Big Compute Update
STAC-ML and STAC-A2

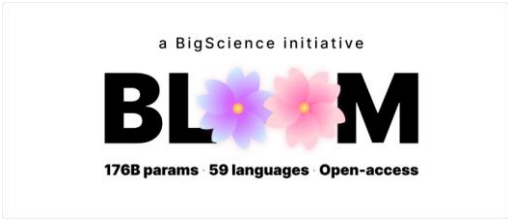
Bishop Brock
Head of Research, STAC
bishop.brock@STACresearch.com

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

Copyright © 2023 Strategic Technology Analysis Center LLC

STAC-ML Updates

- No new STAC-ML Markets (Inference) results to announce today
- But we have been busy with LLMs:
  - Discussions with end-users and vendors
  - Research and prototyping different techniques
  - Analysis of current LLM benchmarks



- We believe that currently available LLM benchmarking is incomplete
- There is no one place to find a comprehensive analysis of
  - Performance, scalability and quality of models across model capabilities and full system stacks
  - Quantization impacts on both representational knowledge and generative quality

Copyright © 2023 Strategic Technology Analysis Center LLC

STAC®
STRATEGIC TECHNOLOGY ANALYSIS CENTER

STAC-ML LLM Directions

- Working Group meeting held October 18
  - Will try for a bi-weekly cadence going forward
- Directions / Questions?
  - Financial workloads
  - Multiple model sizes
  - Generative (vs. simple classification) workloads
  - Workloads that would not be outsourced
  - How to measure Quality?
  - STAC-tuned vs. vendor-tuned models?
  - Dealing with embedded tables?
- **Join the Working Group!**

Example Use Cases	Example Workloads
Standalone Content Generation	General QA Customer Support Bot
Topic Analysis	Email routing Communications Compliance
Single Document Analysis	Interactive Q/A Sentiment Analysis
Multiple Document Retrieval/Analysis	Batch Summarization Interactive Query / Summarization KYC / AML
Code Creation	General Programming Database Queries

www.STACresearch.com/ML

Copyright © 2023 Strategic Technology Analysis Center LLC

STAC®
STRATEGIC TECHNOLOGY ANALYSIS CENTER

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

Copyright © 2023 Strategic Technology Analysis Center LLC

STAC®
STRATEGIC TECHNOLOGY ANALYSIS CENTER

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
- Greeks run times are ~O(Assets^3)

Copyright © 2023 Strategic Technology Analysis Center LLC

STAC®
STRATEGIC TECHNOLOGY ANALYSIS CENTER

STAC-A2 / 2 x Intel® Xeon® Platinum 8480+ (Sapphire Rapids)

- STAC-A2 Pack for Intel® oneAPI (Rev N)
  - Same binary used for Ice Lake audit (INTC210315)
- Stack:
  - Intel® oneAPI Base Toolkit 2023.1 runtime
  - Intel® oneAPI HPC Toolkit 2023.1 runtime
  - Fedora Linux 38 (Server Edition)
  - QuantaGrid D54Q-2U system
    - 2 x Intel® Xeon® Platinum 8480+ (Sapphire Rapids) CPU @ 2.0 GHz
    - 16 x 64GiB DDR5 DIMMs @ 4800MHz (1TiB total)

www.STACresearch.com/INTC230524

Copyright © 2023 Strategic Technology Analysis Center LLC

STAC®
STRATEGIC TECHNOLOGY ANALYSIS CENTER

Compared to a similar Intel® Xeon® 8380 (Ice Lake) System (INTC210315)

- Previous system included:
  - Previous versions of oneAPI toolkits
  - 2 x Intel® Xeon® 8380 (Ice Lake) CPUs @ 2.30 GHz
  - 512 GiB of memory
- Comparisons:
  - 1.43x the throughput[1]
  - 1.53x the speed in warm runs of the large problem size[2]
  - 1.26x the speed in warm runs of the baseline problem size[3]
  - 1.38x the space efficiency[4]
  - 1.10x the energy efficiency[5]
  - 1.13x the number of correlated assets simulated in 10 minutes[6]

