for supporting the STAC community by providing credits for this research!

- Tested two configs for each VM (latency optimized, tput optimized)
- All 6 reports are in the STAC Vault & comparison report will be available soon
- No vendors participated in the setup and optimization of the SUTs

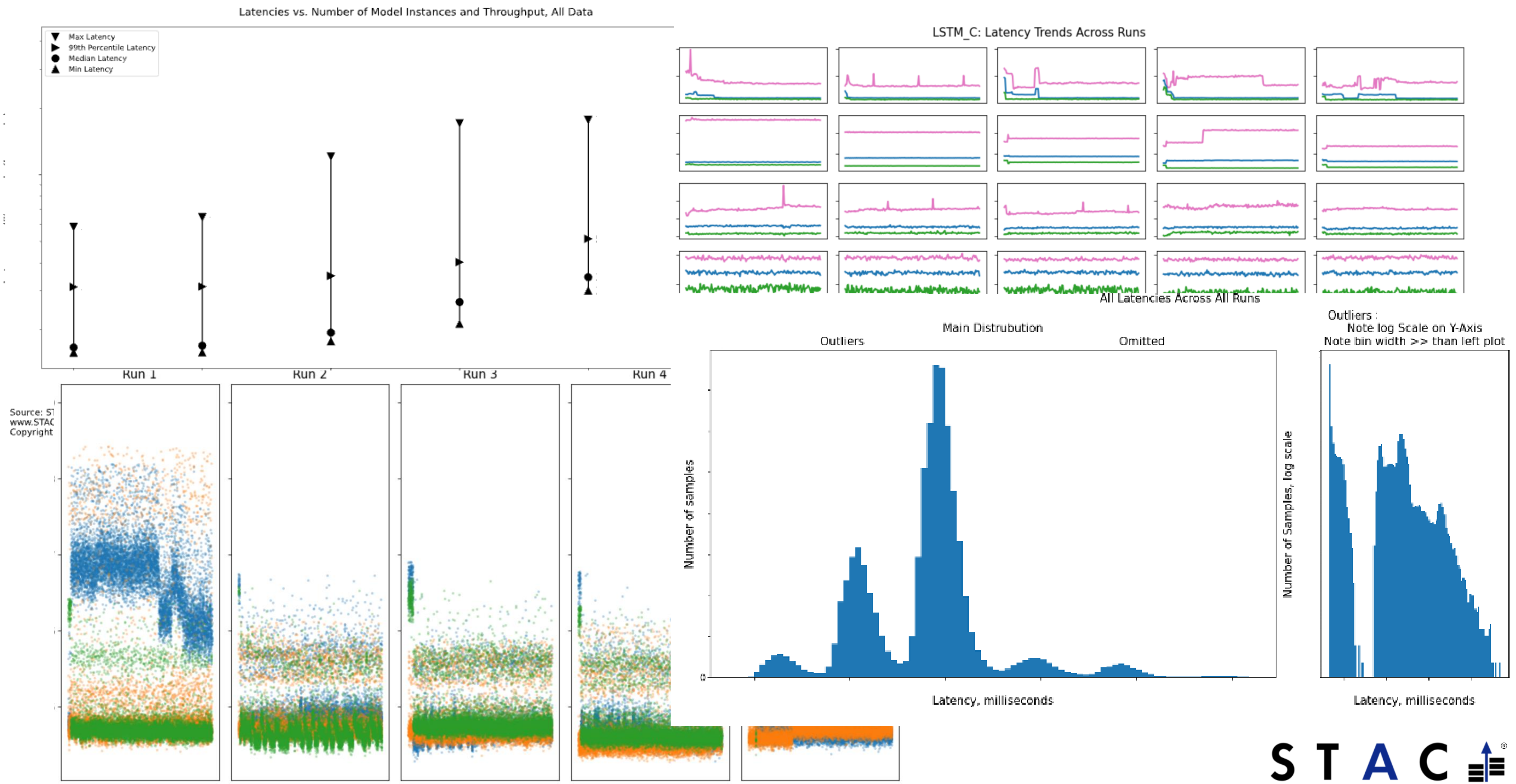
STAC® SECURITIES TECHNOLOGY ANALYSIS CENTER

