STAC Update:
Big Compute

Peter Nabicht
President, STAC

peter.nabicht@STACresearch.com

STAC AI:
Inference Benchmark

History

- Driven by user firms
  - Motivation: market making, hedging, customer pricing, etc.
  - STAC did a POC benchmark at the request of some trading firms: www.STACresearch.com/lstm_inference_poc (STAC Vault)
  - Additional banks, hedge funds, exchanges have since refined the POC specs
- Tech vendors have provided crucial input
- But control ultimately rests with users – i.e., those who must deliver business value from technology in the real world
  - Like all STAC Benchmarks

Latest status

- STAC AI Working Group has agreed on all major components
- Final specifications and documentation are underway
  - Including the official benchmark name
- STAC is finishing the test harness software and reference implementation
- Final approval expected this month (November 2021)

www.STACresearch.com/ai

STAC® SECURITIES TECHNOLOGY ANALYSIS CENTER

Basics

- LSTM models that simulate real models derived from market data
- 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, power efficiency, space efficiency, error
- Benchmarks allow any level of precision (including mixed-precision)
- Some sub-benchmarks decompose performance

Scale dimensions

- Model size
  - Three are currently specified
- Number of simultaneous model instances
  - Some are specified, the rest is open
- Optimization tradeoffs (latency vs throughput vs efficiency vs error) are up to the SUT provider
  - The tests collect all metrics every time, no matter the optimization goal

Next steps

- Ready for use November 2021
- Analytics STAC Track subscribers will have access to the specs and software
- Vendor members interested in running the benchmarks:
  - Contact council@STACresearch.com
- Users and vendors who want to influence this and future AI benchmarks:
  - Join the working group!

www.STACresearch.com/ai

STAC SECURITIES TECHNOLOGY ANALYSIS CENTER

STAC-A2:
Derivatives risk

STAC-A2: Risk computation

- Non-trivial Monte Carlo calculations
  - Heston-based Greeks for multi-asset, path-dependent options with early exercise
  - Metrics: Speed, capacity, quality, efficiency
- Numerous reports
  - Some public, some in the STAC Vault
- Premium STAC members get:
  - Reports in STAC Vault
  - Detailed config info on public and private reports
  - Code from vendor implementations of the benchmarks

www.STACresearch.com/**a2**

A few points on STAC-A2 for the uninitiated

- Some tests measure **response time** for a single option of given problem size
- **Throughput** measures time to handle a portfolio of options
- **Efficiency** relates throughput to power and space
- Each response-time workload is tested 5 times, back-to-back:
  - First run is the **COLD** run
  - Subsequent 4 are **WARM** runs
- COLD relates to real-world systems that must respond to heterogeneous problem classes
  - COLD time includes building memory structures, loading kernels, etc.
- WARM relates to real-world systems configured to handle numerous requests for the same problem class

STAC-A2 Pack for C (Naive Implementation)

- Developed by STAC
- C-language implementation
  - Compiled & built using GCC 9.3.1
  - Uses standard, widely available open-source mathematical and parallelization libraries
- No hardware specific optimizations
- No proprietary libraries
- Reasonably efficient
- Enables useful comparisons but results achieved are not the last word on the absolute performance attainable from the underlying SUT components

STAC-A2 / AMD / Comparing SEV-ES enabled and disabled

- Used STAC-A2 Pack for C (Naive Implementation) on two SUTs
- Only difference was AMD Secure Encrypted Virtualization-Encrypted State (SEV-ES) enabled/disabled
- SEV-ES is used with virtual machines
  - Encrypts CPU register contents when VM stops
  - Provides guest memory encryption
- Audit was run to illustrate the impact of SEV-ES on compute performance

www.STACresearch.com/NAIV210520

STAC-A2 / AMD / Comparing SEV-ES enabled and disabled

- Hardware:
  - Dell PowerEdge R6525
  - 2 x AMD EPYC 72F3 8-Core CPU @ 3.7GHz
  - 32 x 64GiB DDR4 DIMM @ 2933MT/s
- VM:
  - VMware ESXi 7.0 Update 2
  - 16 vCPUs
  - 1.5 TiB VMware Virtual RAM DIMM
  - SUSE Linux Enterprise Server 15 SP2 with kernel changed to 5.9.0-rc2-SEV-ES-orig-24.9-default+
- NAIV210520a SEV-ES disabled
- NAIV210520b SEV-ES enabled

www.STACresearch.com/NAIV210520

With SEV-ES enabled

- No change in the maximum paths or maximum assets handled
  - STAC-A2.β2.GREEKS.{MAX_PATHS/MAX_ASSETS}
- A 0.0% increase* in elapsed time for warm runs of the large Greeks benchmark
  - (STAC-A2.β2.GREEKS.10-100k-1260.TIME.WARM)
- Less than 1.2% increase in elapsed time for warm and cold runs of the baseline Greeks benchmark
  - STAC-A2.β2.GREEKS.TIME.{WARM/COLD}

www.STACresearch.com/NAIV210520

* rounded to the nearest tenth of a percent

STAC SECURITIES TECHNOLOGY ANALYSIS CENTER

With SEV-ES enabled

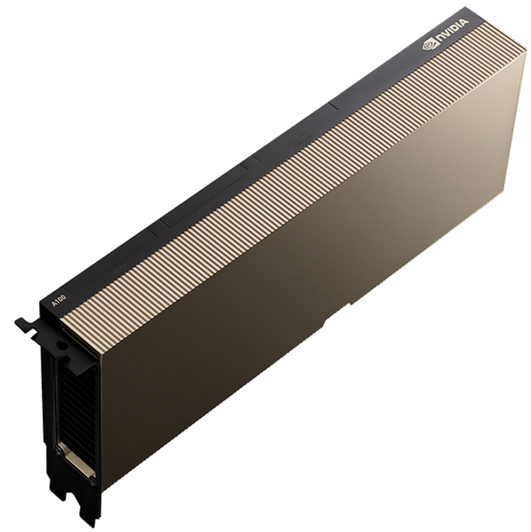
- Less than 1.3% reduction in throughput, energy efficiency, and space efficiency
  - STAC-A2.β2.HPORTFOLIO.SPEED
  - STAC-A2.β2.HPORTFOLIO.ENERGY_EFF
  - STAC-A2.β2.HPORTFOLIO.SPACE_EFF
- No change in quality benchmark results

