STAC Update:
Big Compute

Peter Nabicht
President, STAC

peter.nabicht@STACresearch.com

STAC AI:
Inference and Training

STAC AI Update: Inference

- Some user firms on the STAC Benchmark Council asked for & provided input to proposed inference benchmarks
- STAC proposed a first version of the specs
- Demonstrated the efficacy of those specs via a POC
  - as described in the talk “Benchmarking realtime LSTM inference on time series”, Global STAC Live, Fall 2020 (https://STACresearch.com/GSL-Fall2020-stac-lstm)

www.STACresearch.com/**ai**

STAC®
SECURITIES TECHNOLOGY ANALYSIS CENTER

STAC AI Update: Inference

- The benchmark seeks to isolate inference performance
  - Inference engine software
  - Underlying processors, memory, etc.
  - Anything required to optimally use the former with the latter (e.g., data transfer to processor memory)
- LSTM model on financial time series data
- Targeting co-location / edge deployment, so SUTs are a single server or subset of a single server
- Measuring throughput and latency of inference across multiple model sizes

www.STACresearch.com/**ai**

STAC®
SECURITIES TECHNOLOGY ANALYSIS CENTER

STAC AI Update: Inference

- Working group is now finalizing specifications for an official STAC benchmark
  - Not too late to participate in the specification process
  - Next working group meeting in early June
- Express interest and join the working group at:

www.STACresearch.com/**ai**

STAC®
SECURITIES TECHNOLOGY ANALYSIS CENTER

STAC AI Update: Training

- User firms asked for & provided input to proposed training benchmarks
- STAC is still creating a proposal for the working group
- Creating a problem that meets all the requirements is … iterative
- Working group will be finalizing specifications this summer
- Interested? Tell us!

www.STACresearch.com/**ai**

STAC®
SECURITIES TECHNOLOGY ANALYSIS CENTER

STAC-A2

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**

STAC®
SECURITIES TECHNOLOGY ANALYSIS CENTER

STAC-A2: Risk computation

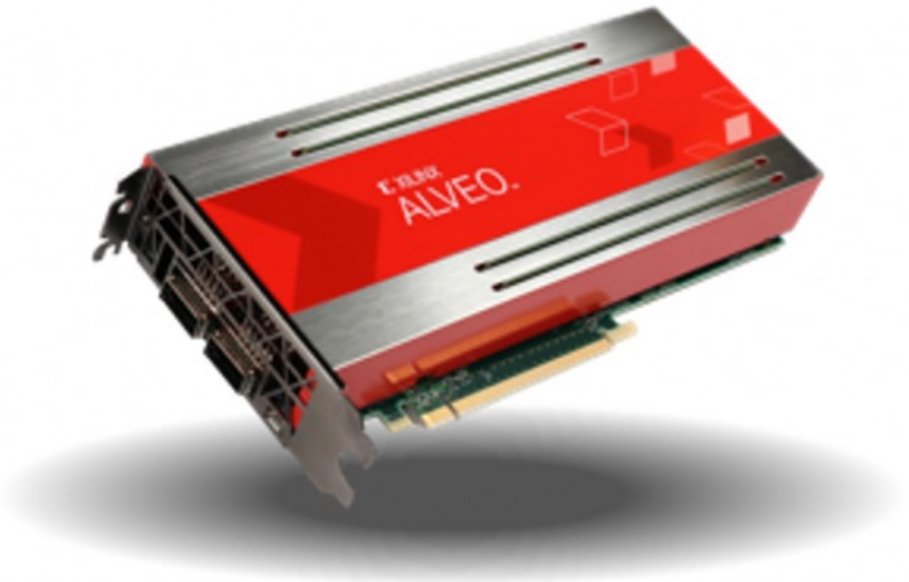
- 2Q21: Four new sets of results came out in rapid succession
- I will describe the significance of each
- Comparisons are as of the time each STAC Report was released

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 / Xilinx Vitis / 8x Xilinx 64GiB Alveo U250

- First STAC-A2 using FPGAs as co-processors
- Implementation written in C++, using software libraries that control the FPGA
- Effectively created a STAC-A2 hardware device
- Stack:
  - STAC-A2 Pack for Xilinx Vitis (Rev B)
  - Xilinx Vitis Unified Software Platform 2020.2
  - Xilinx Vitis Quantitative Finance Library 2020.2
  - Xilinx Runtime 2.5.278
  - 2 x AMD EPYC 7551 32-core CPUs @ 2.0 GHz
  - 8 x Xilinx 64GiB Alveo U250 FPGA cards
  - BOXX GX8-M

