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

STAC®
SECURITIES TECHNOLOGY ANALYSIS CENTER

STAC-ML Markets (Inference)

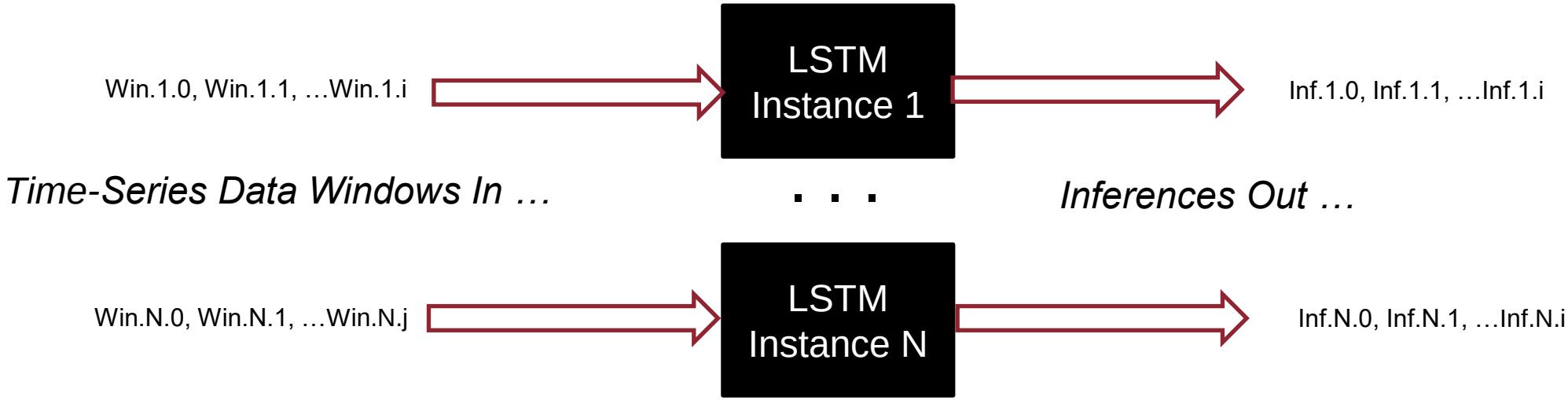
NEWS FLASH

- First audited results from Groq published today!
- Will get to that shortly…

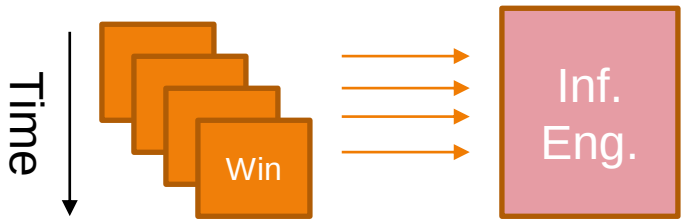
Background - STAC-ML Markets (Inference)

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

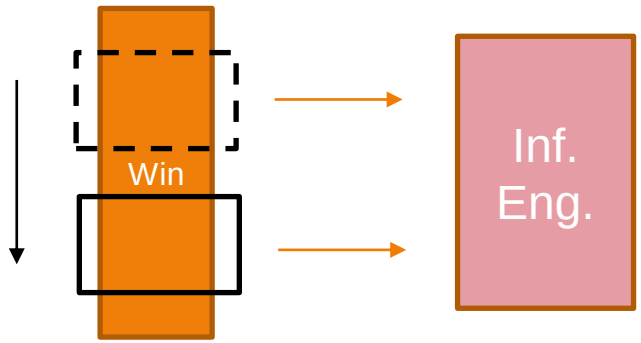
Time-Series Inference using LSTM Models: Perf./Eff./Scalability



- 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!

