Министерство науки и высшего образования Российской Федерации
Федеральное государственное бюджетное образовательное учреждение
высшего образования
«Московский государственный технический университет имени Н.Э. Баумана
(национальный исследовательский университет)»
(МГТУ им. Н.Э. Баумана)

ФАКУЛЬТЕТ **ИНФОРМАТИКА И СИСТЕМЫ УПРАВЛЕНИЯ**

КАФЕДРА **«ПРОГРАММНОЕ ОБЕСПЕЧЕНИЕ ЭВМ И ИНФОРМАЦИОННЫЕ ТЕХНОЛОГИИ» (ИУ7)**

Л А Б О Р А Т О Р Н А Я Р А Б О Т А № 1

Дисциплина: Программирование специализированных вычислительных устройств

Студент	ИУ7-41М		В.Д. Коноваленко
	(Группа)	(Подпись, дата)	(И.О. Фамилия)
Преподаватель			А.П. Ковтушенко
		(Подпись, дата)	(И.О. Фамилия)

Москва, 2025

Постановка задачи

Исследовать зависимость времени работы процесса умножения матриц с помощью технологии CUDA от размера матриц, соотношения их сторон и расположения в памяти.

Для перемножения рассматривается 4 варианта расположения матриц в памяти (зависит от транспонирования матриц):

0. Обе матрицы не транспонированы;
1. Транспонирована первая матрица;
2. Транспонирована вторая матрица;
3. Транспонированы обе матрицы.

Первая матрица имеет размерность m x n, вторая – n x p.

В ходе исследования замеряются следующие величины:

- CPU (ms) – время выполнения перемножения матриц на процессоре;
- GPU Total (ms) – время перемножения матриц с учётом времени на выделение памяти для их хранения, передачи матриц с хоста на устройство и запись результата графическим процессором;
- GPU Clean (ms) – время перемножения матриц и записи результата графическим процессором;

Результаты

Были проведены замеры времени перемножения квадратных матриц с разными размерностями: 100, 200, 400, 800, 1600 и 3200. Результаты представлены в таблице ниже:

	n	m	p	CPU, мс	GPU Clean, мс	GPU Total, мс
Расположение 0	100	100	100	5,2546	0,5484	0,8315
	200	200	200	41,7273	4,6404	5,1801
	400	400	400	362,5425	32,9302	34,6172
	800	800	800	3334,9024	234,7671	239,0220
	1600	1600	1600	41767,0490	1876,7070	1890,5756
	3200	3200	3200	341395,0434	15015,3232	15067,7109
Расположение 1	100	100	100	5,2960	0,6339	0,8948
	200	200	200	41,4609	5,1089	5,6134

	400	400	400	364,6656	34,8592	36,5102
	800	800	800	3321,6715	251,0295	255,3349
	1600	1600	1600	41418,8302	2006,3694	2020,0931
	3200	3200	3200	345444,8734	16058,8389	15047,7783
Расположение 2	100	100	100	5,2908	1,5957	1,8553
	200	200	200	41,0972	11,3328	11,8438
	400	400	400	364,6584	84,4583	86,16
	800	800	800	3276,0019	660,6942	664,9536
	1600	1600	1600	41352,9704	5309,8462	5323,5024
	3200	3200	3200	346505,9908	42833,5312	42885,4375
Расположение 3	100	100	100	5,2256	1,6156	1,8836
	200	200	200	41,2233	11,5932	12,0938
	400	400	400	366,252	86,6108	88,2654
	800	800	800	3282,5318	671,4696	675,6373
	1600	1600	1600	42511,1474	5369,4824	5383,2446
	3200	3200	3200	369005,9058	42891,4375	42949,0312

Результаты измерений для прямоугольных матриц с размерностями 100x50x25, 400x200x100, 800x400x200, 1600x800x400, 3200x1600x800 представлены в таблице ниже:

	n	**m**	**p**	**CPU, мс**	**GPU Clean, мс**	**GPU Total, мс**
Расположение 0	100	50	25	0,6478	0,0905	0,4004
	200	100	50	8,5253	0,6281	1,1138
	400	200	100	41,3265	5,5241	6,1702
	800	400	200	363,6747	35,6248	37,0999
	1600	800	400	3228,0014	252,857	256,8089
	3200	1600	800	40602,7164	1873,8878	1886,5929
Расположение 1	100	50	25	0,6782	0,1076	0,4118
	200	100	50	5,1071	0,8716	1,266
	400	200	100	41,2633	5,8301	6,4704
	800	400	200	360,7034	38,0156	39,4856
	1600	800	400	3271,9791	269,7343	273,6806
	3200	1600	800	41079,1517	2004,8848	2017,7054
Расположение 2	100	50	25	0,6443	0,2148	0,5245
	200	100	50	5,1711	1,585	2,0082
	400	200	100	41,5071	11,446	12,1032
	800	400	200	361,2378	84,2112	85,6764
	1600	800	400	3307,9814	664,9275	669,6256
	3200	1600	800	41184,4041	5355,6992	5368,5527
Расположение 3	100	50	25	0,6883	0,2161	0,5686
	200	100	50	5,1745	1,6163	2,0034
	400	200	100	43,2762	12,2496	12,9475
	800	400	200	363,2811	87,5933	89,0571
	1600	800	400	3268,3682	673,8858	677,9208
	3200	1600	800	41036,338	5360,7939	5375,397

Зависимость времени выполнения от размерности матриц

На рисунке 1 и 2 приведены графики зависимости времени выполнения (CPU и GPU Total) перемножения матриц в зависимости от размерности матриц для квадратных и прямоугольных матриц с вариантом расположения 0.

Рисунок 1. Графики зависимости времени CPU и GPU Total от размерности квадратных матриц.

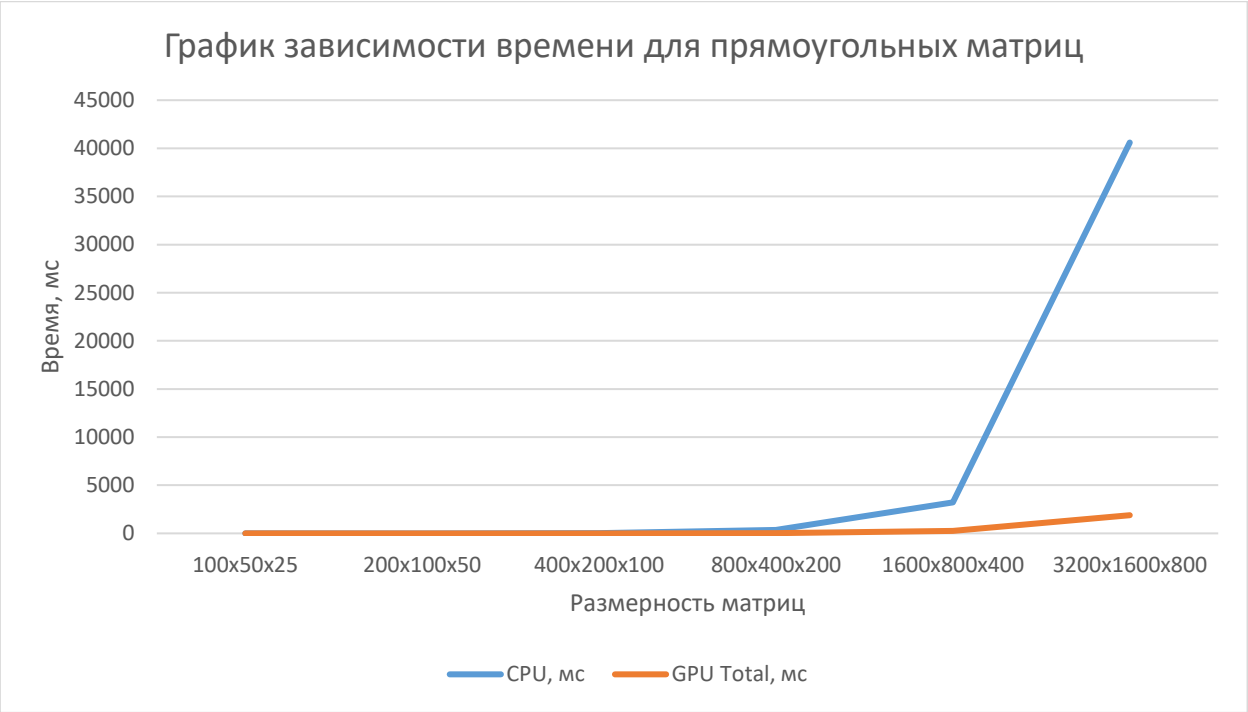

Рисунок 2. График зависимости времени CPU и GPU Total от размерности прямоугольных матриц.

