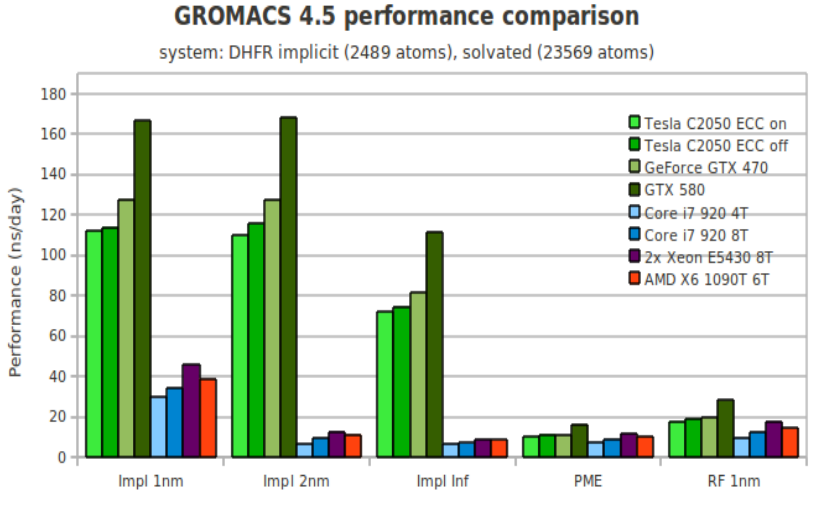
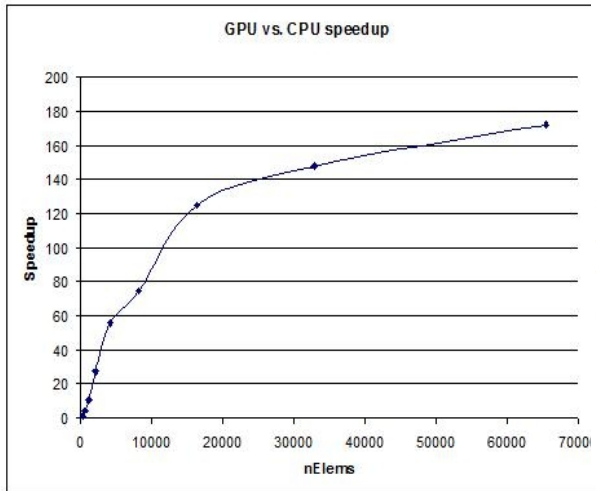
Журнал «Суперкомпьютеры»
Московский Физико-Технический Институт

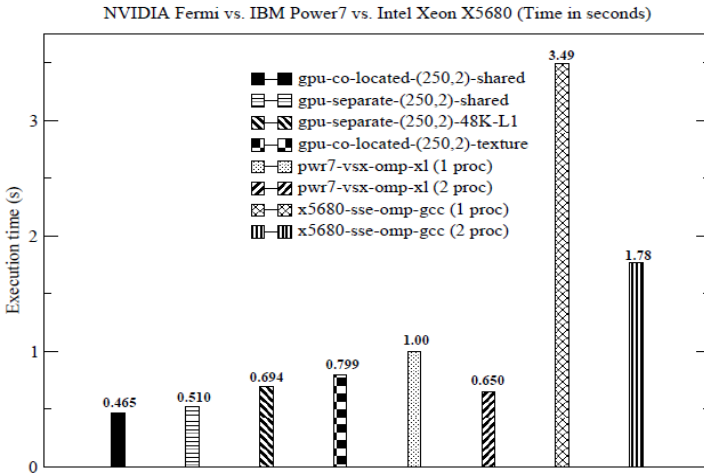
Сравнение вычислительных возможностей графических ускорителей NVidia

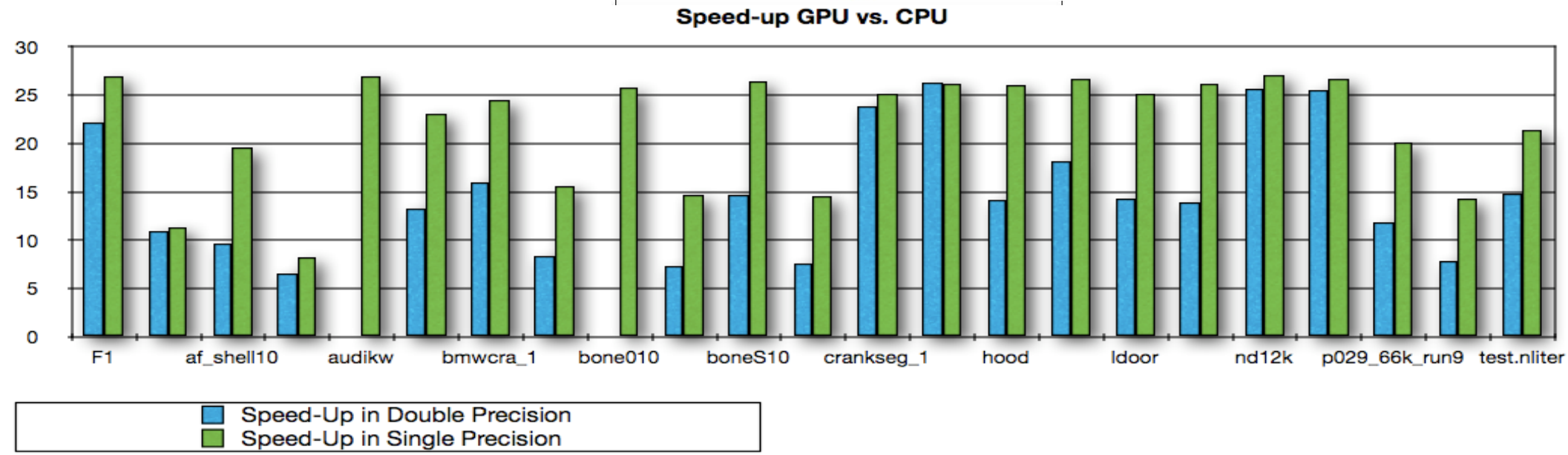
Авторы:
Кривов М.А.
Казеннов А.М.

Нижний Новгород, 2011

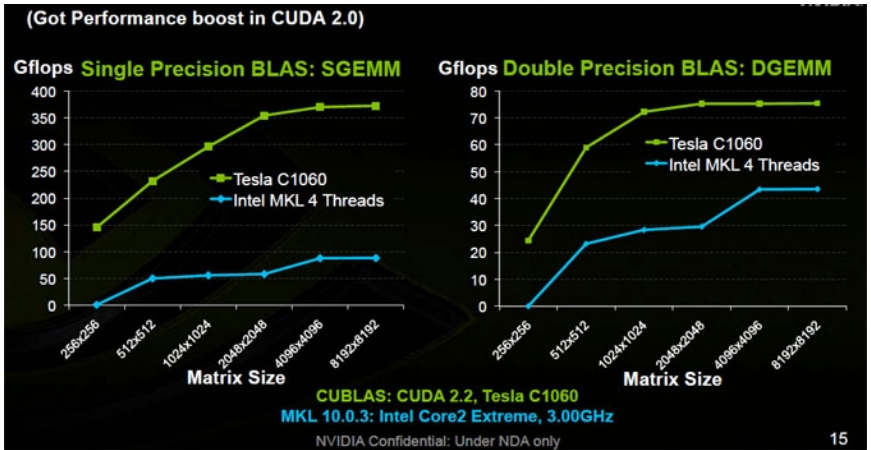

GROMACS 4.5 performance comparison
system: DHFR implicit (2489 atoms), solvated (23569 atoms)
Performance (ns/day)
Tesla C2050 ECC on
Tesla C2050 ECC off
GeForce GTX 470
GTX 580
Core i7 920 4T
Core i7 920 8T
2x Xeon E5430 8T
AMD X6 1090T 6T
Impl 1nm
Impl 2nm
Impl Inf
PME
RF 1nm