Research Summary Note: Business-oriented comparisons



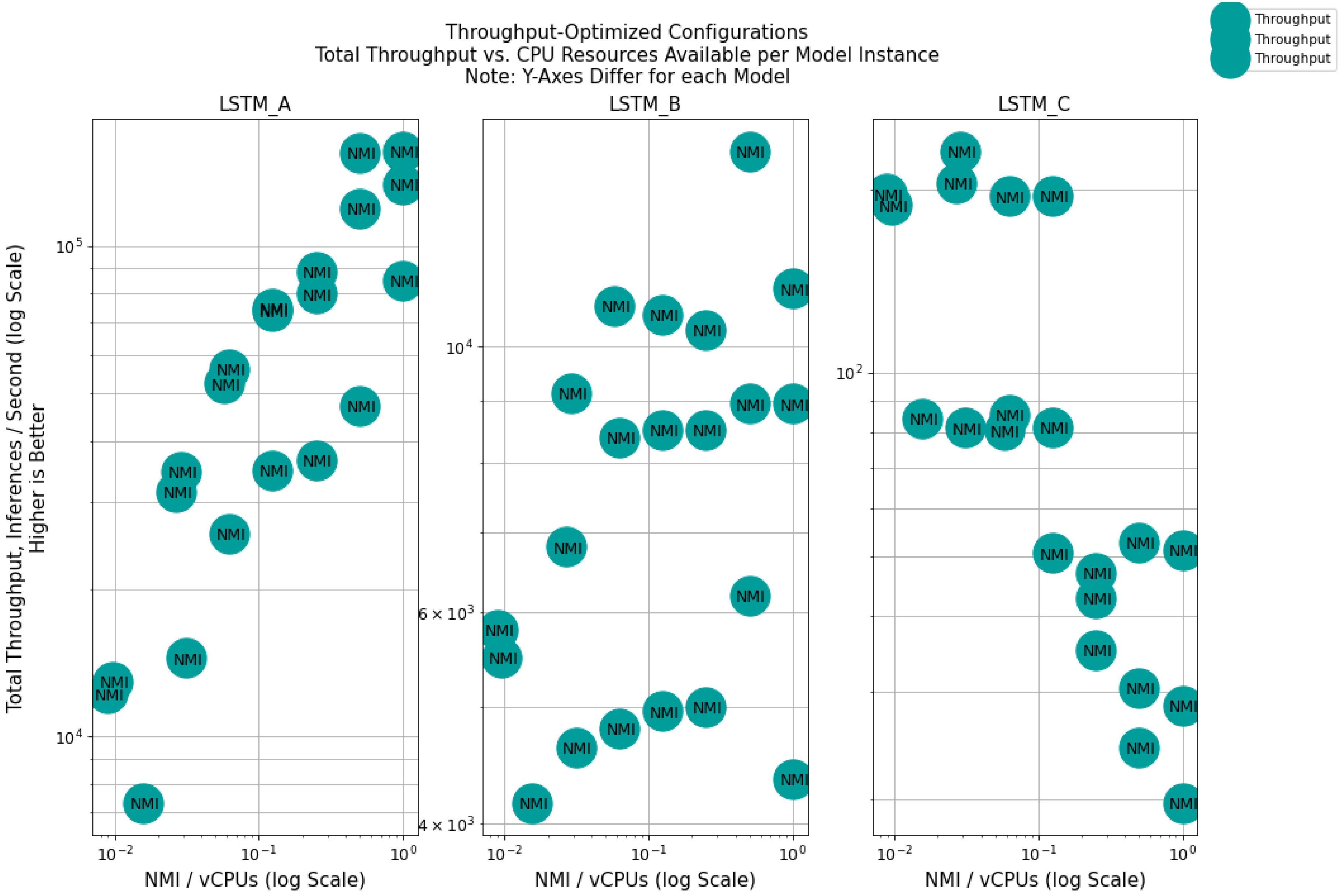
STAC SECURITIES TECHNOLOGY ANALYSIS CENTER

Detailed analysis available for each SUT



STAC
SECURITIES TECHNOLOGY ANALYSIS CENTER

Research Summary Note: Throughput Performance Comparisons



STAC SECURITIES TECHNOLOGY ANALYSIS CENTER

Groq was first public tested SUT!