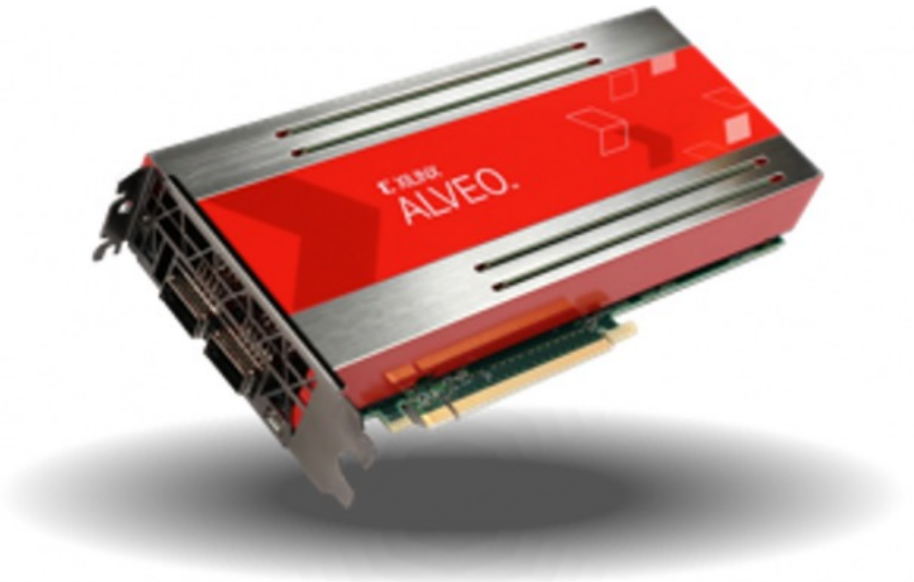


www.STACresearch.com/XLNX210301

STAC®
SECURITIES TECHNOLOGY ANALYSIS CENTER

Versus the most recent CPU-based solution at the time

- 1.48x speed-up in cold run of the baseline problem size
  - (STAC-A2.β2.GREEKS.TIME.COLD vs SUT ID INTC190903)
- 14.1% faster in cold run of the large problem size
  - STAC-A2.β2.GREEKS.10-100k-1260.TIME.COLD vs SUT ID INTC190903
- 20% higher maximum paths
  - STAC-A2.β2.GREEKS.MAX_PATHS vs. SUT ID INTC190903

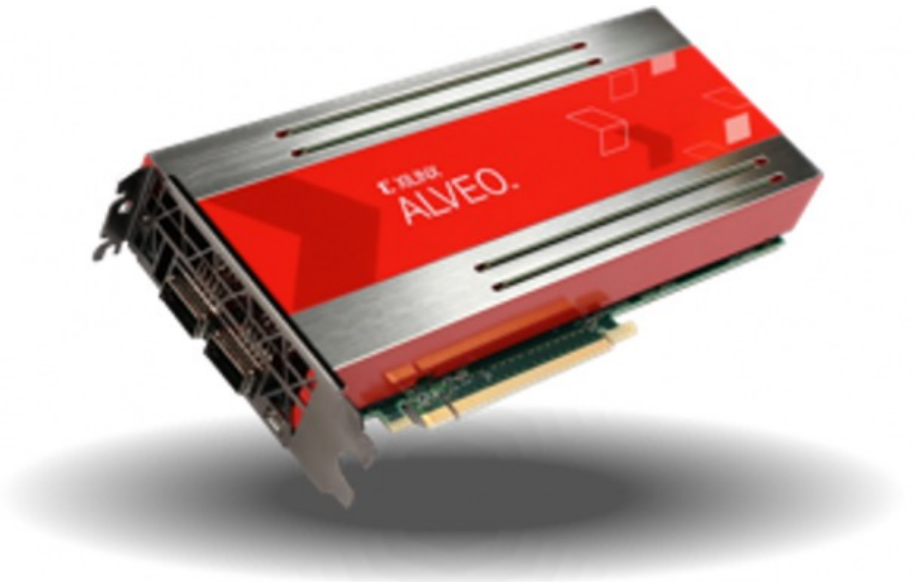


www.STACresearch.com/XLNX210301

STAC®
SECURITIES TECHNOLOGY ANALYSIS CENTER

Verus the most recent GPU-based solution at the time

- Within 11.3% on the cold run of the baseline problem size
  - STAC-A2.β2.GREEKS.TIME.COLD vs. SUT ID NVDA200909
- Within 31.9% on the cold run of the large problem size
  - STAC-A2.β2.GREEKS.10-100k-1260.TIME.COLD vs. SUT ID NVDA200909

