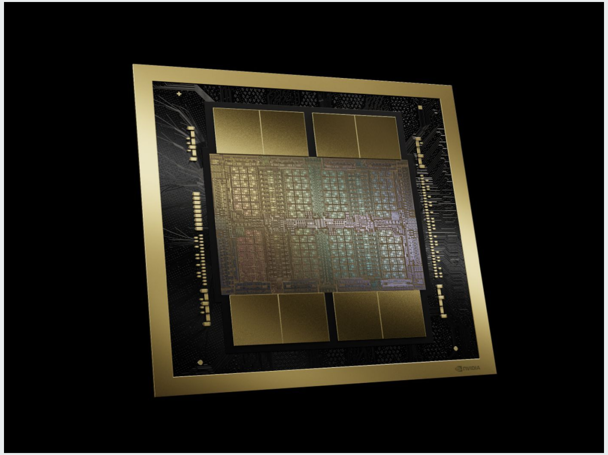
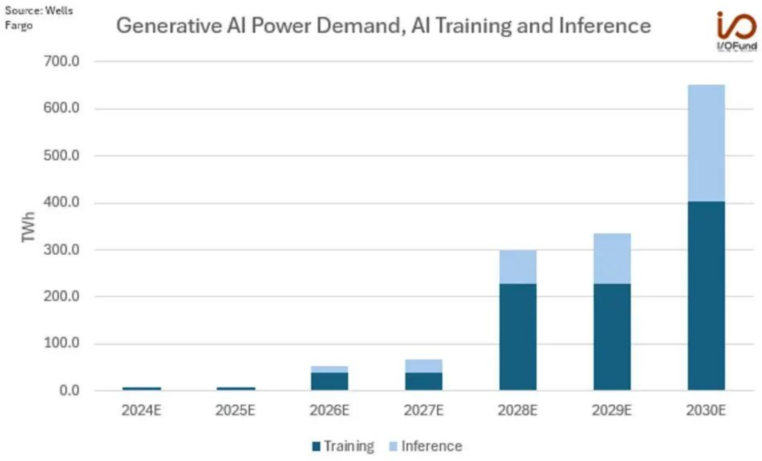
GPU Sparse Matrix Multiplication and Power Profiling

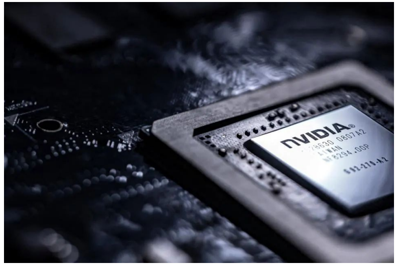
Abdul Muizz

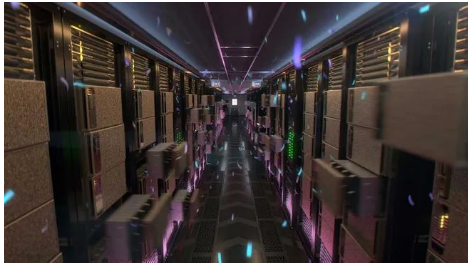


Inspiration for this project

- The recent boom in AI training and accelerators has caused an increased demand for power.
- While new-gen chips are more power efficient, they being used in computation-heavy workloads, increasing power draw
- Training GPT-3 took an estimated 1,300 megawatt hours of electricity (as much power as consumed annually by 130 homes).
- A single modern AI GPU consumes up to 3.7 MWh per year.
- The upcoming Blackwell B200 consumes up to 1,200 Watts.
- **Can we benchmark GPU performance for a given matrix operation?**