- STAC-ML Pack for GroqWare™ (Rev A)
  - Version of STAC “Naive” implementation adapted for GroqWare™ APIs
  - Effectively FP16
- GroqWare™ SDK 0.9.0.5 devtools and runtime
- Python 3.8.15; NumPy 1.23.4
- Ubuntu Linux 22.04.1 LTS
- GroqNode™ GN1-B8C-ES:
  - 8 x GroqCard™ 1 Accelerators (GC1-010B)
  - 2 x AMD EPYC™ 7413 24-core CPUs @ 2650 MHz
  - 16 slots x 64GiB DDR4 - 1024GiB Total

www.STACresearch.com/GROQ221014

STAC® SECURITIES TECHNOLOGY ANALYSIS CENTER

Result highlights – Groq

- For small model LSTM_A, across 1, 2 and 4 simultaneously running model instances (NMI):
  - Worst case 99th percentile latency was 56.4 µsec (STAC-ML.Markets.Inf.S.LSTM_A.4.LAT.v1)
  - 99th percentile latencies varied 1% (from 55.9 to 56.4 µsec) (STAC-ML.Markets.Inf.S.LSTM_A.[1,2,4].LAT.v1)
  - The widest spread from minimum to 99th percentile latency was 6% (53.4 to 56.4 µsec) (STAC-ML.Markets.Inf.S.LSTM_A.4.LAT.v1)

www.STACresearch.com/GROQ221014

STAC®
SECURITIES TECHNOLOGY ANALYSIS CENTER

Results highlights – Groq

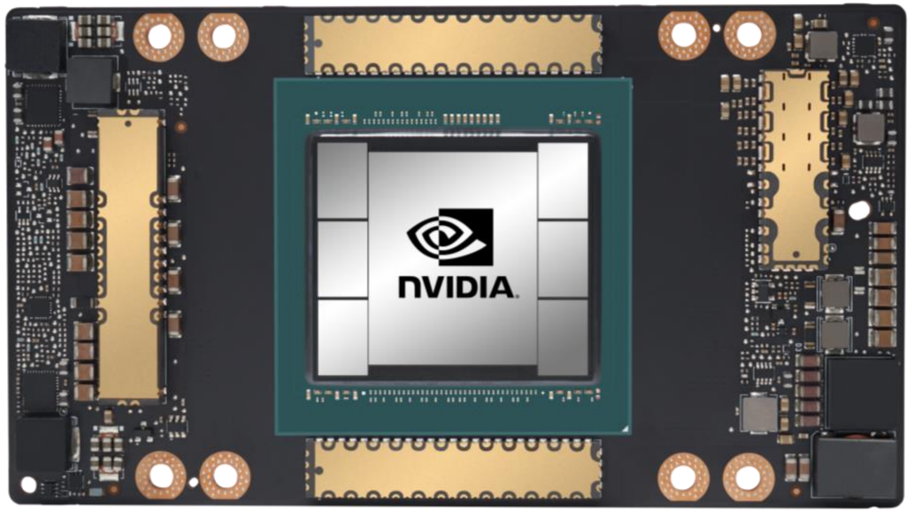
- For large model LSTM_C, across all NMI tested:
  - Worst case 99th percentile latency was 2.27 ms (STAC-ML.Markets.Inf.S.LSTM_C.8.LAT.v1)
  - 99th percentile latencies varied by 2% (from 2.72 to 2.77 ms) (STAC-ML.Markets.Inf.S.LSTM_C.[1,2,4,8].LAT.v1)
  - The widest spread from minimum to 99th percentile latency was 3% (2.68 to 2.77 ms) (STAC-ML.Markets.Inf.S.LSTM_C.8.LAT.v1)

www.STACresearch.com/GROQ221014

STAC SECURITIES TECHNOLOGY ANALYSIS CENTER

NVIDIA – 2 SUTs with same GPU-based stack

- STAC-ML Pack for CUDA and cuDDN (Rev A)
- NVIDIA CUDA Toolkit 11.7
- NVIDIA CUDA Deep Neural Network library (cuDNN) 8.4.1.50
- Ubuntu 20.04.5 LTS
- SuperMicro Ultra SuperServer SYS-620U-TNR
  - NVIDIA A100 80GB PCIe Tensor Core GPU
  - 2 x Intel Xeon Gold 6354 CPU @ 3.00GHz
  - 512GiB of memory