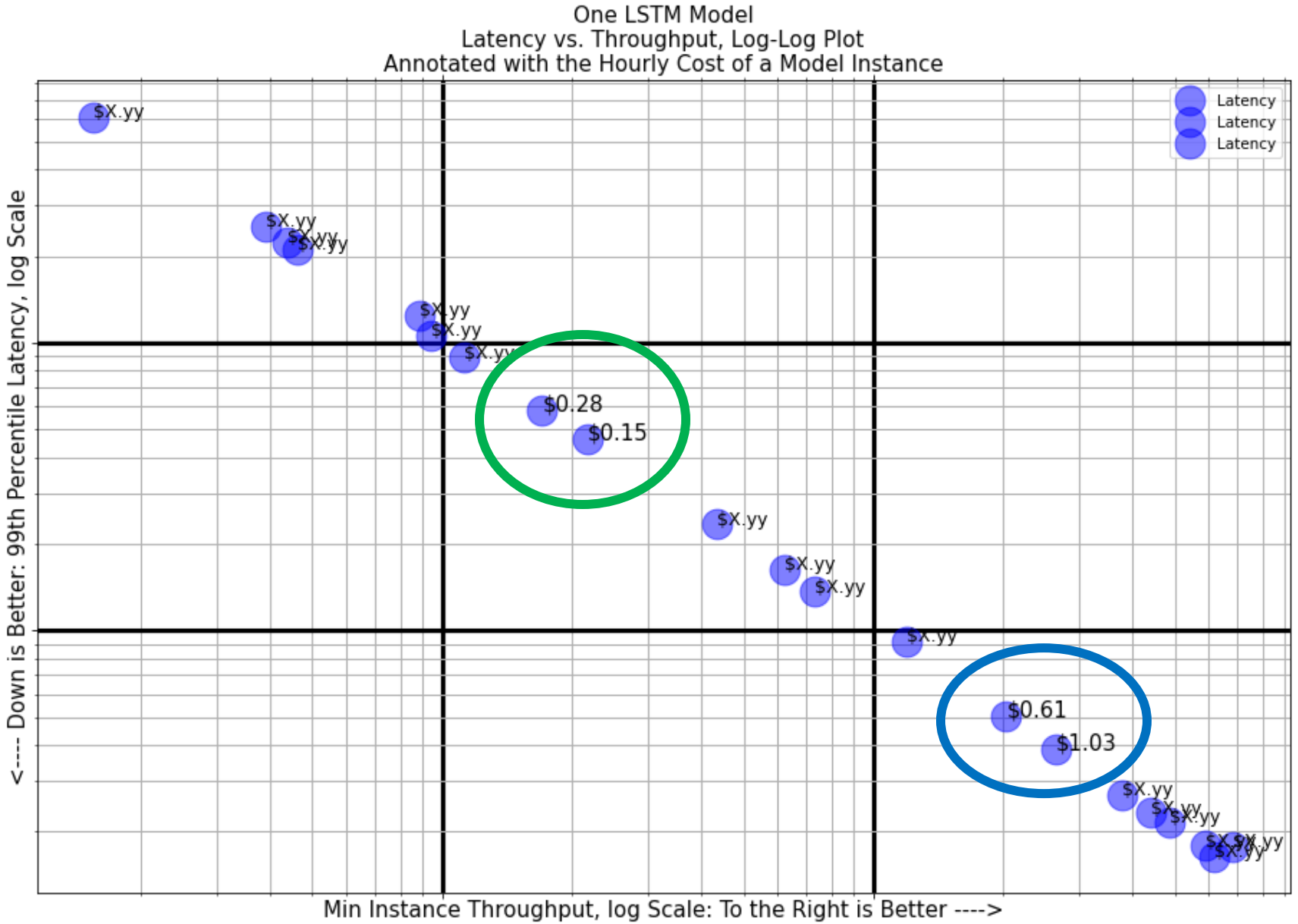
- Goal: compare 3 CPU architectures for inference
  - Intel, AMD, Ampere (ARM)
- Used the STAC "Naive" Python implementation with ONNX
- Tested on Microsoft Azure

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 & comparison report will soon be in the STAC Vault
- No vendors participated in the setup and optimization of the SUTs