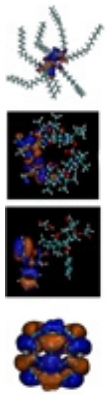

GPU vs. CPU speedup
Speedup
nElems

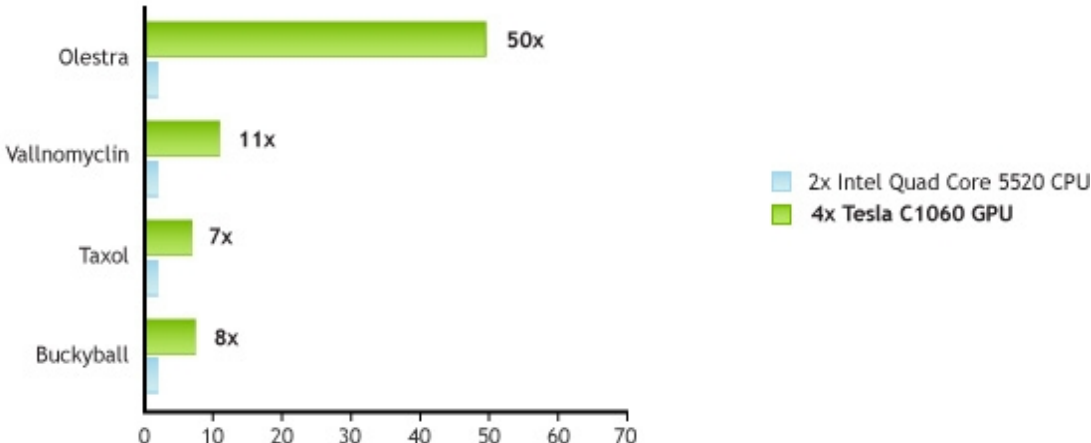

NVIDIA Fermi vs. IBM Power7 vs. Intel Xeon X5680 (Time in seconds)
Execution time (s)
gpu-co-located-(250,2)-shared
gpu-separate-(250,2)-shared
gpu-separate-(250,2)-48K-L1
gpu-co-located-(250,2)-texture
pwr7-vsx-omp-xl (1 proc)
pwr7-vsx-omp-xl (2 proc)
x5680-sse-omp-gcc (1 proc)
x5680-sse-omp-gcc (2 proc)
0.465
0.510
0.694
0.799
1.00
0.650
3.49
1.78


Speed-up GPU vs. CPU
F1
af_shell10
audikw
bmwcra_1
bone010
boneS10
crankseg_1
hood
ldoor
nd12k
p029_66k_run9
test.nliter

Speed-Up in Double Precision
Speed-Up in Single Precision


(Got Performance boost in CUDA 2.0)
Gflops Single Precision BLAS: SGEMM
Gflops Double Precision BLAS: DGEMM
Tesla C1060
Intel MKL 4 Threads
Matrix Size
CUBLAS: CUDA 2.2, Tesla C1060
MKL 10.0.3: Intel Core2 Extreme, 3.00GHz
NVIDIA Confidential: Under NDA only
15




Olestra
50x
Vallnomyclin
11x
Taxol
7x
Buckyball
8x
2x Intel Quad Core 5520 CPU
4x Tesla C1060 GPU

«Суперкомпьютеры» Зима-2010

GeForce или Tesla?

Авторы Максим Кривов, Андрей Казанков

Графические ускорители — одна из популярнейших тем в области высокопроизводительных вычислений. Почти все знают, что графические процессоры эффективны, дешевы и имеют колоссальную производительность. Но когда встает проблема выбора конкретной модели, то всплывают разные мелочи — эти карты не поддерживают операции с двойной точностью, у этих недостаточно памяти, а вот эти слишком дороги. В данном мини-обзоре приведены результаты тестирования 5 топовых видеокарт от NVIDIA, и показаны их реальные возможности, что позволит подобрать оптимальный вариант именно для ваших задач

Специализированный ускоритель или игровая карта?

Принято считать, что для вычислений идеально подходят именно карты серии Tesla. Они обладают и повышенной надежностью, и двойную точность поддерживают, и памяти у них побольше. Данное мнение активно поддерживается маркетологами NVIDIA, постоянно забывающими упомянуть о том факте, что «игровой» аналог каждой карты Tesla стоит раз так в 5 дешевле.

Как показали тесты, в большинстве случаев массовые GeForce обходят специализированные Tesla. Например, при выполнении Быстрого Преобразования Фурье с одинарной точностью GeForce GTX 480 оказалась в два раза быстрее, чем Tesla C2050. Самое интересное, что и при переходе к двойной точности лидером остается GeForce.

Fermi или не-Fermi?

Архитектура Fermi, представите-

№	Модель	Пиковая производительность (float/double), GFlops	Объем памяти, GB	Количество ядер	Примерная цена, руб.	Цена за GFlops (float/double), руб.	Max Flops тест (float/double), GFlops
1	GeForce GTX 480	1344.96/168	1.5	480	16000	11.8/95.2	1280/167.8
2	**Tesla C2050**	**1030.4/515.2**	**3**	**448**	**85000**	**82.5/165**	**977.8/406.2**
3	GeForce GTX 295	2x804.2/2x74.5	1.75	2x240	15000	8.3/100.6	691.8/74.2
4	**Tesla C1060**	**933/78**	**4**	**240**	**50000**	**53.6/643**	**722/77.4**
5	GeForce GTX 275	1008/84	1.75	240	11000	11/131	782.1/83.9
6	*GeForce 210*	*67.2/-*	*1*	*16*	*1900*	*28.27/-*	*51.9/-*

лями которой являются GeForce GTX480 и Tesla C2050, позиционируется компанией NVIDIA как самая передовая архитектура для вычислений. Но если посмотреть на технические детали, то ее превосходство становится не так очевидно — пиковая производительность возросла «всего» на какие-то 100-300 Гигафлопсов, а объем памяти даже уменьшился. Так, более новая Tesla C2050 оснащается 3 Гигабайтами, в то время как ее предшественник, Tesla C1060, комплектовался 4 Гигабайтами.

Согласно тестам, преимущества у новой архитектуры все-таки есть, и весьма значительные. Во-первых, использование более современного стандарта памяти GDDR5 позволило значительно повысить пропускную способность, в результате чего глобальная память карты GeForce GTX480 оказалась даже в 1.5 раза быстрее, чем разделяемая память GeForce GTX 295. Во-вторых, в новых картах появилось большее количество ядер и улучшенная поддержка операций с двойной точностью. Все это вместе позволило обойти предыдущее поколение во всех тестах. И хотя в большинстве случаев отрыв не превышает десятков процентов, на задачах типа DGMEM использование новых ускорителей может повысить производительность более чем в 4 раза.

Результаты

Подводя итог, стоит озвучить мысль, которая будет витать в GP-GPU сообществе еще не один год, — все зависит от задач. Если требуется реализовать алгоритм, работающий с небольшими объемами данных, то однозначно стоит предпочесть карты массовой серии GeForce, обеспечивающие за гораздо меньшую стоимость большую производительность. Если же возникает необходимость оперировать большими объемами данных (которыми могут оказаться как разностные сетки, так и матрицы) или на первое место выходят соображения надежности при работе в режиме 24/7, то предпочтение стоит отдать профессиональным картам серии Tesla. Иначе резко возрастает

того, что для прием
сти памяти просто
в самый неподходя
что-то случится и
один месяц вычис
При выборе графи
теля также имеет с
внимание и на вид
дущего поколения
ускорители можно
большой уценкой,
задачах по вычисл
можностям они не
играют своим п
Для оценки, а стои
переносить свои п
графические ускор
воспользоваться д
ными решениями
почти в каждом ко
добрав правильны
пропорционально
GeForce 210, испол

ном обзоре, подобный коэффициент варьируется от 5 и примерно до 15.