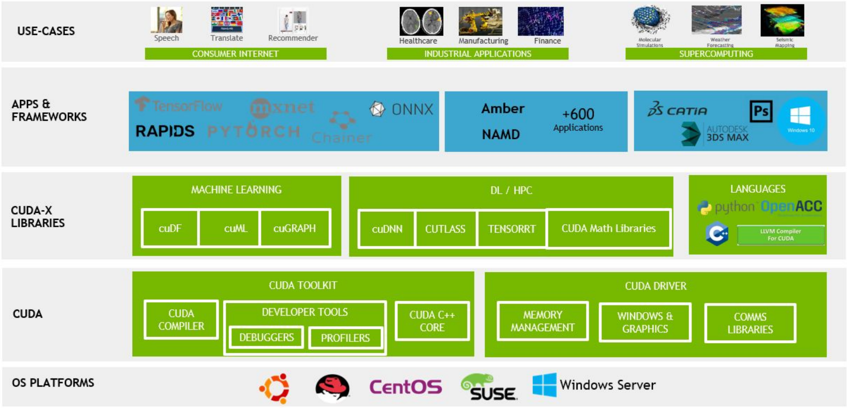






Inspiration for this project

- First opportunity to use NVidia's CUDA Development Toolkit.
- Specifically, we can use **nvidia-smi** (System Management Interface) to **monitor** NVidia devices
- **cuSPARSE library** for optimized sparse operations
- https://developer.nvidia.com/cuda-toolkit



CUDA Toolkit

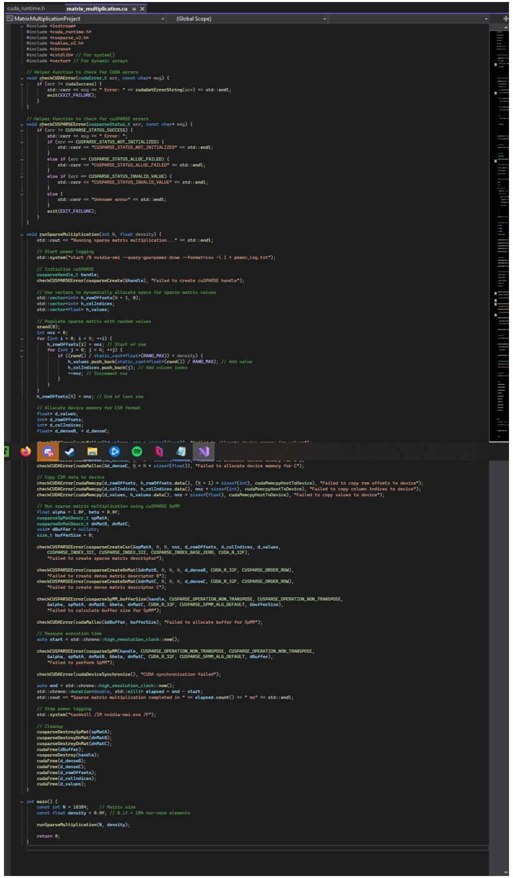
The NVIDIA® CUDA® Toolkit provides a development environment for creating high-performance, GPU-accelerated applications. With it, you can develop, optimize, and deploy your applications on GPU-accelerated embedded systems, desktop workstations, enterprise data centers, cloud-based platforms, and supercomputers. The toolkit includes GPU-accelerated libraries, debugging and optimization tools, a C/C++ compiler, and a runtime library.

Download Now

Project Overview

Key features of matrix_multiplication.cu (CUDA File)

- Generate a random sparse matrix with user-defined density and dimension.
    - Dimension “N” =1024, 2048, 4096, 8192, 16384
    - Density = 10%, 50%, 90% (Percent of non-zero elements)
- Log power at regular intervals to power_log.txt (Average these for our average power)
- Record execution time to console



Project Hardware

- This project is being run on my PC
- RTX 3070-Ti (MSI SUPRIM X)
- Idles around 87W
- 310W max power consumption
- 8GB GDDR6X VRAM
- 1860 MHz clockspeed
- This GPU is being paired with an Intel i7-12700k, 850W PSU

Results Outline

For each density and dimension (N) combination, the following is record

- Execution time (milliseconds)
- Average GPU Power (W)

These values are used to compute

- Power efficiency (GFLOPs/W)
- watt-hours (Wh)
- milliwatt-hours (mWh)

