CUDA — Laboratorium 03

Sprawozdanie

Mateusz Łopaciński

23 czerwca 2025

1 Cel

Porównać wydajność dwóch implementacji mnożenia macierzy na GPU Tesla T4:

1. **Simplest** — naiwny kernel bez pamięci współdzielonej.
2. **CUDA Samples** — kernel z tilingiem i shared memory.

Pomiar wykonano dla $N = 256, 512, 1024, 2048$.

2 Środowisko

Google Colab, CUDA 12.5, GPU Tesla T4 (sm_75), `nvcc -std=c++17 -O3`.

3 Wyniki

N	Simplest		Samples	
	czas [s]	GFLOPS	czas [s]	GFLOPS
256	0.185	0.18	0.156	0.22
512	0.433	0.62	0.243	1.10
1024	0.942	2.28	0.103	20.8
2048	17.35	0.99	0.110	156

4 Komentarz

Simplest: wydajność spada dla dużych N z powodu nadmiaru dostępów do pamięci globalnej.

Samples: dzięki tilingowi czas praktycznie nie rośnie od $N = 1024$, a GFLOPS sięga 150 (ok. 2% peak T4).

5 Wnioski

- Shared memory poprawia wydajność do 150× przy $N = 2048$.
- Naiwny kernel służy głównie celom dydaktycznym.