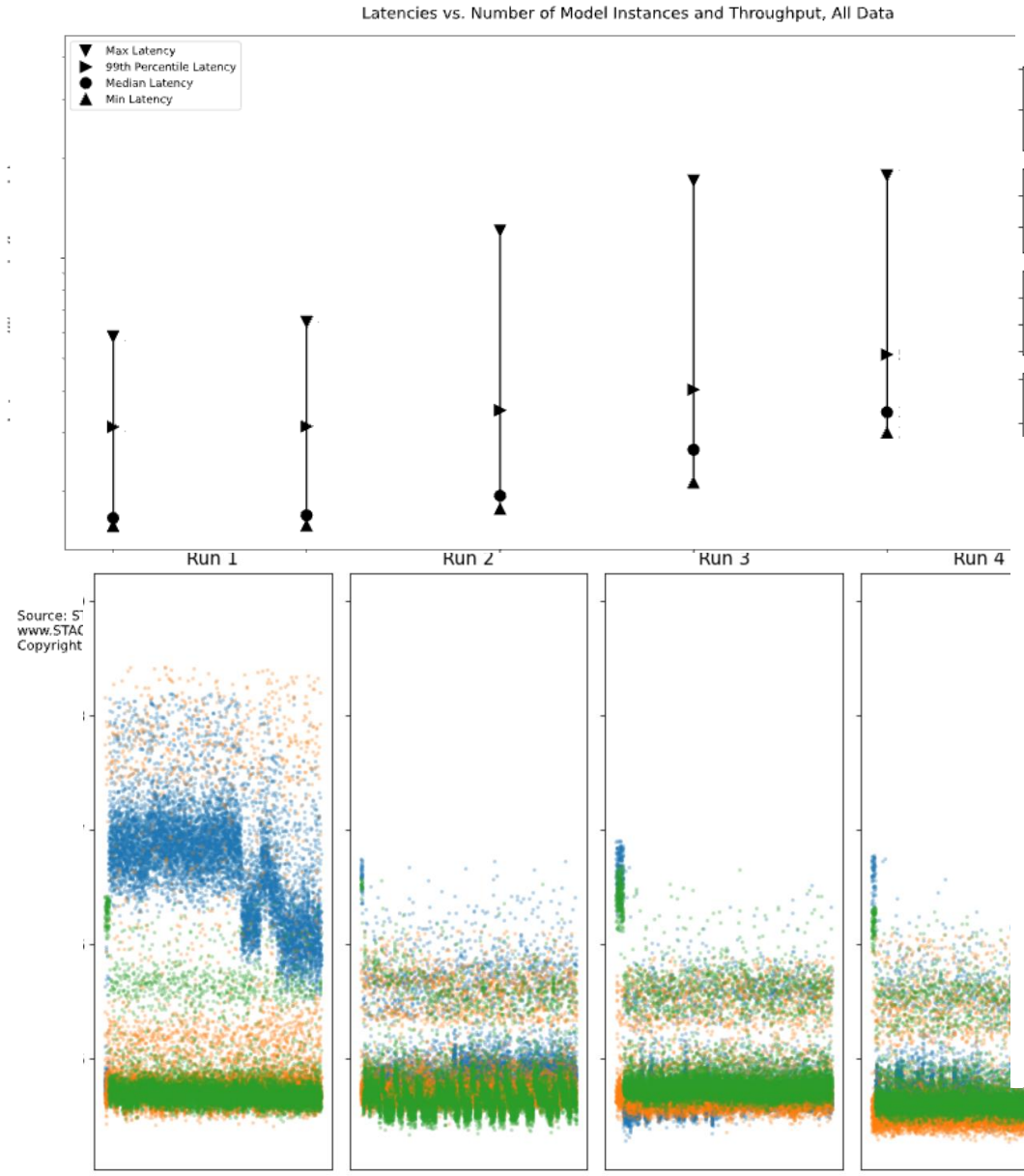
STAC®
SECURITIES TECHNOLOGY ANALYSIS CENTER