О бенчмарке

В качестве метрики в данном обзоре использовался бенчмарк SHOC (Scalable HeterOgeneous Computing), разрабатываемый американской лабораторией ORNL.

Для тестирования использовалась первая стабильная версия 1.0, выпущенная в начале декабря.

Данный бенчмарк состоит из следующего набора синтетических тестов, измеряющих производительность каждой подсистемы

FFT (float/double), GB/s	GEMM (float/double), GFlops	S3D (float/double), GFlops	Тест пропуск... Global memory (coalesced)
192.6/65.5	318.5/85.1	45.5/25.0	151.8
102.5/52.3	**261.3/71.1**	**34.2/18.8**	**90.2**
96.2/26.9	210.5/15.6	27.7/14.7	64.0
92.7/36.4	**212.3/16.0**	**25.8/13.9**	**69.7**
101.0/29.7	233.8/17.7	31.2/16.6	70.4
15.3/-	*25.7*	*3.2*	*6.2*

План выступления

Описание тестовых систем

Бенчмарк SHOC

Тестирование MAGMA и CUBLAS

Тестирование RODINIA и примеров из CUDA-SDK

Результаты

Тестовая система

- CPU
  - ***2x*** AMD Opteron 2427 @2.2 GHz (12 ядер)
- GPU
  - Tesla C2050
  - Tesla C1060
  - GeForce 480 GTX
  - GeForce 295 GTX
  - ~~GeForce 9800 GTX~~
- RAM
  - 16 GB DDR3
- OS
  - Linux RedHat
  - NVidia CUDA Toolkit 3.1

Тестируемые ускорители

Tesla C2050

Производительность: 1030,4 / 515,2 GFlops
Количество ядер: 448 @ 1147 MHz
Память: 3 GB

Tesla C1060

Производительность: 933 / 78 GFlops
Количество ядер: 240 @ 1296 MHz
Память: 4 GB

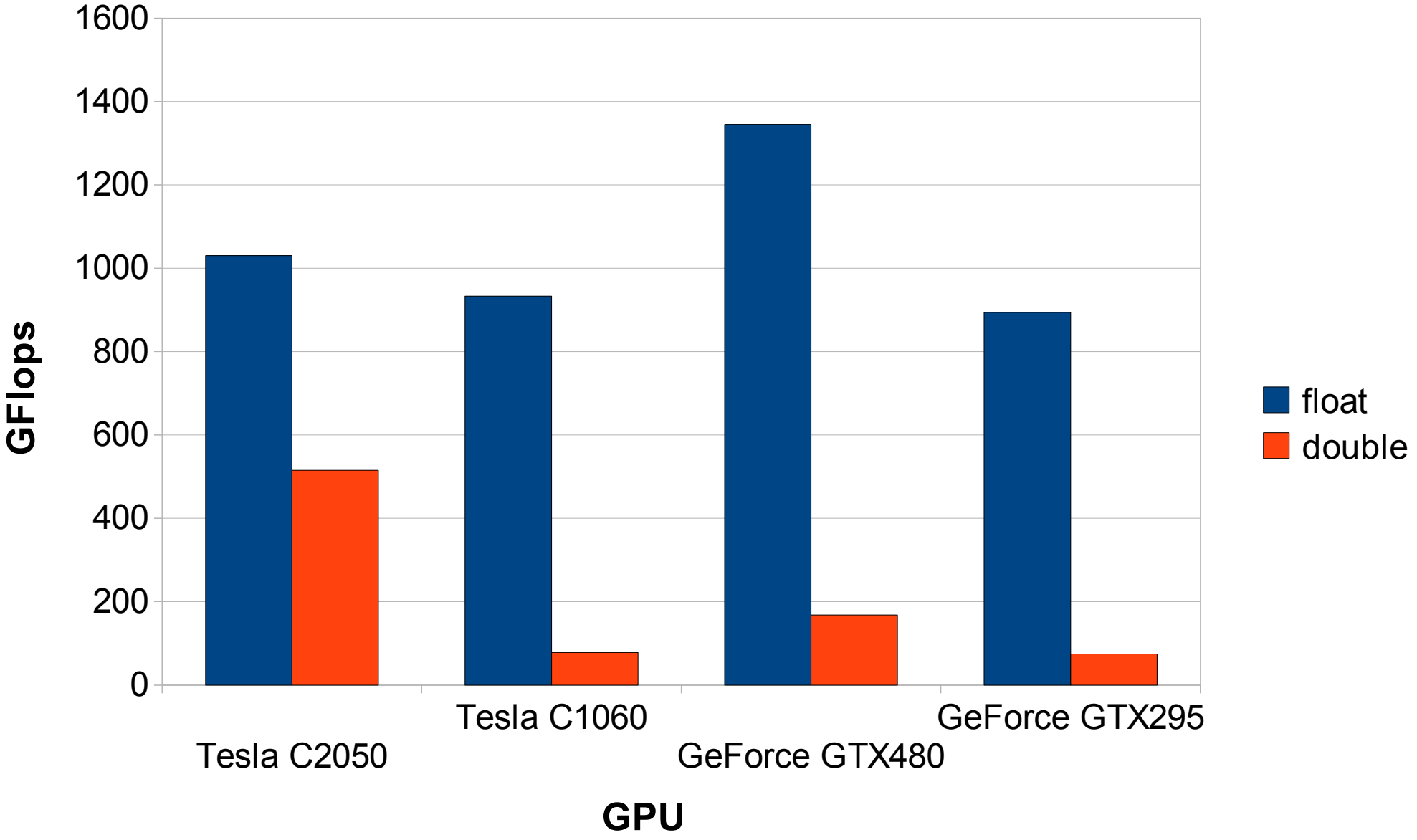
GeForce 480 GTX

Производительность: 1344,96 / 168 GFlops
Количество ядер: 480 @ 1401 MHz
Память: 1,5 GB

GeForce 295 GTX

Производительность: **2x**894,2 / **2x**74,5 GFlops
Количество ядер: **2x**240 @ 1242 MHz
Память: **2x**0,87 GB

Пиковая производительность



План выступления

Описание тестовых систем

Бенчмарк SHOC

Тестирование MAGMA и CUBLAS

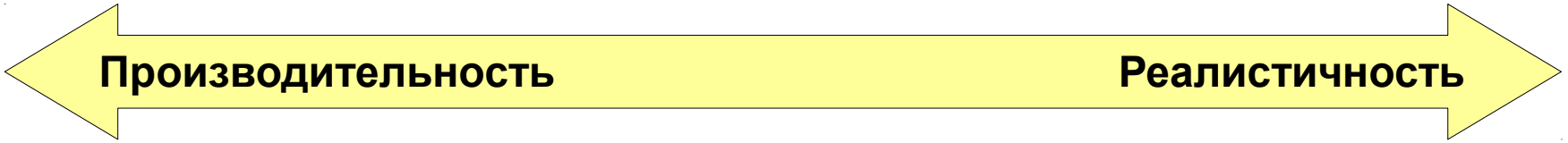
Тестирование RODINIA и примеров из CUDA-SDK