www.STACresearch.com/NAIV210520

STAC
SECURITIES TECHNOLOGY ANALYSIS CENTER

STAC-A2 / NVIDIA STAC Pack / A100 SXM4 80GB / OpenShift

- STAC-A2 Pack for CUDA, Compatibility Rev G
  - Major update of STAC Pack
- First published fully containerized STAC-A2 SUT
- Stack:
  - NVIDIA CUDA 11.2
  - Red Hat OpenShift 4.8.3 with RHEL CoreOS 48.84
    - NVIDIA GPU Operator for Red Hat OpenShift
  - NVIDIA DGX A100 server
    - 8 x NVIDIA A100 SXM4 80GB GPUs
    - 2 x AMD EPYC 7742 64-core processors @ 2.25 GHz
    - 32 x 64GiB Dual Rank ECC DDR4 DIMMs @ 2933 MT/s

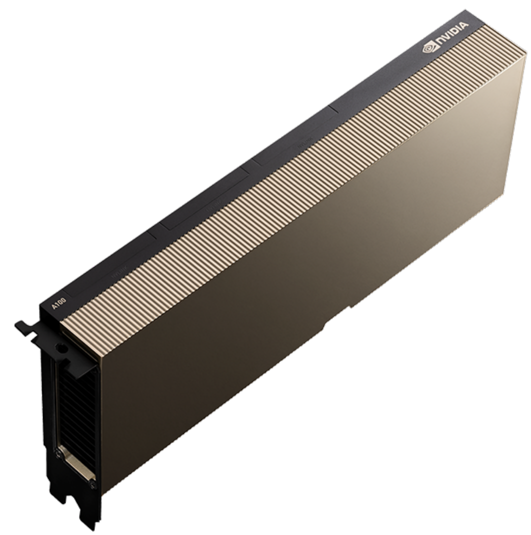


www.STACresearch.com/NVDA210914

STAC® SECURITIES TECHNOLOGY ANALYSIS CENTER

Compared to all publicly reported solutions to date

- Set 8 records:
  - highest energy efficiency (STAC-A2.β2.HPORTFOLIO.ENERG_EFF)
  - highest space efficiency (STAC-A2.β2.HPORTFOLIO.SPACE_EFF)
  - highest throughput (STAC-A2.β2.HPORTFOLIO.SPEED)
  - fastest warm & cold times in the large Greeks benchmark (STAC-A2.β2.GREEKS.TIME.{WARM,COLD})
  - fastest warm time in the baseline Greeks benchmark (STAC-A2.β2.GREEKS.TIME.WARM)
  - highest maximum paths (STAC-A2.β2.GREEKS.MAX_PATHS)
  - highest maximum assets (STAC-A2.β2.GREEKS.MAX_ASSETS)



www.STACresearch.com/NVDA210914

Compared to the best results for any non-NVIDIA-based solution

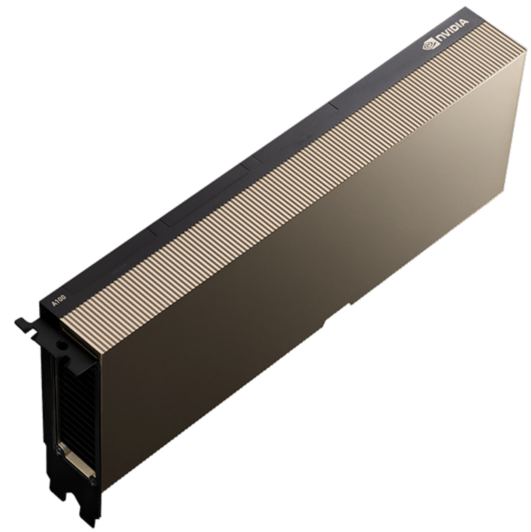
- 3.0x the throughput
  - STAC-A2.β2.HPORTFOLIO.SPEED vs. SUT ID INTC210331
- 2.6x the energy efficiency
  - STAC-A2.β2.HPORTFOLIO.ENERGY_EFF vs. SUT ID INTC210315
- 2.6x faster in warm baseline Greeks benchmark
  - STAC-A2.β2.GREEKS.WARM vs. SUT ID NEC210422
- 2.3x faster in warm large Greeks benchmark
  - STAC-A2.β2.GREEKS.10-100k-1260.TIME.WARM vs. SUT ID INTC181012
- 2.1x the maximum assets
  - STAC-A2.β2.GREEKS.MAX_ASSETS vs. SUT ID INTC181012

www.STACresearch.com/NVDA210914

STAC®
SECURITIES TECHNOLOGY ANALYSIS CENTER

Compared to NVIDIA’s first STAC-A2 audit (NVDA131118)

- 54x faster in the warm baseline Greeks benchmark
  - STAC-A2.β2.GREEKS.WARM
- 24x the maximum paths
  - STAC-A2.β2.GREEKS.MAX_PATHS



www.STACresearch.com/NVDA210914

STAC®
SECURITIES TECHNOLOGY ANALYSIS CENTER