Power Efficiency is found by taking the number of non-zero elements,

$$nnz = density \times N^2$$

And multiplying it by 2N to get FLOPS,

$$FLOPs = 2 \times N \times nnz$$

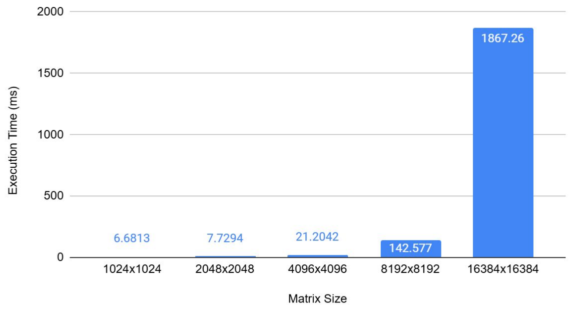
And then computing power efficiency,

$$Power\ Efficiency = \frac{FLOPs}{Execution\ Time \times Average\ Power}$$

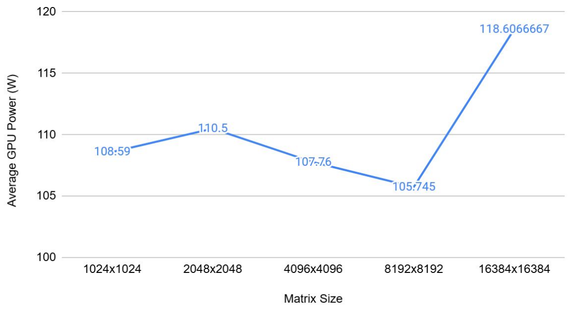
10% Density Results

Density 0.1	// 10% non-zero elements				
Matrix Size	Execution Time (ms)	Average GPU Power (W)	Power Efficiency (GFLOPs/W)	Watt-hours	Milliwatt-Hours
1024x1024	6.6813	108.59	0.295991	0.00020153	0.20153
2048x2048	7.7294	110.5	2.011462	0.00023725	0.23725
4096x4096	21.2042	107.76	6.014927	0.0006347	0.6347
8192x8192	142.577	105.745	7.292736	0.004188	4.188
16384x16384	1867.26	118.6066667	3.971695	0.06152	61.52

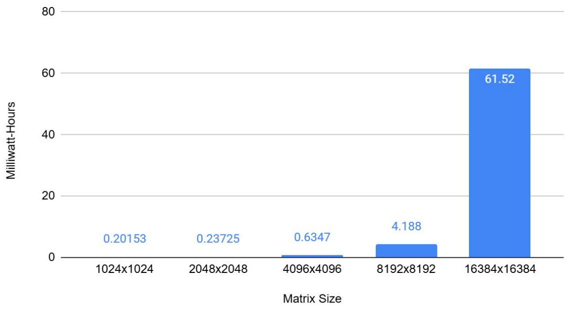
Execution Time (ms) vs. Matrix Size [10% Density]



Average GPU Power (W) vs. Matrix Size [10% Density]

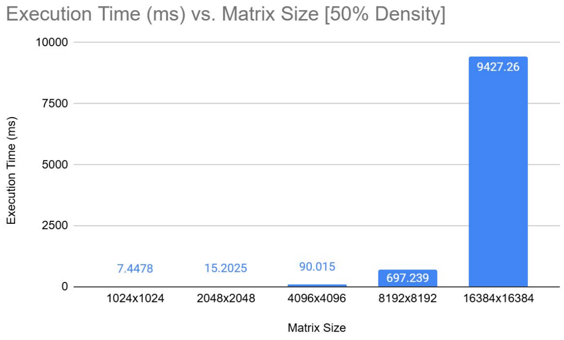


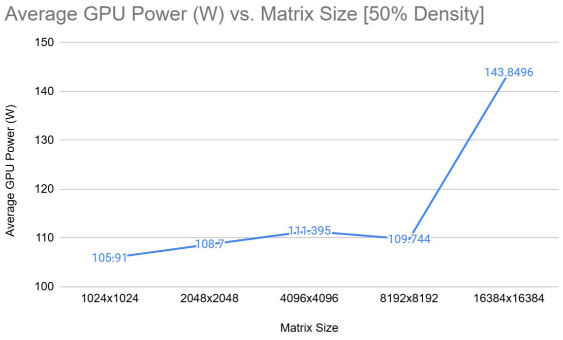
Milliwatt-Hours vs. Matrix Size [10% Density]