На основе представленных графиков можно сделать вывод о том, что при больших размерах перемножаемых матриц время выполнения вычислений на CPU значительно превышает время выполнения на GPU. В то же время Данная зависимость прослеживается и для квадратных, и для прямоугольных матриц.

Таким образом, применение GPU для вычисления операций над матрицами особенно актуально на больших размерностях, так как в данном случае накладные расходы становятся несущественными по сравнению с выигрышем в скорости выполнения операций.

Зависимость времени выполнения от расположения матриц

На рисунках 3 и 4 приведены графики зависимости времени выполнения перемножения (GPU Clean) от размерностей квадратных и прямоугольных матриц для различных расположений матриц.

Рисунок 3. График зависимости времени GPU Clean от размерности квадратных матриц для каждого варианта расположения.

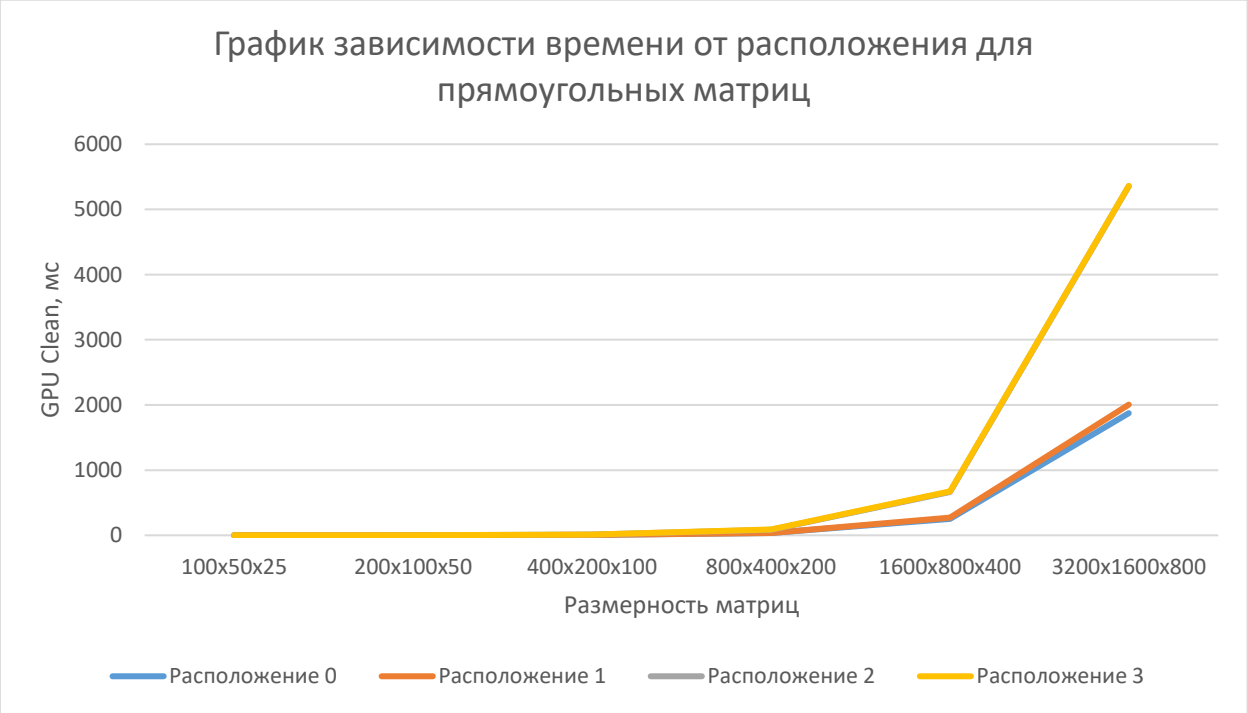

Рисунок 4. График зависимости времени GPU Clean от размерности прямоугольных матриц для каждого варианта расположения.

На основе приведённых графиков можно сделать вывод о том, что наиболее лучшие показатели времени выполнения наблюдаются при варианте расположения 0 (обе матрицы не транспонированы), на втором месте вариант расположения 1 (транспонирована первая матрица), худшие показатели у вариантов расположения 2 (транспонирована вторая матрица) и 3 (транспонированы обе матрицы).

Вывод

На основе проведённого исследования можно сформировать следующие выводы:

- Вычисление произведения матриц на GPU происходит быстрее, чем на CPU. Чем больше размерность матрицы, тем большая разница в скорости выполнения между этими вариантами.
- Наиболее эффективным из вариантов расположения матриц в памяти является обычный, когда обе матрицы не транспонированы.