Результаты

SHOC - Scalable HeteroGeneous Computing

- Сайт
  - http://ft.ornl.gov/doku/shoc
- Используемая версия
  - SHOC-1.01 (OpenCL + CUDA)
- Структура

Level 0	Level 1	Level 2
MaxFlops DeviceMemory BusSpeed	FFT MD / NBody SGEMM Reduction	S3D



Результаты тестирования

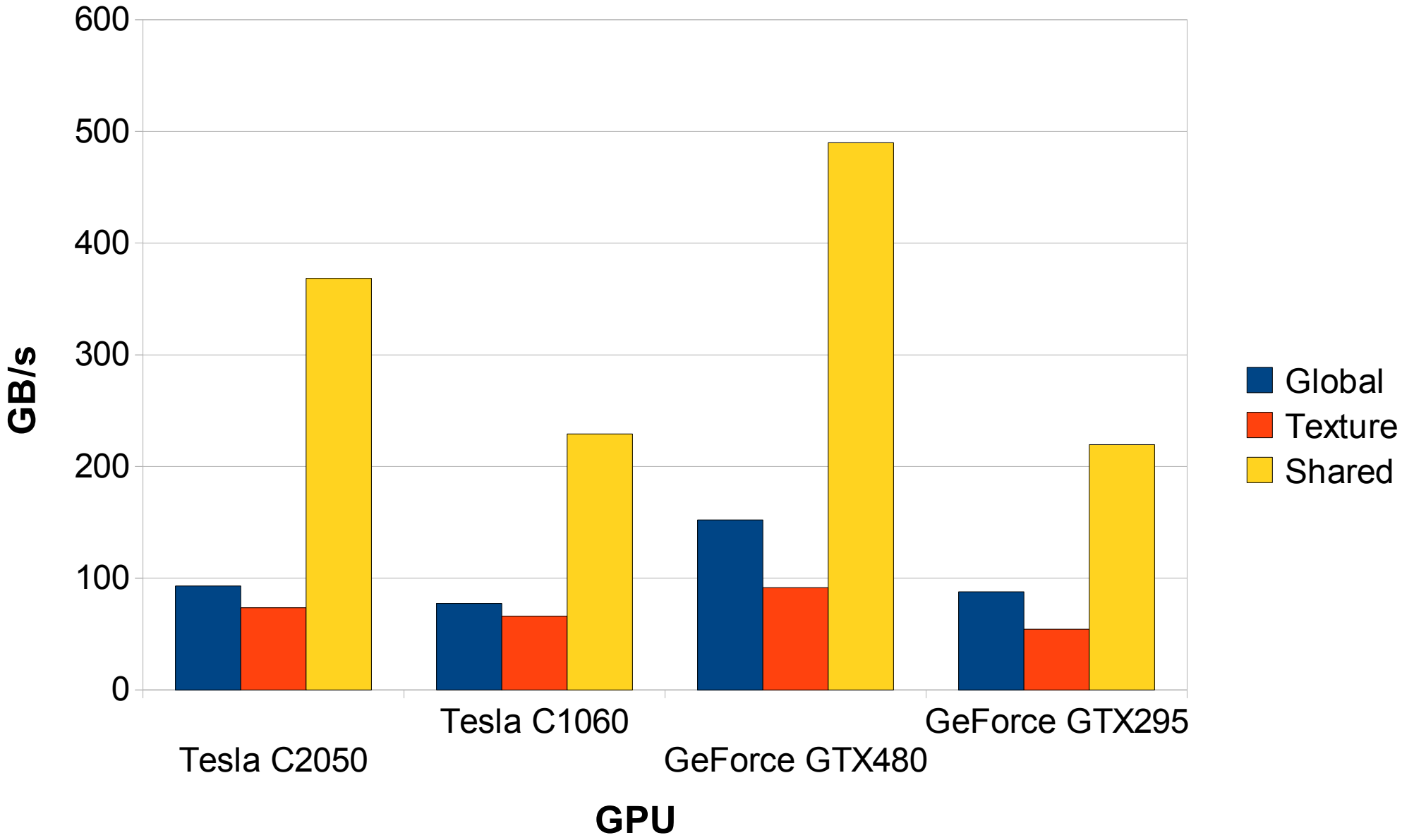
SHOC-CUDA

GPU	**MaxFlops** (float/double) GFlops	**FFT** (float/double) GFlops	**GEMM** (float/double) GFlops	**Bandwidth** GB/s			**MD** (float/double) GB/s	**Reduction** (float/double) GB/s
				Global	Texture	Shared		
Tesla C2050	1002 / 501	69 / 34,3	301,5 / 68	93,1	73,6	368,5	74 / 84,5	61,2 / 65
Tesla C1060	721,9 / 77	94 / 26,7	202 / 38,1	77,5	66	229,2	59,1 / 27	46 / 42,4
GeForce 480 GTX	1313 / 168	202 / 63,5	366,6 / 84	152,3	91,6	489,9	91,5 / 75	83 / 86,6
GeForce 295 GTX	691 / 74,2	93 / 25,7	200,2 / 38	87,9	54,4	219,6	51,5 / 26	50 / 42,2