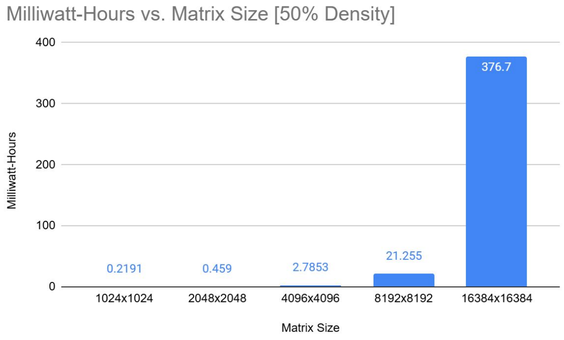


50% Density Results

Density 0.5	// 50% non-zero elements				
Matrix Size	Execution Time (ms)	Average GPU Power (W)	Power Efficiency (GFLOPs/W)	Watt-hours	Milliwatt-Hours
1024x1024	7.4478	105.91	1.361241	0.0002191	0.2191
2048x2048	15.2025	108.7	5.198108	0.000459	0.459
4096x4096	90.015	111.395	6.853292	0.0027853	2.7853
8192x8192	697.239	109.744	7.184679	0.021255	21.255
16384x16384	9427.26	143.8496	3.24314	0.3767	376.7

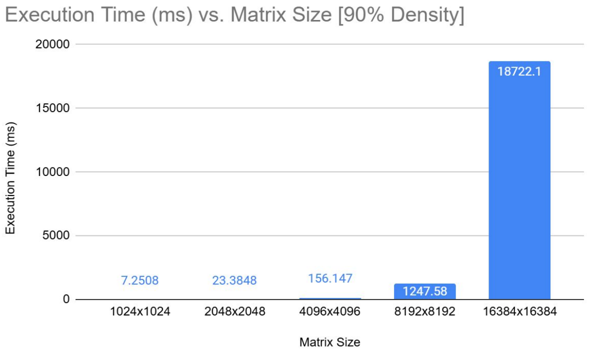


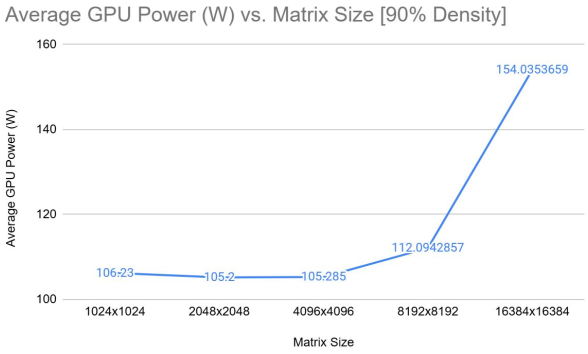


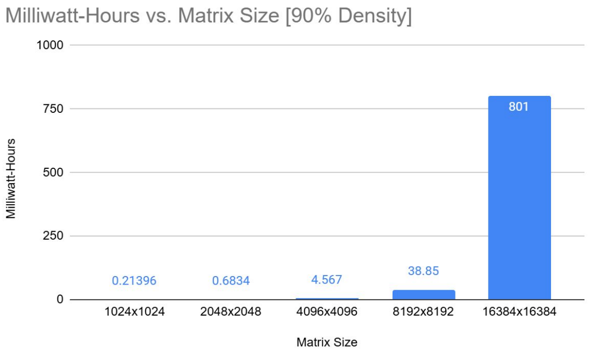


90% Density Results

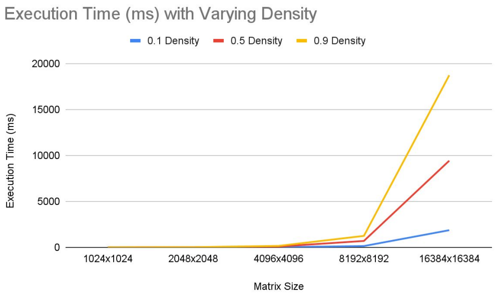
Density 0.9	// 90% non-zero elements				
Matrix Size	Execution Time (ms)	Average GPU Power (W)	Power Efficiency (GFLOPs/W)	Watt-hours	Milliwatt-Hours
1024x1024	7.2508	106.23	2.509223	0.00021396	0.21396
2048x2048	23.3848	105.2	6.285111	0.0006834	0.6834
4096x4096	156.147	105.285	7.524059	0.004567	4.567
8192x8192	1247.58	112.0942857	7.076043	0.03885	38.85
16384x16384	18722.1	154.0353659	2.745095	0.801	801