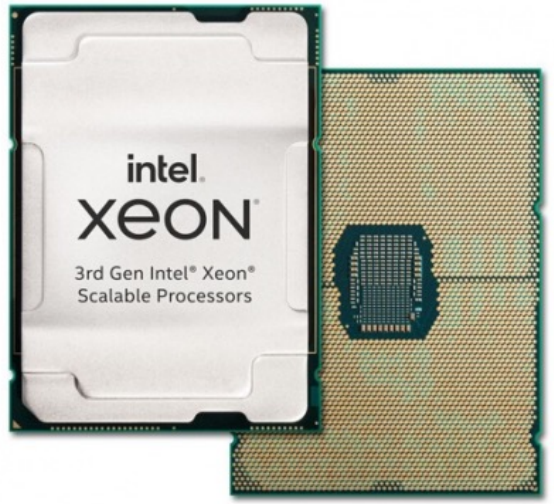


www.STACresearch.com/XLNX210301

STAC®
SECURITIES TECHNOLOGY ANALYSIS CENTER

STAC-A2 / Intel STAC Pack / Intel Ice Lake

- First STAC results on new Ice Lake processors
- Stack:
  - STAC-A2 Pack for Intel oneAPI (Rev N)
  - Intel® oneAPI Base Toolkit 2021.1 Gold
  - Intel® oneAPI HPC Toolkit 2021.1 Gold
  - 2 x Intel® Xeon® Platinum 8380 (Ice Lake) CPU @ 2.30GHz
  - Intel® Server System M50CYP Software Development Platform
  - 16 x 32GiB DDR4 DIMM @ 3200MHz (512GiB total)
  - CentOS Linux 8.3

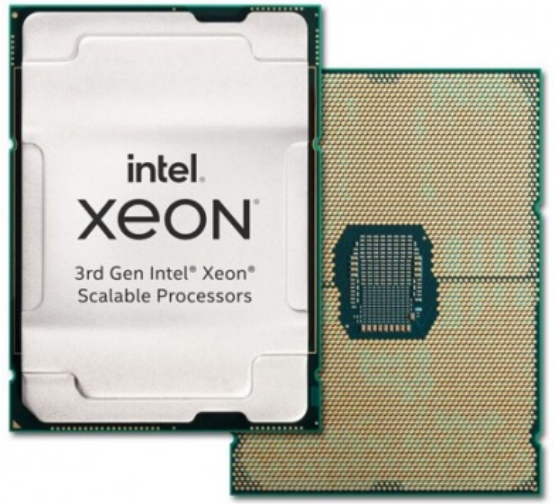


www.STACresearch.com/INTC210315

STAC®
SECURITIES TECHNOLOGY ANALYSIS CENTER

Versus all publicly reported single-server solutions at the time

- Fastest cold time in the large problem size STAC-A2.β2.GREEKS.10-100k-1260.TIME.COLD