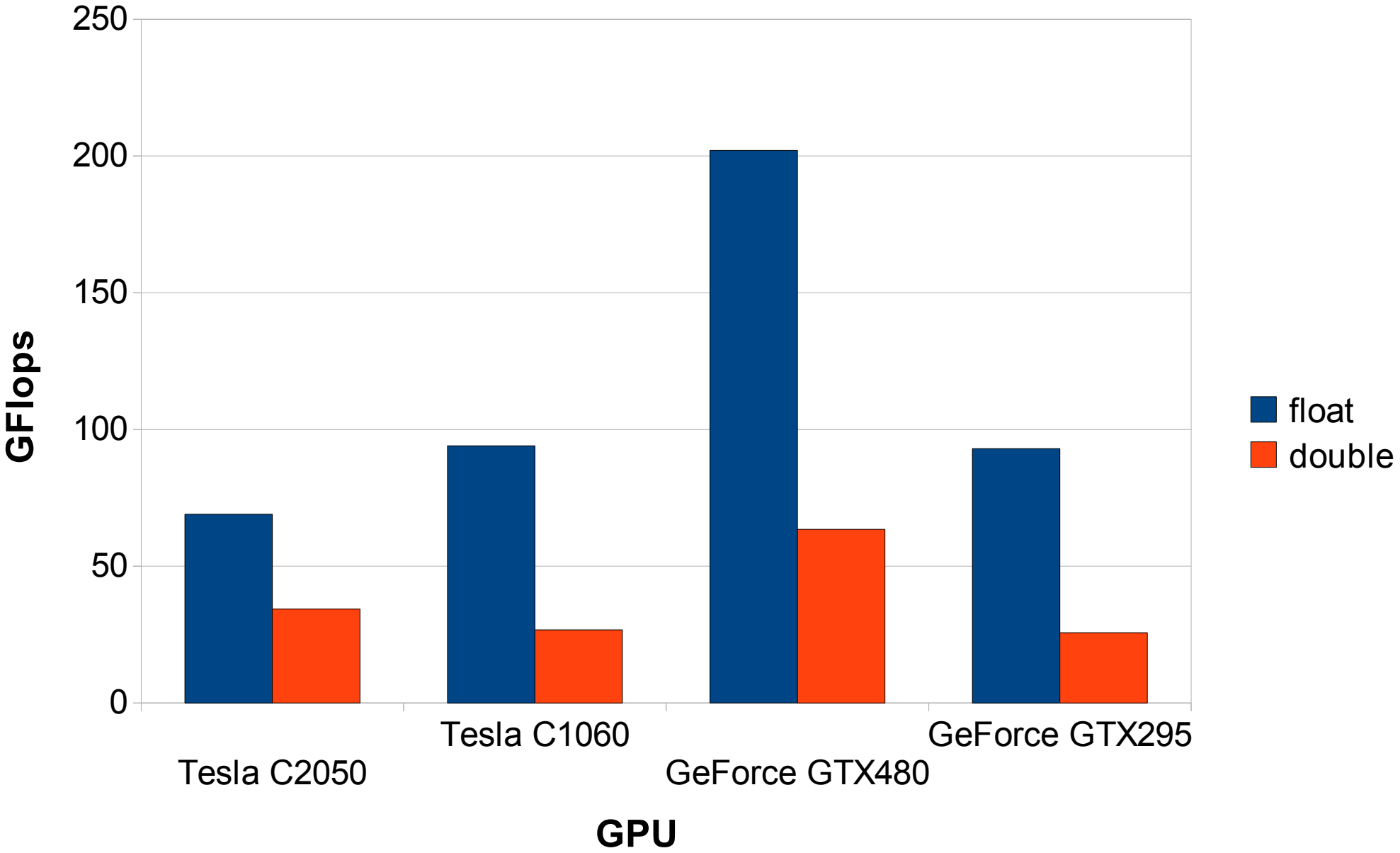
SCHOC-OpenCL

GPU	**MaxFlops** (float/double) GFlops	**FFT** (float/double) GFlops	**GEMM** (float/double) GFlops	**Bandwidth** GB/s			**MD** (float/double) GB/s	**Reduction** (float/double) GB/s
				Global	Image	Local		
Tesla C2050	1005 / 503	40,4 / 17	274 / 51,6	95	72,4	371,8	25 / 27,7	34 / 36,6
Tesla C1060	727 / 77,6	23 / 7, 1	133 / 20,5	81,9	66,2	262,6	23 / 22,6	26 / 24,9
GeForce 480 GTX	1317 / 168	52,1 / 17	334 / 52,9	164	98,2	486,5	30,7 / 34	45 / 47,2
GeForce 295 GTX	696 / 74,3	26 / 6, 9	136,6 / 20	94,3	54,7	251,7	22 / 21,7	26 / 23,8

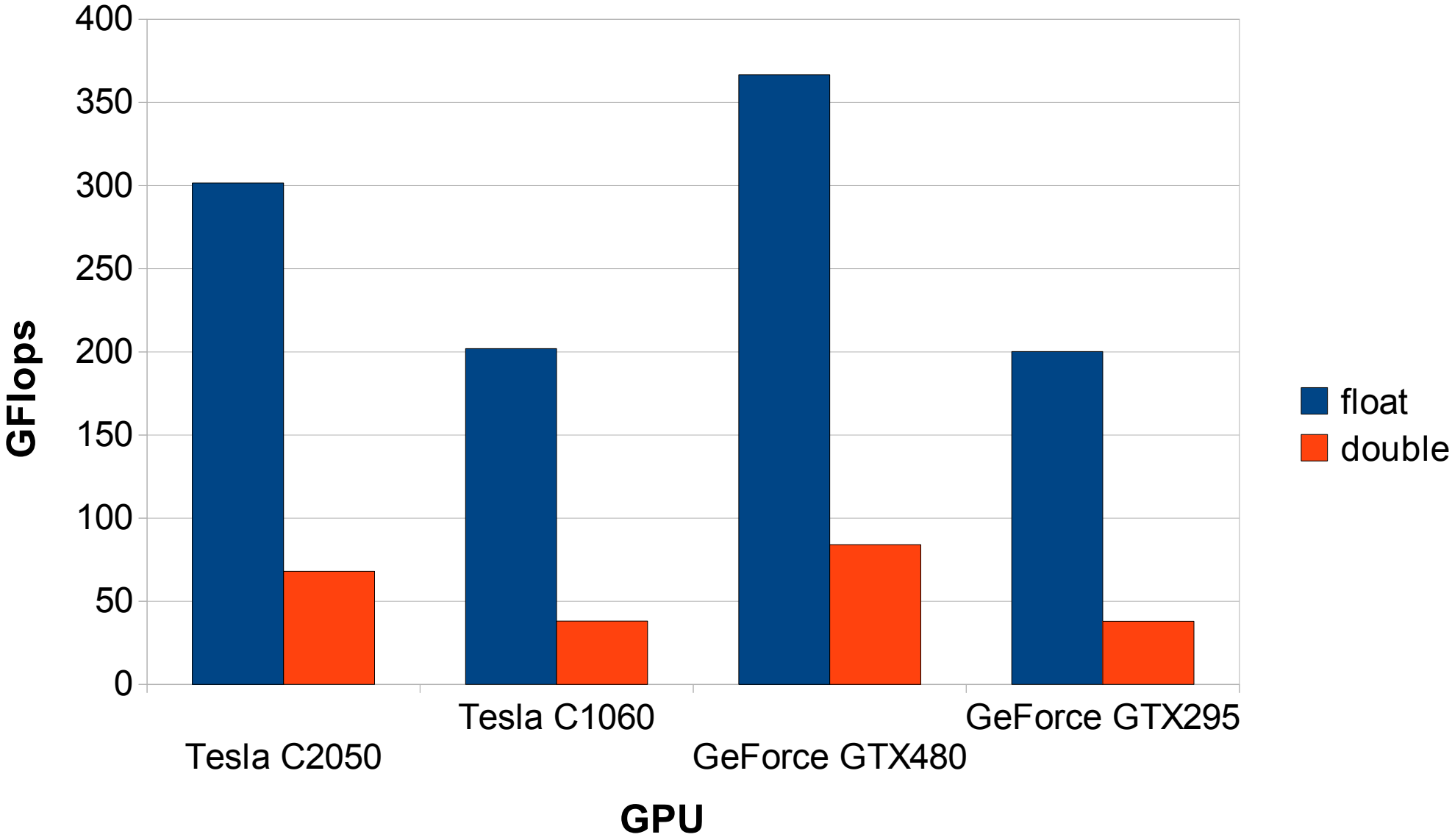
Пропускная способность памяти



FFT


250
200
150
100
50
0
GFlops
Tesla C2050
Tesla C1060
GeForce GTX480
GeForce GTX295
GPU
float
double

SGEMM / DGEMM


400
350
300
250
200
150
100
50
0
GFlops
Tesla C2050
Tesla C1060
GeForce GTX480
GeForce GTX295
GPU
float
double

