Матрицы размерностью N делились на подматрицы размером 32х32 и такими блоками загружались в блоки нитей.

Две матрицы умножаются вместе, а далее блочная подматрица С записывается в глобальную память. При этом каждый поток записывает один элемент подматрицы.

Таблица результатов скорости умножения:

|  |  |  |  |
| --- | --- | --- | --- |
| Размерность | Время на GPU (мс) | Время на CPU (мс) | Ускорение |
| 320 | 5.32 | 310 | 58.27 |
| 640 | 39.7 | 3356 | 84.53 |
| 1024 | 151.8 | 14178 | 93.39 |
| 1600 | 505.04 | 57170 | 113.21 |
| 1920 | 848.613 | 100219 | 118.09 |