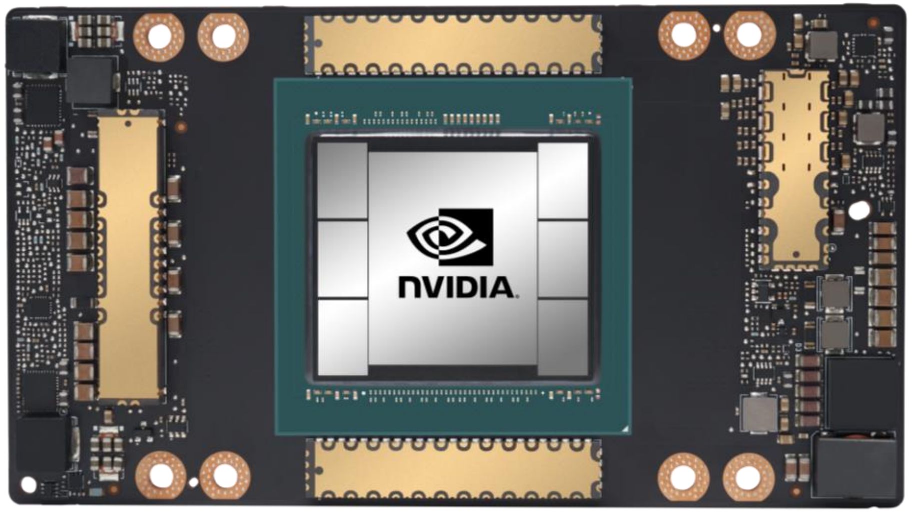


- Publishing results on two SUTs
  - Throughput optimized, Sumaco suite, FP16
  - Latency optimized, Tacana suite, FP32

Reports coming soon

Throughput optimized, Sumaco suite, FP16

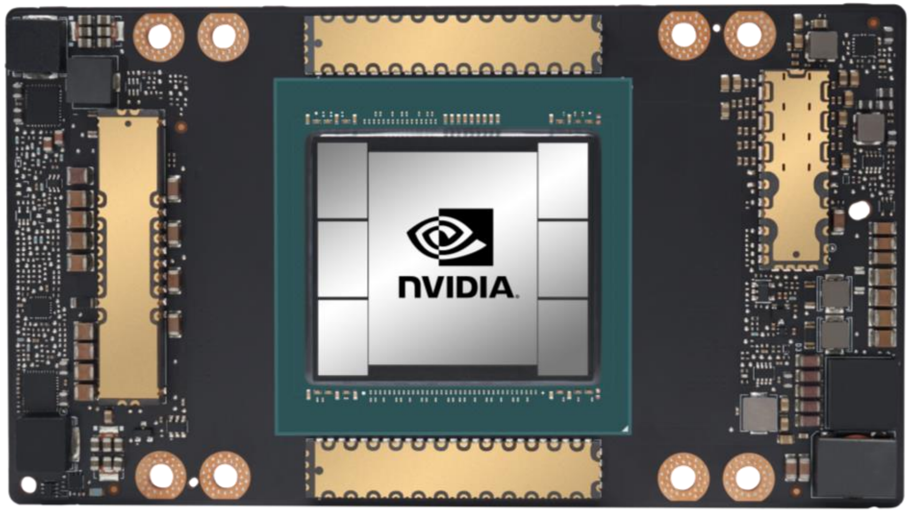
- Same stack configured to
  - Operate on a fixed window of unique updates (Sumaco)
  - Maximize throughput
  - Use FP16
- For LSTM_A, across all NMI tested:
  - Total throughput ranged from 1.63 to 1.71 M inf/sec (STAC-ML.Markets.Inf.S.LSTM_A.[1,2,4].TPUT.v1)
  - Energy efficiency ranged from 1.72 to 1.8 M inf/sec/kW (STAC-ML.Markets.Inf.S.LSTM_A.2.ENERG_EFF.v1)