План выступления

Описание тестовых систем

Бенчмарк SHOC

Тестирование MAGMA и CUBLAS

Тестирование RODINIA и примеров из CUDA-SDK

Результаты

Описание

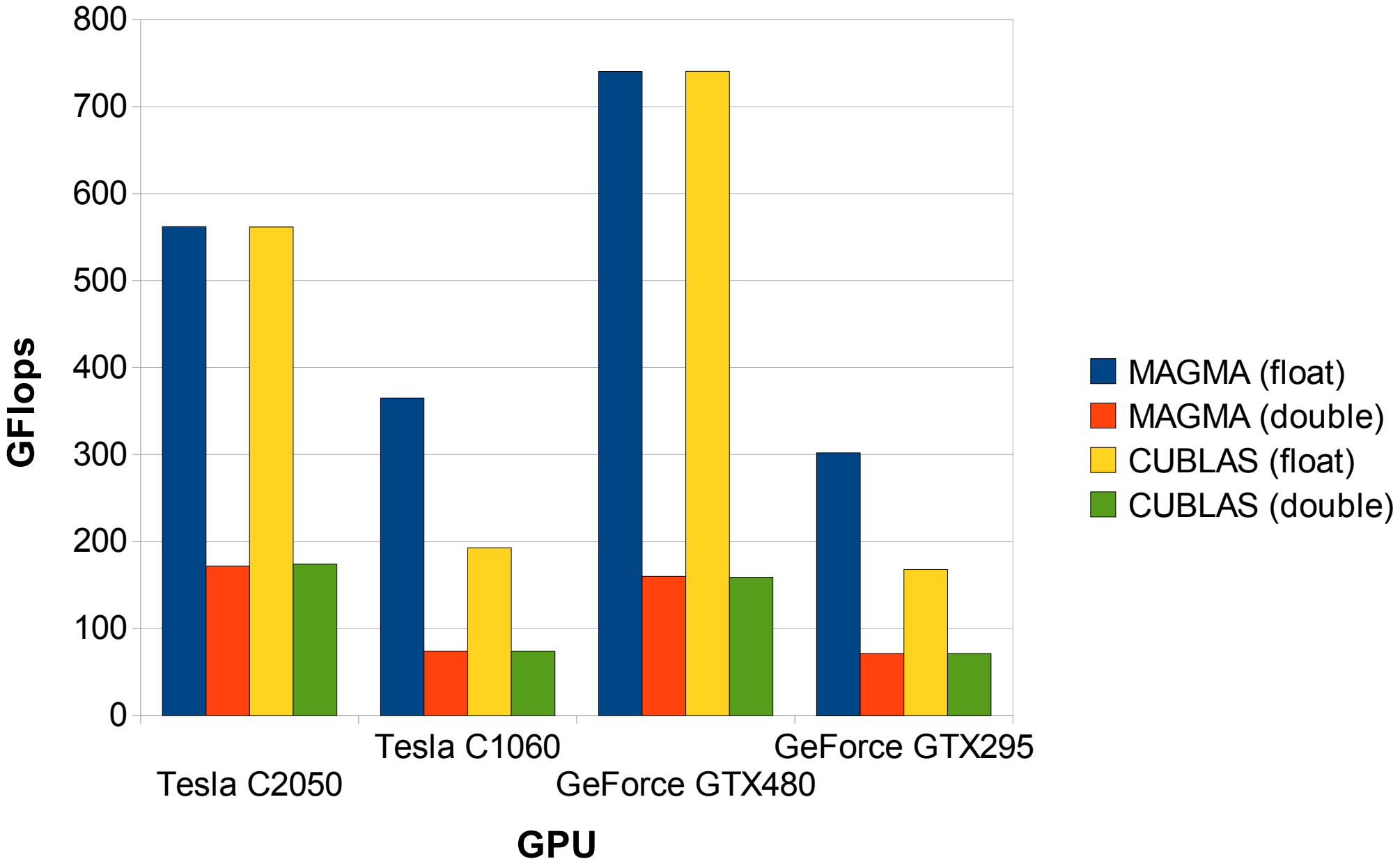
- MAGMA
  - Реализация LAPACK для систем GPU+CPU
  - Сайт - http://icl.cs.utk.edu/magma/
- CUBLAS
  - Реализация BLAS1 и BLAS2 на CUDA
  - Поставляется с CUDA Toolkit
- Тесты
  - ***GEMM***: `X = A * B`
  - ***GEMV***: `x = A * u`
  - ***GESV***: `x : A * x = b`
  - ***SYMV***: `x = S * u + r`

Результаты тестирования

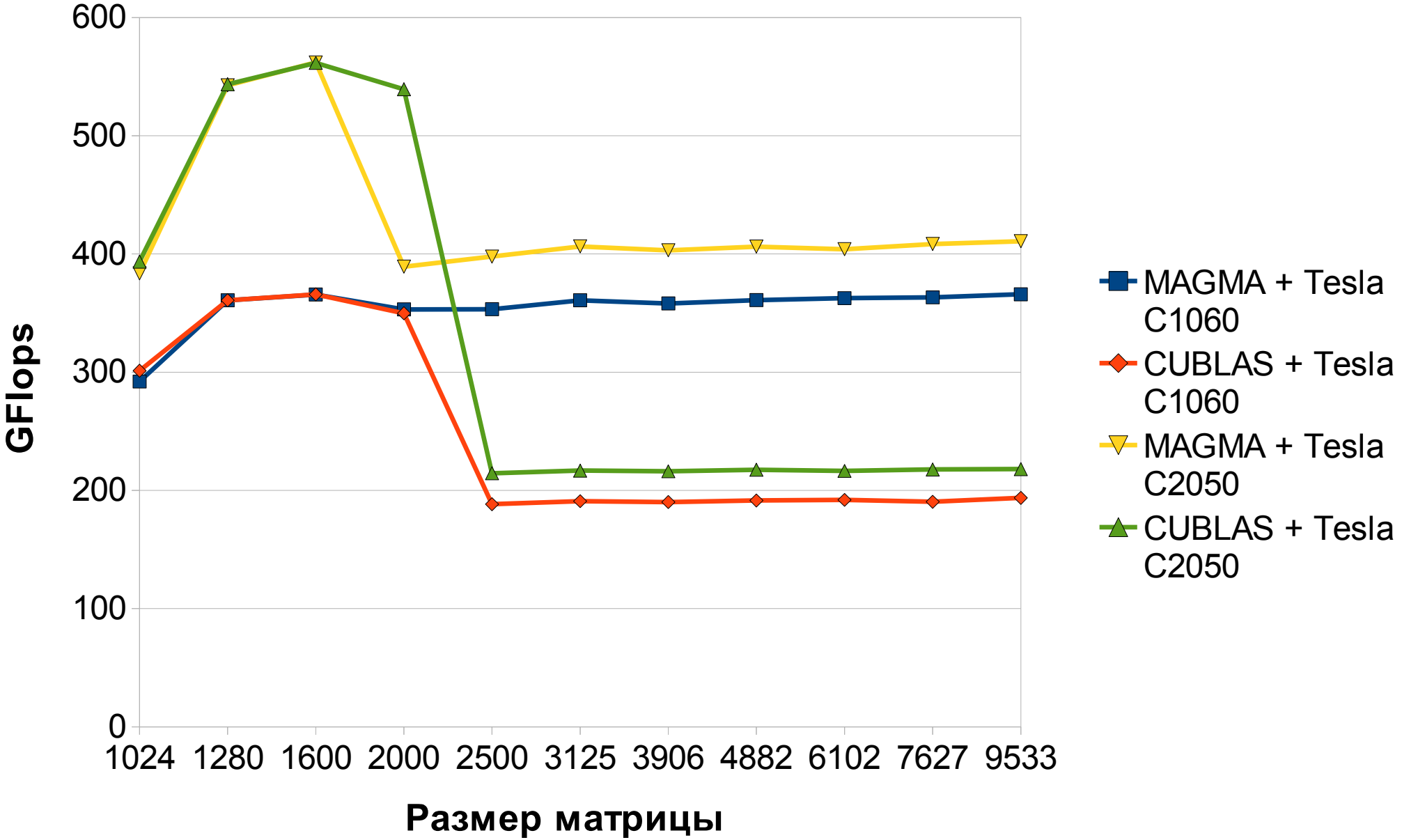
MAGMA vs CUBLAS

GPU	**GEMM** (float / double), GFlops		**GEMV** (float / double), GFlops		**GESV** (float / double), GFlops		**SYMV** (float / double), GFlops	
	MAGMA	CUBLAS	MAGMA	CUBLAS	MAGMA	CUBLAS	MAGMA	CUBLAS
Tesla C2050	562 / 172	561,6 / 174	42,7 / 22	45,7 / 20	320 / 142,5	- / -	50 / 31	15 / 12, 4
Tesla C1060	365 / 74	193 / 74	39 / 21,3	40 / 16,5	288,3 / 66	- / -	60 / 23,4	16,4 / 3,7
GeForce 480 GTX	740,2 / 160	740,6 / 159	75 / 30	64 / 32	408,5 / 140	- / -	67 / 43,1	22, 5 / 15,2
GeForce 295 GTX	302 / 71,4	168 / 71,3	49 / 23	48 / 18	280,9 / 64	- / -	67 / 22,8	16,2 / 3,6