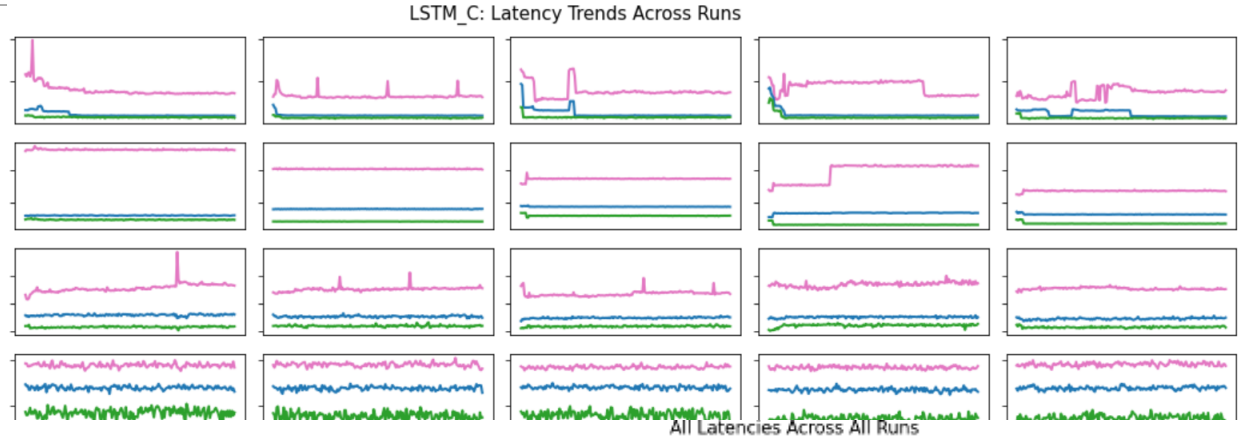
Research Summary Note: Business-oriented comparisons