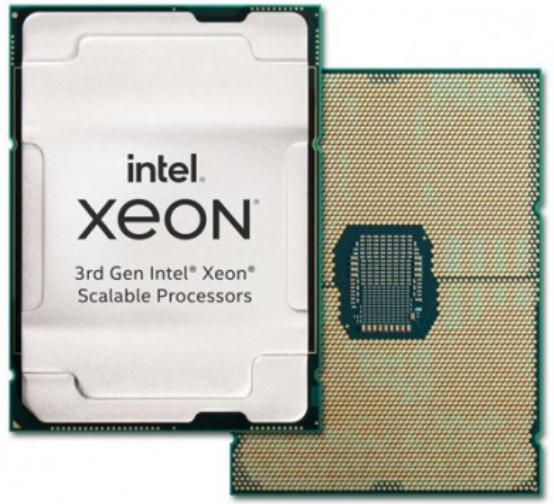


www.STACresearch.com/INTC210315

STAC®
SECURITIES TECHNOLOGY ANALYSIS CENTER

Versus solution with previous Intel generation (Cascade Lake)*

- 4.0x the maximum paths
  - STAC-A2.β2.GREEKS.MAX_PATHS
- > 1.6x the space efficiency
  - STAC-A2.β2.HPORTFOLIO.SPACE_EFF
- > 1.3x the energy efficiency
  - STAC-A2.β2.HPORTFOLIO.ENERG_EFF
- > 1.5x the throughput
  - STAC-A2.β2.HPORTFOLIO.SPEED
- Large problem size:
  - 2.1x in cold runs, 1.5x in warm runs (STAC-A2.β2.GREEKS.10-100k-1260.TIME.COLD/WARM)
- Baseline problem size
  - 1.9x in cold runs, 1.5x in warm runs (STAC-A2.β2.GREEKS.TIME.COLD/WARM)



www.STACresearch.com/INTC210315

** SUT ID INTC190402*

STAC®
SECURITIES TECHNOLOGY ANALYSIS CENTER

STAC-A2 / Intel STAC Pack / Google Cloud / Intel CPUs

- First public cloud solution with publicly released STAC-A2 results
- Used a 10-node cluster in Google Cloud
- Solution configured by AppsBroker
- Used STAC-A2 Pack for Intel OneAPI (Rev N) without code changes
  - Same as Ice Lake SUT discussed above

www.STACresearch.com/INTC210331

STAC®
SECURITIES TECHNOLOGY ANALYSIS CENTER

STAC-A2 / Intel STAC Pack / Google Cloud / Intel CPUs

- Stack:
  - STAC-A2 Pack for Intel oneAPI (Rev N)
  - Intel® oneAPI Base Toolkit 2021.1 Gold
  - Intel® oneAPI HPC Toolkit 2021.1 Gold
  - CentOS Linux release 8.3.2011
  - Google Cloud's VPC Premium Network Tier
  - Google Cloud Compact resource placement policy
  - 10 x Google Cloud Compute Engine "On Demand" c2-standard-60 compute optimized VM instances each with:
    - 60 vCPUs based on 2nd Generation Intel® Xeon® Scalable Processor (Cascade Lake) @ 3.1GHz
    - 240 GiB DRAM
    - 20 GB Boot disk

www.STACresearch.com/INTC210331

STAC®
SECURITIES TECHNOLOGY ANALYSIS CENTER

Compared to all publicly reported solutions at the time

- Highest throughput
  - STAC-A2.β2.HPORTFOLIO.SPEED
- Fastest cold time in the large problem size
  - STAC-A2.β2.GREEKS.10-100k-1260.TIME.COLD

www.STACresearch.com/INTC210331

STAC®
SECURITIES TECHNOLOGY ANALYSIS CENTER

Compared to a solution with 8-node on-premises cluster*

- 5x the maximum paths
  - STAC-A2.β2.GREEKS.MAX_PATHS
- 10% greater throughput
  - STAC-A2.β2.HPORTFOLIO.SPEED
- 18% faster in cold runs of the large problem size
  - STAC-A2.β2.GREEKS.10-100k-1260.TIME
- 9% faster in cold runs of the baseline problem size
  - STAC-A2.β2.GREEKS.TIME.COLD