Report coming soon

Throughput optimized, Sumaco suite, FP16

- For LSTM_B, across all NMI tested:
  - Total throughput was 191 K inf/sec (STAC-ML.Markets.Inf.S.LSTM_B.[1,2,4].TPUT.v1)
  - Energy efficiency was 206 K inf/sec/kW (STAC-ML.Markets.Inf.S.LSTM_B.[1,2,4]. ENERG_EFF.v1)
- For LSTM_C, across all NMI tested:
  - Total throughput was 12.8 K inf/sec (STAC-ML.Markets.Inf.S.LSTM_C.[1,2,4].TPUT.v1)
  - Energy efficiency was 17.7 K inf/sec/kW (STAC-ML.Markets.Inf.S.LSTM_C.[1,2,4]. ENERG_EFF.v1)

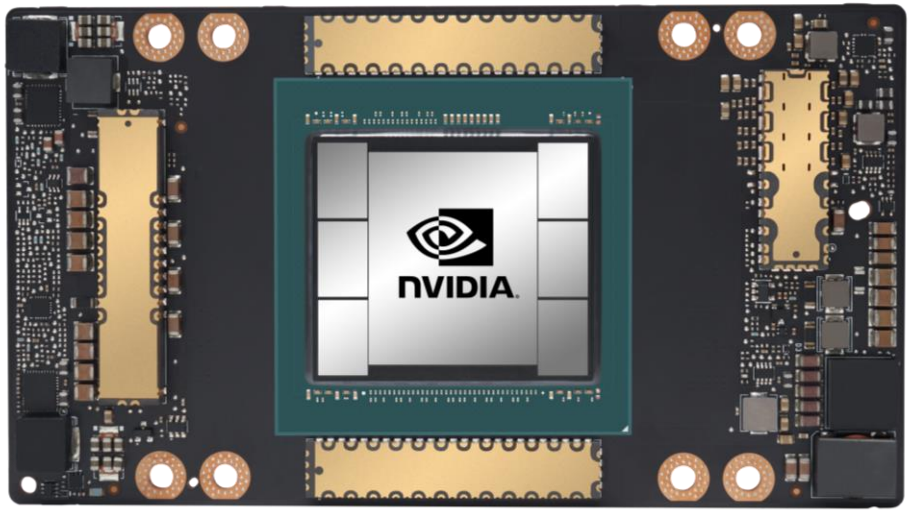


Report coming soon

STAC®
SECURITIES TECHNOLOGY ANALYSIS CENTER

Latency optimized, Tacana suite, FP32

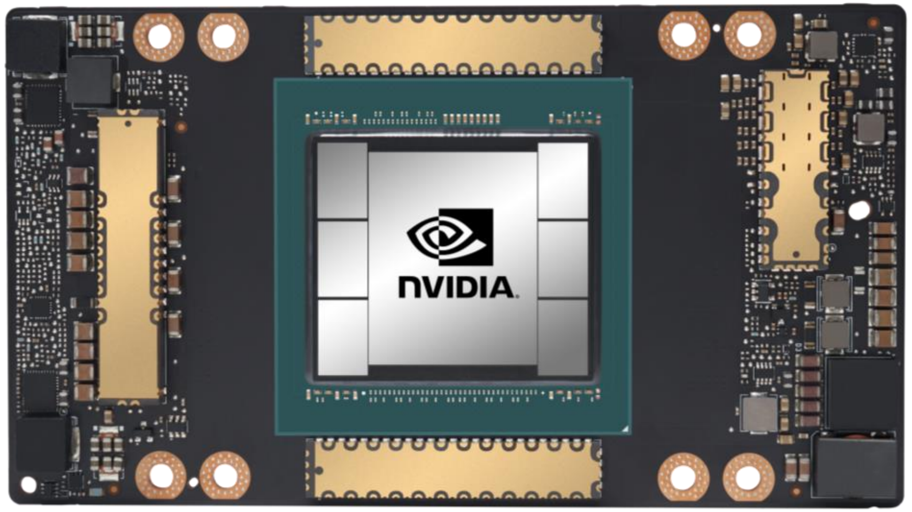
- Same stack configured to
  - Operate on a sliding window of updates (Tacana)
  - Minimize latency
  - Use FP32
- For LSTM_A the 99p latency :
  - With 1 NMI was 35.2 µsec (STAC-ML.Markets.Inf.T.LSTM_A.1.LAT.v1)
  - With 32 NMI was 58.8 µsec (STAC-ML.Markets.Inf.T.LSTM_A.32.LAT.v1)
- For LSTM_B the 99p latency:
  - With 1 NMI was 68.5 µsec (STAC-ML.Markets.Inf.T.LSTM_B.1.LAT.v1)
  - With 32 NMI was 149 µsec (STAC-ML.Markets.Inf.T.LSTM_B.32.LAT.v1)



Report coming soon

Latency optimized, Tacana suite, FP32

- For LSTM_C the 99p latency:
  - With 1 NMI was 640 µsec (STAC-ML.Markets.Inf.T.LSTM_C.1.LAT.v1)
  - With 16 NMI was 748 µsec (STAC-ML.Markets.Inf.T.LSTM_C.16.LAT.v1)
- Across all tested LSTM models and NMI, the largest outlier was 2.3x the median latency
  - Median latency 35 µsec, max latency 81 µsec (STAC-ML.Markets.Inf.T.LSTM_A.2.LAT.v1)



Report coming soon