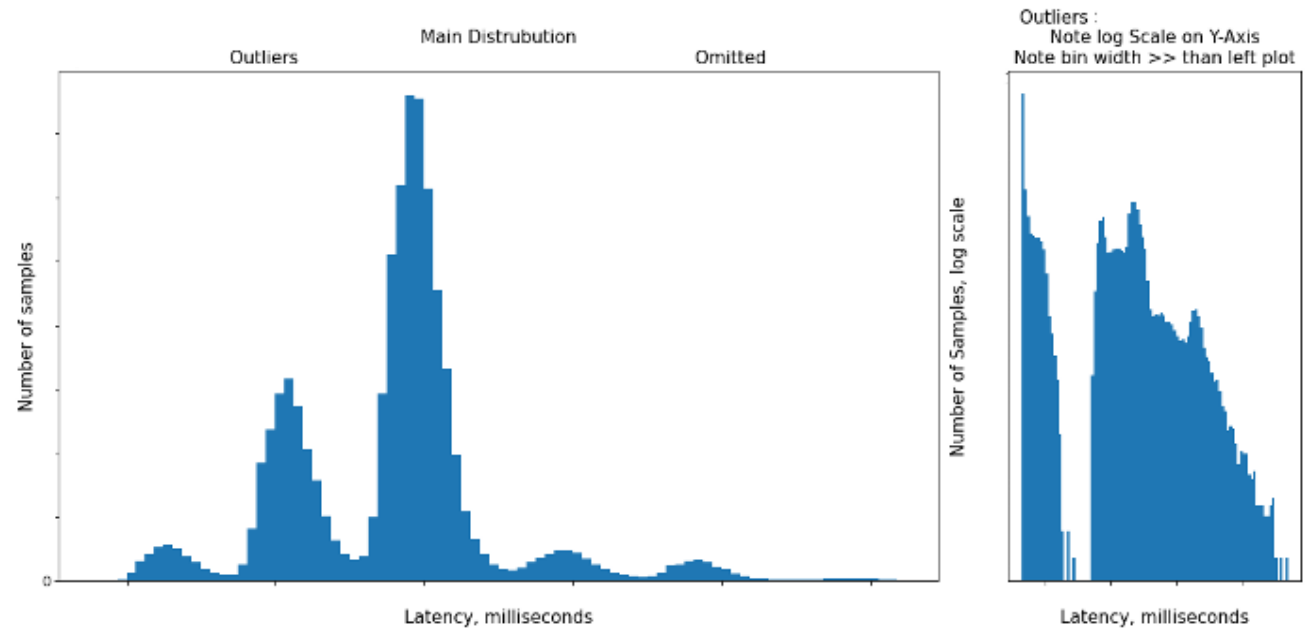


STAC®
SECURITIES TECHNOLOGY ANALYSIS CENTER

Detailed analysis available for each SUT

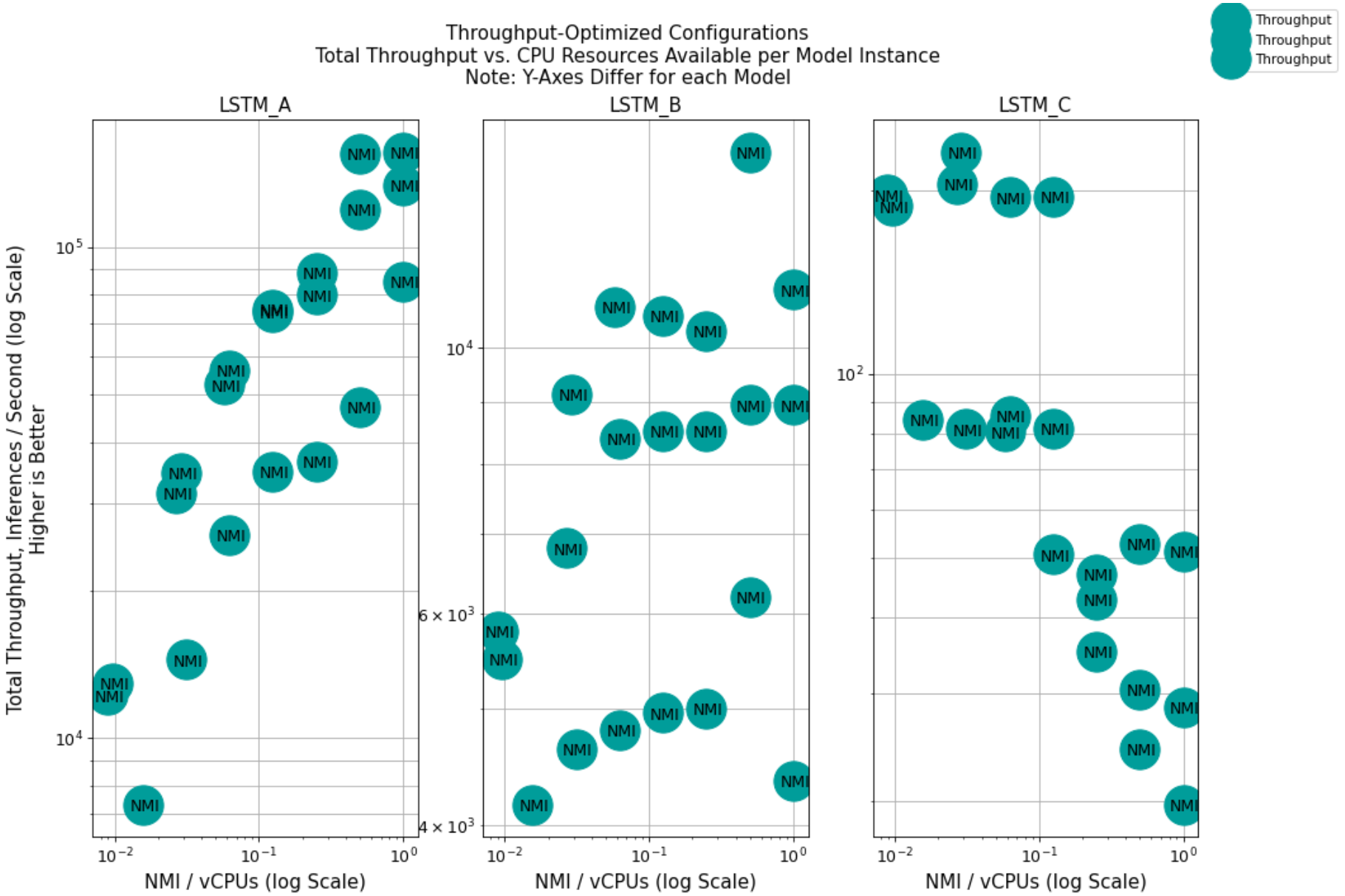






STAC®
SECURITIES TECHNOLOGY ANALYSIS CENTER

Research Summary Note: Throughput Performance Comparisons



STAC®
SECURITIES TECHNOLOGY ANALYSIS CENTER

First public tested SUT!

- STAC-ML Pack for GroqWare™ (Rev A)
  - Version of STAC “Naive” implementation adapted for GroqWare™ APIs
- 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

Results 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.LSTM_A.S.4.LAT.v1)
  - 99th percentile latencies varied 1% (from 55.9 to 56.4 µsec) (STAC-ML.Markets.Inf.LSTM_A.S.[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.LSTM_A.S.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

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

STAC®
SECURITIES TECHNOLOGY ANALYSIS CENTER