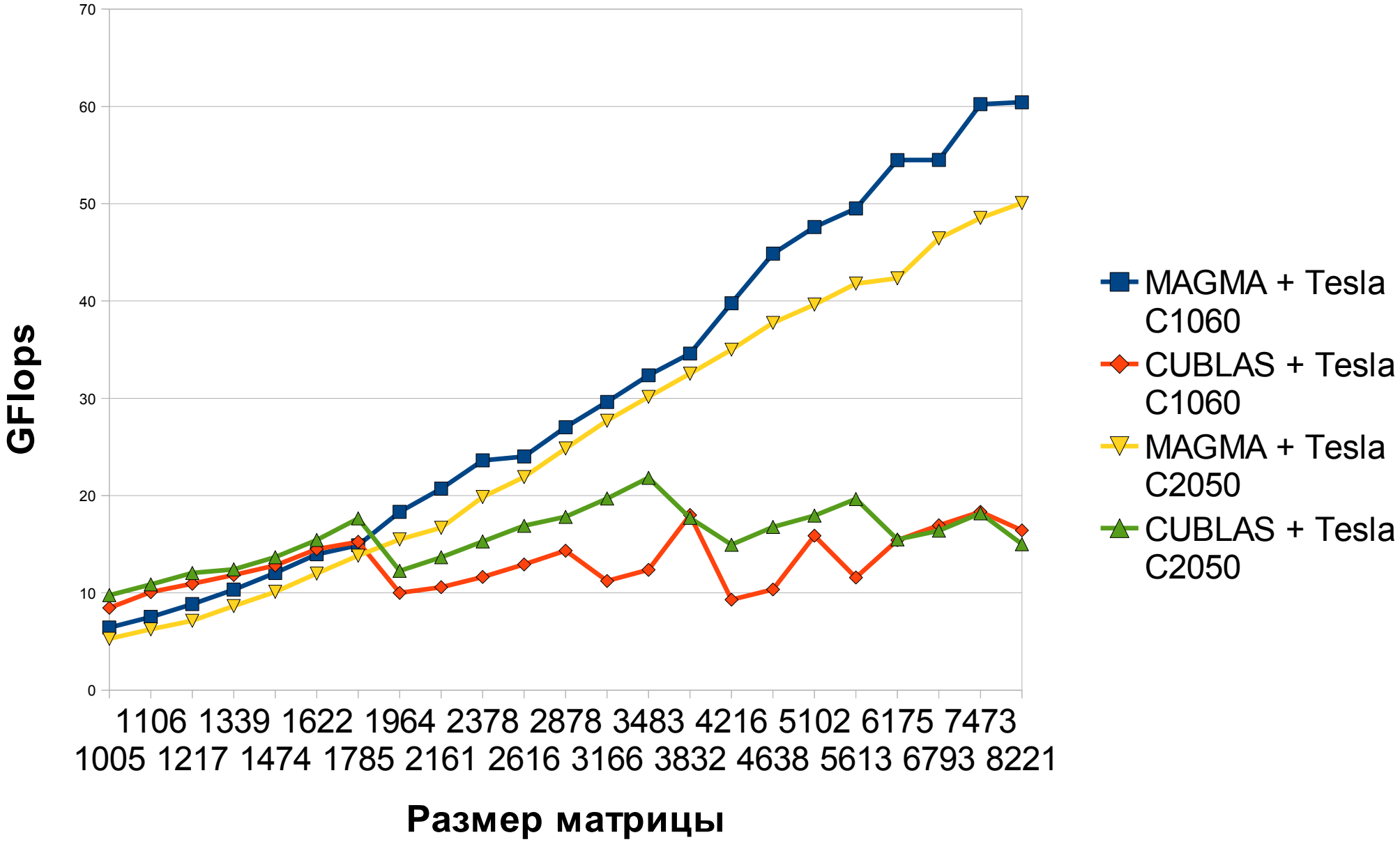
SGEMM / DGEMM


800
700
600
500
400
300
200
100
0
GFlops
Tesla C2050
Tesla C1060
GeForce GTX480
GeForce GTX295
GPU
MAGMA (float)
MAGMA (double)
CUBLAS (float)
CUBLAS (double)

SGEMM


600
500
400
300
200
100
0
GFlops
1024 1280 1600 2000 2500 3125 3906 4882 6102 7627 9533
Размер матрицы
MAGMA + Tesla C1060
CUBLAS + Tesla C1060
MAGMA + Tesla C2050
CUBLAS + Tesla C2050

SSYMV


GFlops
70
60
50
40
30
20
10
0
1106 1339 1622 1964 2378 2878 3483 4216 5102 6175 7473
1005 1217 1474 1785 2161 2616 3166 3832 4638 5613 6793 8221
Размер матрицы
MAGMA + Tesla C1060
CUBLAS + Tesla C1060
MAGMA + Tesla C2050
CUBLAS + Tesla C2050

План выступления

Описание тестовых систем

Бенчмарк SHOC

Тестирование MAGMA и CUBLAS

Тестирование RODINIA и примеров из CUDA-SDK

Результаты

Описание

- RODINIA
  - Набор вычислительно-ёмких ядер на CUDA и OpenMP из разных областей
  - Сайт - https://www.cs.virginia.edu/~skadron/wiki/rodinia/
- NVidia CUDA Samples
  - Black-Scholes
  - Monte Carlo
  - Fast walsh transform
  - Radix sort
  - Sorting networks