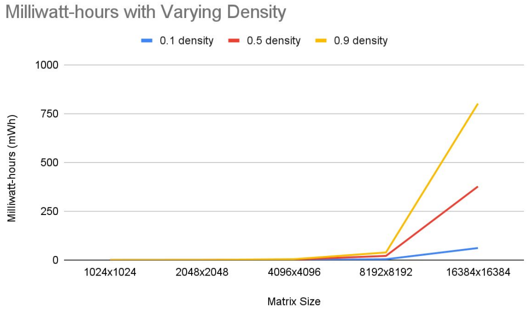


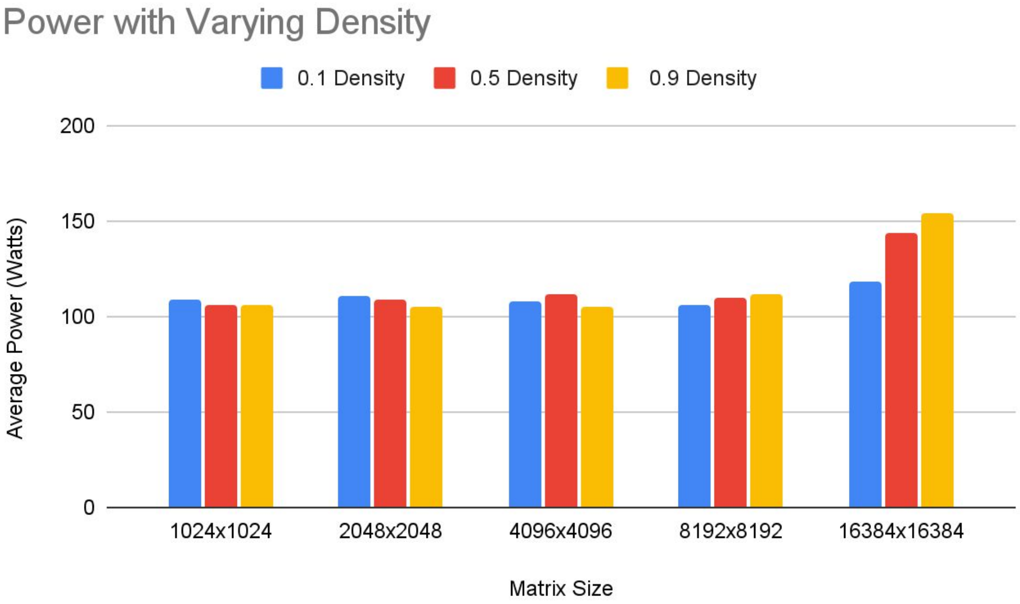




Varied Density Results

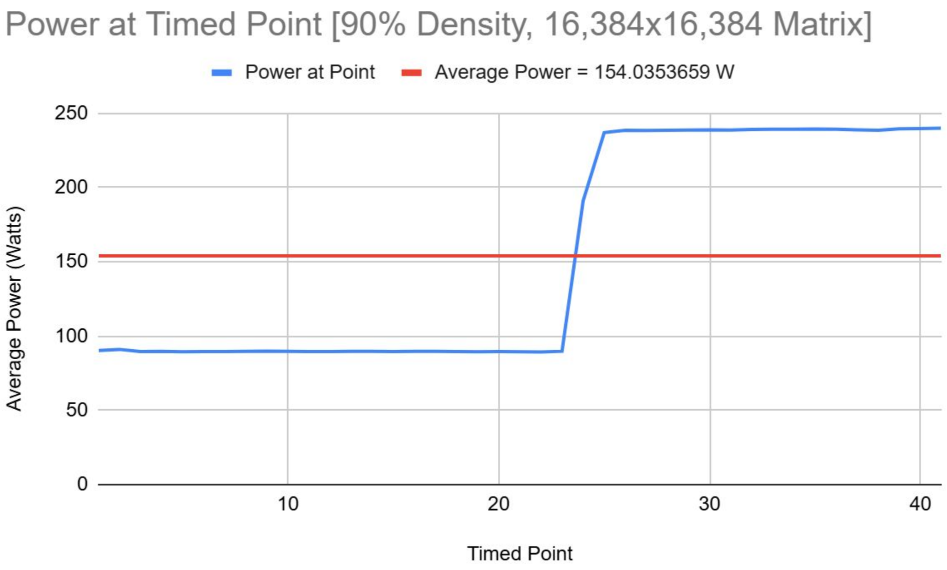






Power Draw over a single Operation

- Run the 16,384x16,384 multiplication with 90% density
- Look at the output of power_log.txt to see the captured power level over fixed time intervals.



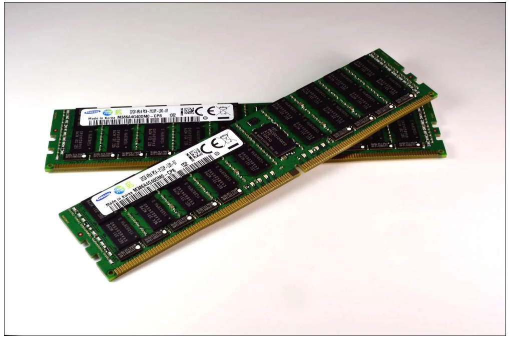
	Power (Watts)
1	90.27
2	91.03
3	89.53
4	89.7
5	89.43
6	89.54
7	89.5
8	89.66
9	89.82
10	89.65
11	89.63
12	89.63
13	89.75
14	89.68
15	89.61
16	89.75
17	89.69
18	89.54
19	89.44
20	89.53
21	89.49
22	89.35
23	89.75
24	191.01
25	236.96
26	238.54
27	238.35
28	238.49
29	238.66
30	238.88
31	238.67
32	239.13
33	239.27
34	239.26
35	239.41
36	239.24
37	238.86
38	238.53
39	239.58
40	239.72
41	239.92

Key Results

- As matrix size increases, time to complete multiplication increases.
- For a given matrix dimension, as density increases, time to complete increases.
- Thus, milliwatt-hours for an operation increase proportionally to dimension and density.
- Power consumption is generally pretty equal, except for the largest matrix dimension.
- For more longer, more complex operations, you can measure a certain power spike.

Moving Forward

- Explore other areas of power consumption, namely retrieving data from memory (Major area of power draw in AI Model Training)