1. STAC-A2.β2.HPORTFOLIO.SPEED
2. STAC-A2.β2.GREEKS.10-100k-1260TIME.WARM
3. STAC-A2.β2.GREEKS.TIME.WARM
4. STAC-A2.β2.HPORTFOLIO.SPACE_EFF
5. STAC-A2.β2.HPORTFOLIO.ENERG_EFF
6. STAC-A2.β2.GREEKS.MAX_ASSETS

www.STACresearch.com/INTC230524

Copyright © 2023 Strategic Technology Analysis Center LLC

STAC®
STRATEGIC TECHNOLOGY ANALYSIS CENTER

STAC-A2 / 8 x NVIDIA H100 PCIe 80 GiB GPUs / HPE ProLiant XL675d

- Stack:
  - NVIDIA STAC Pack
  - CUDA 12.0
  - g++ version 7.5
  - SUSE Linux Enterprise Server 15 SP4
  - HPE ProLiant XL675d Gen10 Plus Server
    - 8 x NVIDIA H100 PCIe 80 GiB GPUs
    - 2 x AMD EPYC 7543 processors @ 2.80 GHz
    - 2 TiB DDR4 @ 2933 MT/s

1. STAC-A2.β2.HPORTFOLIO.SPEED
2. STAC-A2.β2.GREEKS.10-100k-1260TIME.WARM
3. STAC-A2.β2.GREEKS.TIME.WARM
4. STAC-A2.β2.HPORTFOLIO.SPACE_EFF
5. STAC-A2.β2.HPORTFOLIO.ENERG_EFF
6. STAC-A2.β2.GREEKS.MAX_ASSETS

www.STACresearch.com/NVDA230721

Copyright © 2023 Strategic Technology Analysis Center LLC

STAC®
STRATEGIC TECHNOLOGY ANALYSIS CENTER

STAC-A2 / 8 x NVIDIA H100 PCIe 80 GiB GPUs / HPE ProLiant XL675d

- Set numerous new records compared to all previous public reports, including (but not limited to):
  - The first sub-10 ms warm time (8.9 ms) in the baseline Greeks benchmark[1]
  - A cold time in the baseline Greeks benchmark (38ms) over 3x faster than any previously reported[2]
  - The fastest warm[3] (0.51 s) and cold[4] (1.85 s) times in the large Greeks benchmarks
  - The most correlated assets[5] (400) and most paths[6] (310,000,000) simulated in 10 minutes
  - The best energy efficiency[7] (311,045 options / kWh)

1, STAC-A2.β2.GREEKS.TIME.WARM
2. STAC-A2.β2.GREEKS.TIME.COLD
3. STAC-A2.β2.GREEKS.10-100k-1260.TIME.WARM
4. STAC-A2.β2.GREEKS.10-100k-1260.TIME.COLD
5. STAC-A2.β2.GREEKS.MAX_ASSETS
6. STAC-A2.β2.GREEKS.MAX_PATHS
7. STAC-A2.β2.HPORTFOLIO.ENERG_EFF

www.STACresearch.com/NVDA230721

Copyright © 2023 Strategic Technology Analysis Center LLC

STAC®
STRATEGIC TECHNOLOGY ANALYSIS CENTER

STAC-A2 / 8 x NVIDIA H100 PCIe 80 GiB vs. 8 x NVIDIA A100 SXM4 80 Gib

- Compared to a SUT using the previous STAC Pack and A100 GPUs (NVDA210914), this solution:
  - Was 10x the speed in the cold run of the baseline Greeks benchmark[1]
  - Was 1.38x the speed in the warm[2] runs of the baseline Greeks benchmark[2]
  - Was 1.38x the speed in the cold run of the large Greeks benchmark[3]
  - Was 1.40x the speed in the warm runs of the large Greeks benchmark[4]
  - Was 10% more energy efficient[5]

1. STAC-A2.β2.GREEKS.TIME.COLD
2. STAC-A2.β2.GREEKS.TIME.WARM
3. STAC-A2.β2.GREEKS.10-100k-1260.TIME.COLD
4. STAC-A2.β2.GREEKS.10-100k-1260.TIME.WARM
5. STAC-A2.β2.HPORTFOLIO.ENERG_EFF