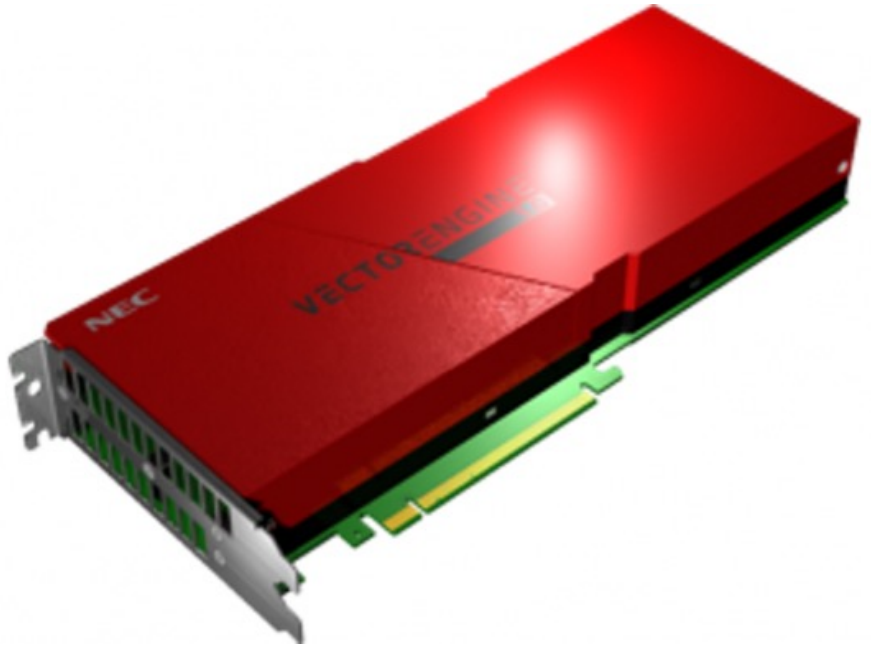
www.STACresearch.com/INTC210331

* SUT ID INTC181012

STAC®
SECURITIES TECHNOLOGY ANALYSIS CENTER

STAC-A2 / NEC STAC Pack / NEC Vector Engine

- First STAC-A2 results using NEC Vector Engines as co-processors
- Stack:
  - STAC-A2 Pack for NEC Vector Engine (Rev A)
  - NEC SDK
  - Red Hat Enterprise Linux 8.2 (Ootpa)
  - NEC VEOS 2.7.5
  - 8 x NEC SX-Aurora TSUBASA/Vector Engine Type 20B
  - 2 x Intel® Xeon® Gold 6226 CPU @ 2.70GHz
  - NEC SX-Aurora TSUBASA B300-8 with 192GiB DRAM

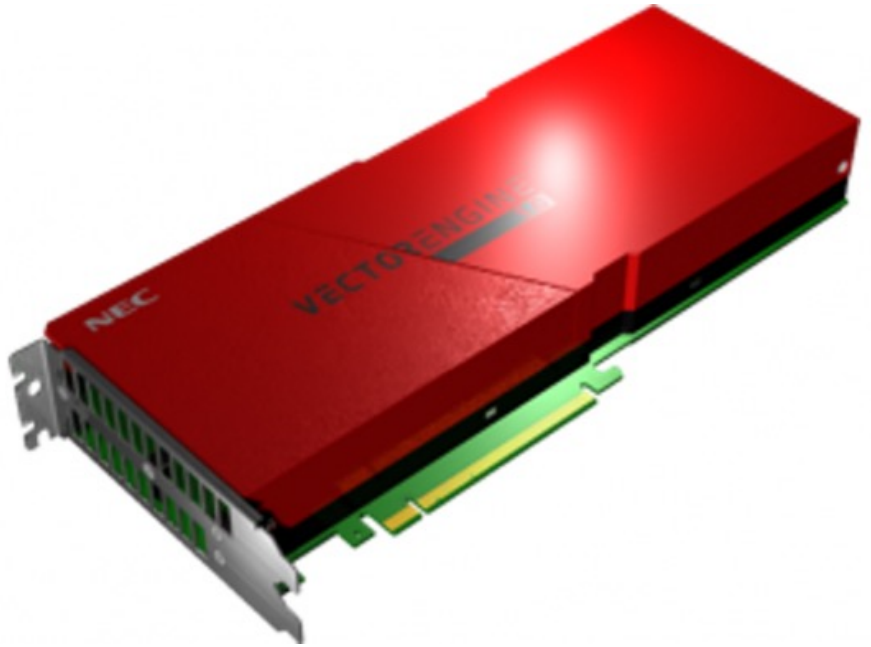


www.STACresearch.com/NEC210422

STAC®
SECURITIES TECHNOLOGY ANALYSIS CENTER

Compared to all publicly reported solutions

- Fastest cold time in the large problem size
  - STAC-A2.β2.GREEKS.10-100k-1260.TIME.COLD

