Performance Analysis of SpMV Hardware Execution Choice

Max Hawkins

mhawkins60@gatech.edu

hpcgarage

Georgia Tech

CPG Computational Physics Group

Scientific Dependence on SpMV: Algebraic Multigrid Example

[1] FY. Lu et al., "AmgT: Algebraic Multigrid Solver on Tensor Cores," in SC24: International Conference for High Performance Computing, Networking, Storage and Analysis, Atlanta, GA, USA, 2024, pp. 1-16, doi: 10.1109/SC41406.2024.00058.

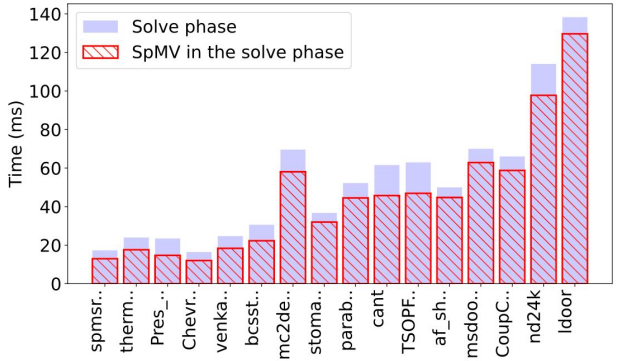
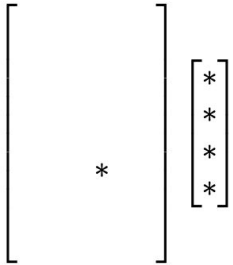


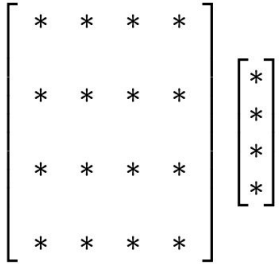
Figure taken from [1]

Sparse



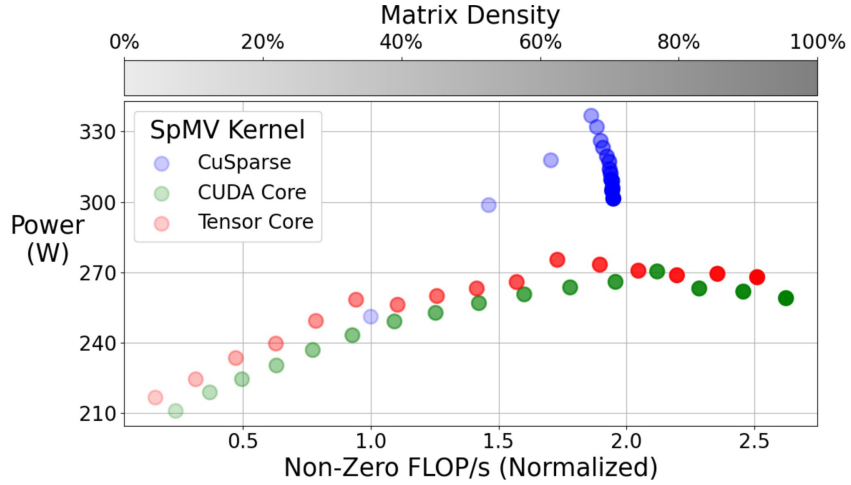
ALU-Like HW
(CUDA Core)

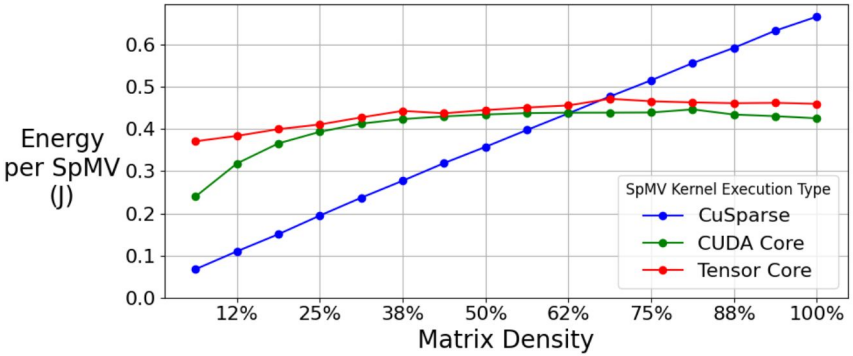
Dense



Dense HW Accelerator
(Tensor Core)

Where is the crossover?





'CUDA Core' & 'Tensor Core' are non-load-balanced SpMV kernels from AmgT. 'CuSparse' is a direct call to NVIDIA's CuSparse library.

Current/Future Work

- Custom microbenchmarking kernels
- Varied data types and vendors
- Structured Sparsity
- Implications in LLMs
  - Sparse Attention
  - Pruning
  - Quantization-induced sparsity
- Sparsity in natural, intelligent systems
- Novel architectures (e.g. neuromorphic)

Let's Chat!

mhawkins60@gatech.edu
linkedin.com/in/maxh/

+ More data & insights on my poster

(including kernel details from NVIDIA Nsight Compute)