STAC® SECURITIES TECHNOLOGY ANALYSIS CENTER

Myrtle.ai tested with FPGA as accelerator

- STAC-ML Pack for Myrtle.ai VOLLO™ (Rev A)
  - bfloat16 precision
- VOLLO SDK 0.1.0
- VOLLO Accelerator 0.1.0
- Ubuntu Linux 22.04.1 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)

Myrtle.ai

Report coming soon

STAC®
SECURITIES TECHNOLOGY ANALYSIS CENTER

Results highlights – Myrtle.ai

- 99p latencies across 1, 2, 3 & 4 NMI for:
  - LSTM_A were 24.0 – 24.1 µsec
  - LSTM_B were 64.8 µsec
  - LSTM_C were 1.35 ms

- For LSTM_A with 48 NMI:
  - Total throughput was 651 K inf/sec (STAC-ML.Markets.Inf.S.LSTM_A.48.TPUT.v1)
  - Space eff. was 647 K inf/sec/cubic foot (STAC-ML.Markets.Inf.S.LSTM_A.48. SPACE_EFF.v1)
  - Energy eff. was 1.2 M inf / sec/ kW (STAC-ML.Markets.Inf.S.LSTM_A.48. ENERG_EFF.v1)
  - The 99p latency was 73.9 µsec, which was 3.1x the 99th percentile latency of 1 NMI (STAC-ML.Markets.Inf.S.LSTM_A.[1, 48].LAT.v1)

Myrtle.ai

Report coming soon

STAC®
SECURITIES TECHNOLOGY ANALYSIS CENTER

Results highlights – Myrtle.ai

- For LSTM_B with 16 NMI:
  - The 99p latency was 147 µsec, which was 2.3x the 99p latency of 1 NMI (STAC-ML.Markets.Inf.S.LSTM_B.[1, 16].LAT.v1)

- Across all Models and NMI:
  - The widest percentage spread from median to 99p latencies was 7% (26.5 µsec to 28.4 µsec) (STAC-ML.Markets.Inf.S.LSTM_A.12.LAT.v1)

Myrtle.ai

Report coming soon

STAC-ML tools are ready for you, too

- Vendor implementations – See how it works
- Test harness software and analysis tools – Test your own stacks
  - In fact, test your own models!

www.STACresearch.com/ML