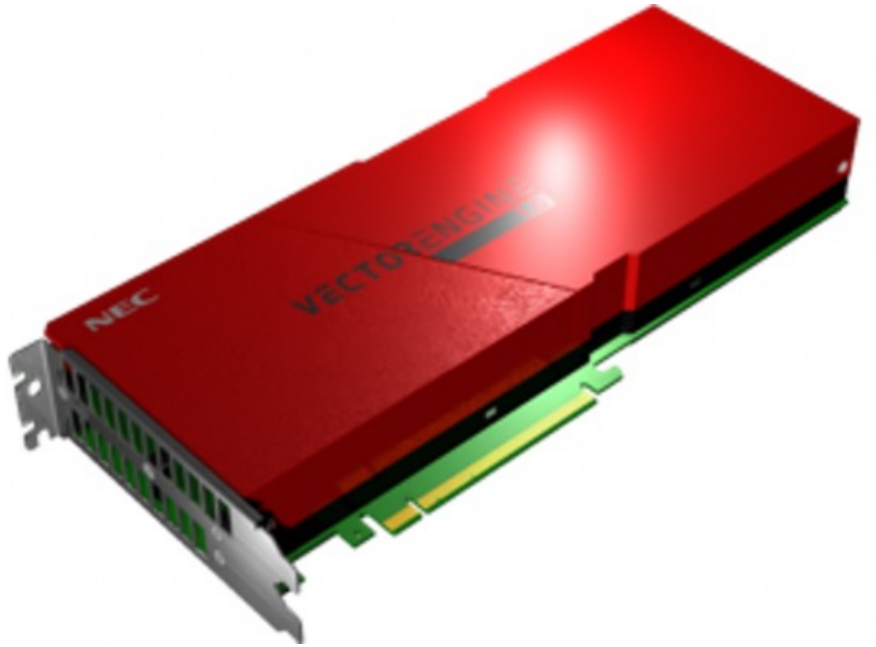


www.STACresearch.com/NEC210422

STAC®
SECURITIES TECHNOLOGY ANALYSIS CENTER

Compared to the previous best for a co-processor based solution

- Versus SUT ID NVDA200909
  - 2.07x speedup in the cold run of the large problem size (STAC-A2.β2.GREEKS.10-100k-1260.TIME.COLD)
  - 1.18x speedup in warm runs for the large problem size (STAC-A2.β2.GREEKS.10-100k-1260.TIME.WARM)
  - Within 10% of the maximum assets (STAC-A2.β2.GREEKS.MAX_ASSETS

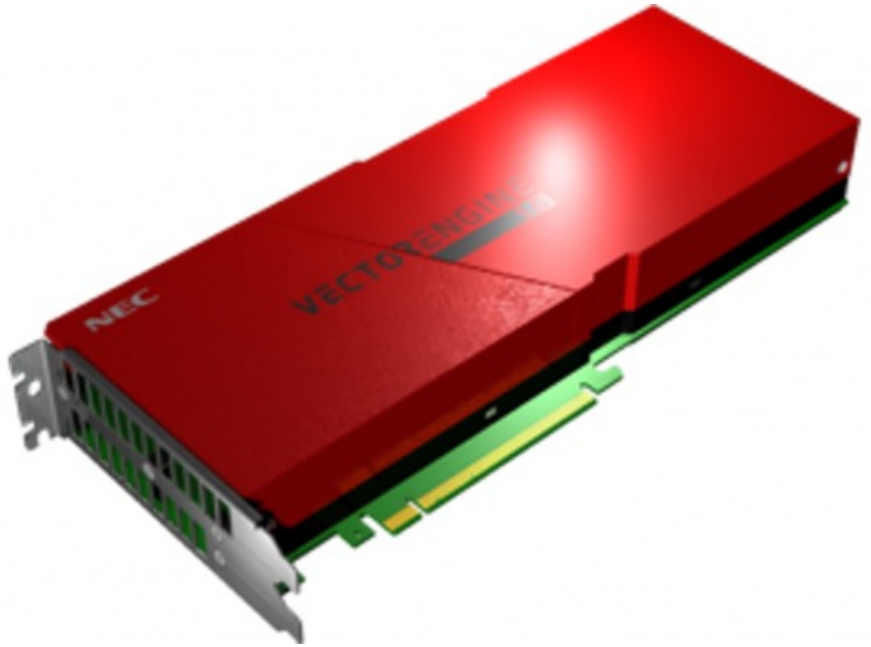


www.STACresearch.com/NEC210422

STAC®
SECURITIES TECHNOLOGY ANALYSIS CENTER

An update to SUT Scaling results

- Due to a configuration error, the initially published report associated SUT Scaling metrics with the wrong SUT scale
- This made the different scaling points look slower than they were
- This has been rectified in v1.1 of the report



www.STACresearch.com/NEC210422

STAC®
SECURITIES TECHNOLOGY ANALYSIS CENTER