- Analyze when a spike triggers, or what triggers it
- Optimal to get a better benchmarking setup, currently the GPU is also running background tasks (Running my computer)



Any Questions?

GPU Sparse Matrix Multiplication and Power Profiling

Abdul Muizz

Works Cited

Kindig, Beth. “AI Power Consumption: Rapidly Becoming Mission-Critical.” *Forbes*, 20 June 2024, www.forbes.com/sites/bethkindig/2024/06/20/ai-power-consumption-rapidly-becoming-mission-critical/.

Morales, Jowi. “A Single Modern AI GPU Consumes up to 3.7 MWh of Power per Year — GPUs Sold Last Year Alone Consumed More Power than 1.3 Million Homes.” *Tom’s Hardware*, 14 June 2024, www.tomshardware.com/desktops/servers/a-single-modern-ai-gpu-consumes-up-to-37-mwh-of-power-per-year-gpus-sold-last-year-alone-consume-more-power-than-13-million-households.

“MSI GeForce RTX 3070 Ti SUPRIM X 8G.” *Msi.com*, 2018, www.msi.com/Graphics-Card/GeForce-RTX-3070-Ti-SUPRIM-X-8G/Specification. Accessed 4 Dec. 2024.

NVIDIA. “CUDA Toolkit.” *NVIDIA Developer*, 2 July 2013, developer.nvidia.com/cuda-toolkit.

Vincent, James. “How Much Electricity Does AI Consume?” *The Verge*, 16 Feb. 2024, www.theverge.com/24066646/ai-electricity-energy-watts-generative-consumption.