www.STACresearch.com/NVDA230721

Copyright © 2023 Strategic Technology Analysis Center LLC

STAC®
STRATEGIC TECHNOLOGY ANALYSIS CENTER

STAC-A2 / 4 x Intel® Data Center GPU Max 1550 / Dell PowerEdge XE9640

- Firsts:
  - First look at Intel Data Center GPU Max accelerators
  - First direct-liquid-cooled (DLC) SUT audited by STAC!
- STAC-A2 Pack for Intel® oneAPI (Rev O)
- Stack:
  - Intel® oneAPI Base Toolkit 2023.2, DPC++/C++ compiler
  - Intel® oneAPI HPC Toolkit 2023.2, C++ compiler
  - Toolkits include Intel® [oneTBB, oneMKL, MPI]
  - Ubuntu 22.0.3 LTS
  - Dell PowerEdge XE9640 Server
    - 2U form factor, DLC
    - 4 x Intel® Data Center GPU Max 1550
    - 2 x Intel® Xeon® Platinum 8468 CPU @ 2.1 GHz
    - 2 x 16 GiB DDR5 DIMMs @ 4800MHz (32 GiB total)

www.STACresearch.com/INTC230927

Copyright © 2023 Strategic Technology Analysis Center LLC

STAC®
STRATEGIC TECHNOLOGY ANALYSIS CENTER

STAC-A2 / 4 x Intel® Data Center GPU Max 1550 / Dell PowerEdge XE9640

- This SUT set numerous new records compared to all previous public reports, including (but not limited to):
  - The fastest warm[1] (0.405 s) and cold[2] (1.09 s) times in the large problem size benchmarks
  - A space efficiency[3] (238 options / hr. / cu. in.) 2.3x better than the previous best result
  - The best energy efficiency[4] (314,493 options / kWh), 1.0% better than the previous record.

1. STAC-A2.β2.GREEKS.10-100k-1260TIME.WARM
2. STAC-A2.β2.GREEKS.10-100k-1260TIME.COLD
3. STAC-A2.β2.HPORTFOLIO.SPACE_EFF
4. STAC-A2.β2.HPORTFOLIO.SPACE_EFF

www.STACresearch.com/INTC230927

Copyright © 2023 Strategic Technology Analysis Center LLC

STAC®
STRATEGIC TECHNOLOGY ANALYSIS CENTER

STAC-A2 / 4 x Intel® Data Center GPU Max 1550 vs. 8 x GPUs

- Comparison SUT included 8 GPUs (NVDA230721)
  - 78% of the throughput[1]
  - 98% of the speed in warm runs in the baseline problem size benchmarks[2]
  - 1.7x the speed in cold runs of the large problem size benchmark[3]
  - 1.2x the speed in warm runs of the large problem size benchmark[4]
  - 4.3x the space efficiency[5]

1 STAC-A2.β2.HPORTFOLIO.SPEED
2 STAC-A2.β2.GREEKS.TIME.WARM
3 STAC-A2.β2.GREEKS.10-100k-1260TIME.COLD
4 STAC-A2.β2.GREEKS.10-100k-1260TIME.WARM
5 STAC-A2.β2.HPORTFOLIO.SPACE_EFF

www.STACresearch.com/INTC230927

Copyright © 2023 Strategic Technology Analysis Center LLC

STAC®
STRATEGIC TECHNOLOGY ANALYSIS CENTER

STAC-A2 / 4 x Intel® Data Center GPU Max 1550 vs. 2 x Sapphire Rapids

- Comparison SUT (INTC230524) included:
  - Previous version of the Intel® STAC Pack (Rev N)
  - 2 x Intel® Xeon® Platinum 8480+ (Sapphire Rapids) CPUs
- Comparisons:
  - 4.0x and 2.0x better performance in warm[1] and cold[2] runs (respectively) of the baseline problem size benchmarks
  - 6.9x and 4.4x better performance in warm[3] and cold[4] runs (respectively) of the large problem size benchmarks
  - 7.9x better throughput[5]
  - 7.0x improvement in space efficiency[6] and 2.7x improvement in energy efficiency[7]

1. STAC-A2.β2.GREEKS.TIME.WARM
2. STAC-A2.β2.GREEKS.TIME.COLD
3. STAC-A2.β2.GREEKS.10-100k-1260.TIME.WARM
4. STAC-A2.β2.GREEKS.10-100k-1260.TIME.COLD
5. STAC-A2.β2.HPORTFOLIO.SPEED
6. STAC-A2.β2.HPORTFOLIO.SPACE_EFF
7. STAC-A2.β2.HPORTFOLIO.ENERG_EFF

www.STACresearch.com/INTC230927

Copyright © 2023 Strategic Technology Analysis Center LLC

STAC®
STRATEGIC TECHNOLOGY ANALYSIS CENTER