Результаты тестирования

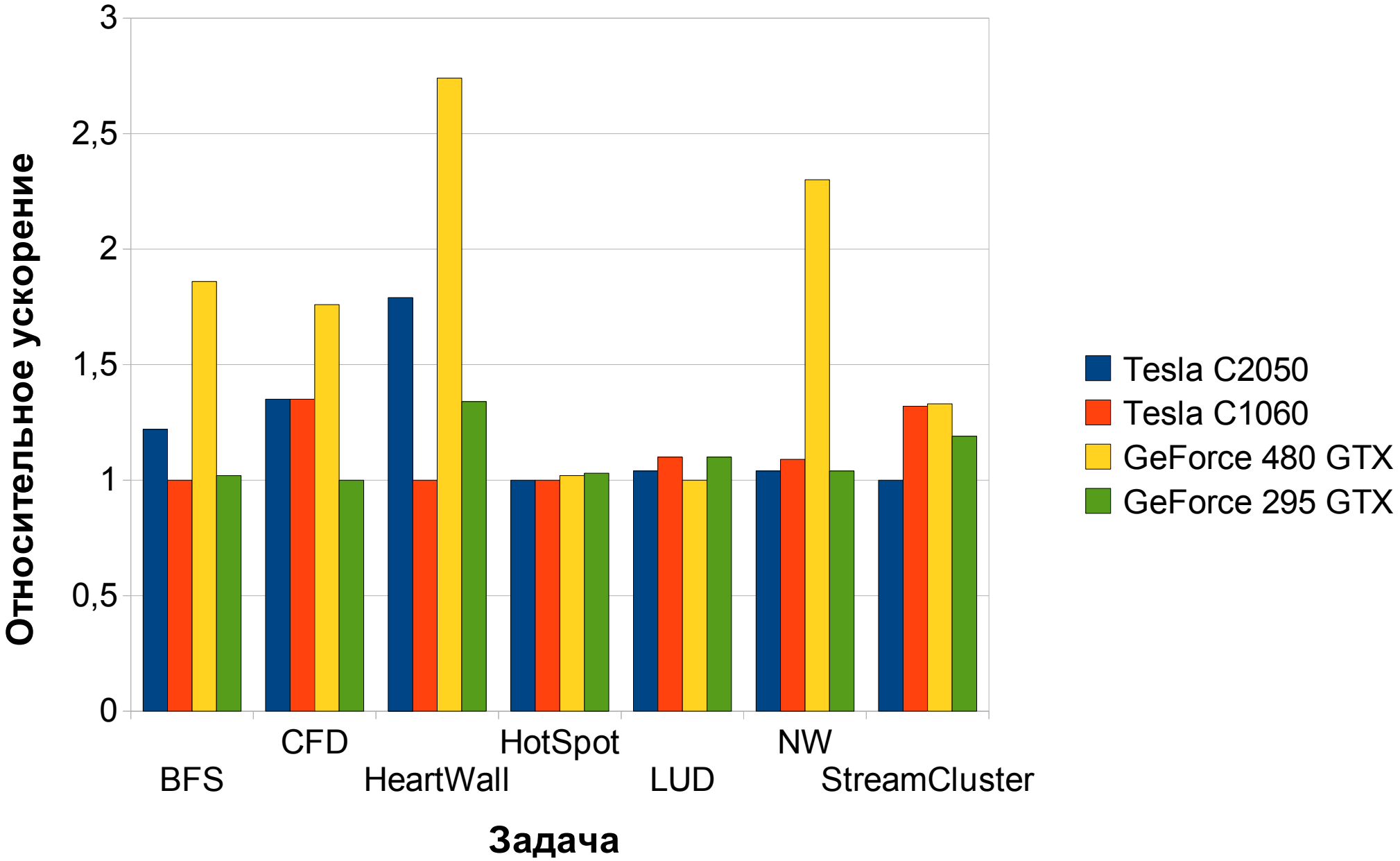
Rodinia

	BFS msec	**CFD** sec	**HeartWall** msec	**HotSpot** msec	**LUD** msec	**NW** sec	**StreamCluster** sec
Предметная область	*Графы*	*Аэро-динамика*	*Медицина*	*Физика*	*Линейная алгебра*	*Био-информатика*	*Поиск данных*
Tesla C2050	15,4	6,6	201	6,64	2,1	2,2	16,4
Tesla C1060	18,8	6,6	360	6,6	2	2,1	12,4
GeForce 480 GTX	10,1	5,05	131,2	6,5	2,2	2,3	12,27
GeForce 295 GTX	18,26	8,9	268	6,4	2	2,2	13,7

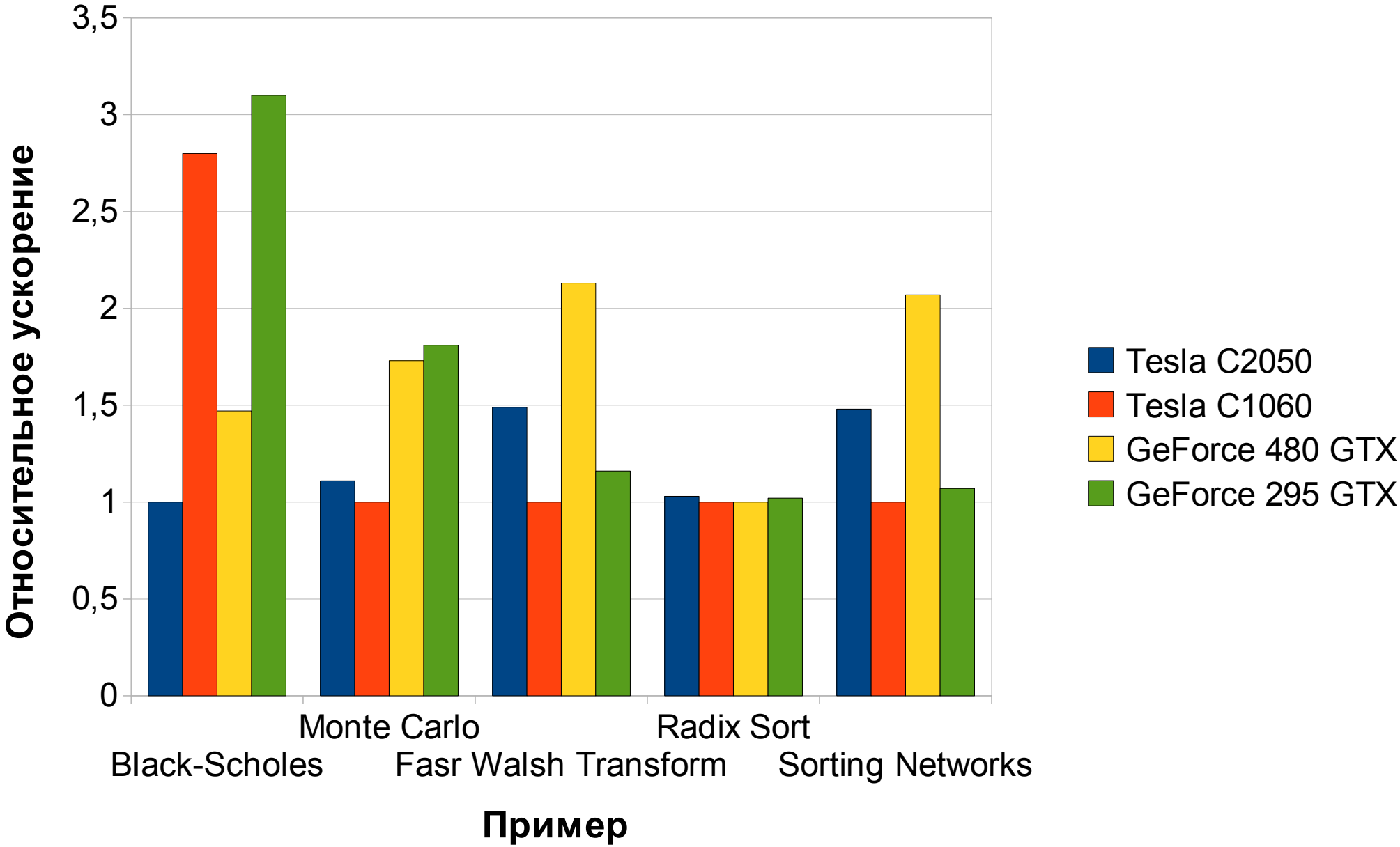
CUDA Samples

GPU	**Black-Scholes**		**Monte Carlo** msec	**Fast Walsh Transform** msec	**Radix sort**		**Sorting networks** msec
	Time, msec	Bandwidth GB/s			Int, msec	Float, msec	
Tesla C2050	2,8	27,6	3,4	17,5	47,5	47,1	14,8
Tesla C1060	1	80,5	3,78	26,2	49,7	48,4	22
GeForce 480 GTX	1,9	40,6	2,18	12,3	47,7	47,4	10,6
GeForce 295 GTX	0,9	86,5	3,2	22,6	50,1	48,6	20,5

RODINIA


Относительное ускорение
3
2,5
2
1,5
1
0,5
0
BFS
CFD
HeartWall
HotSpot
LUD
NW
StreamCluster
Задача
Tesla C2050
Tesla C1060
GeForce 480 GTX
GeForce 295 GTX

CUDA Samples



План выступления

Описание тестовых систем

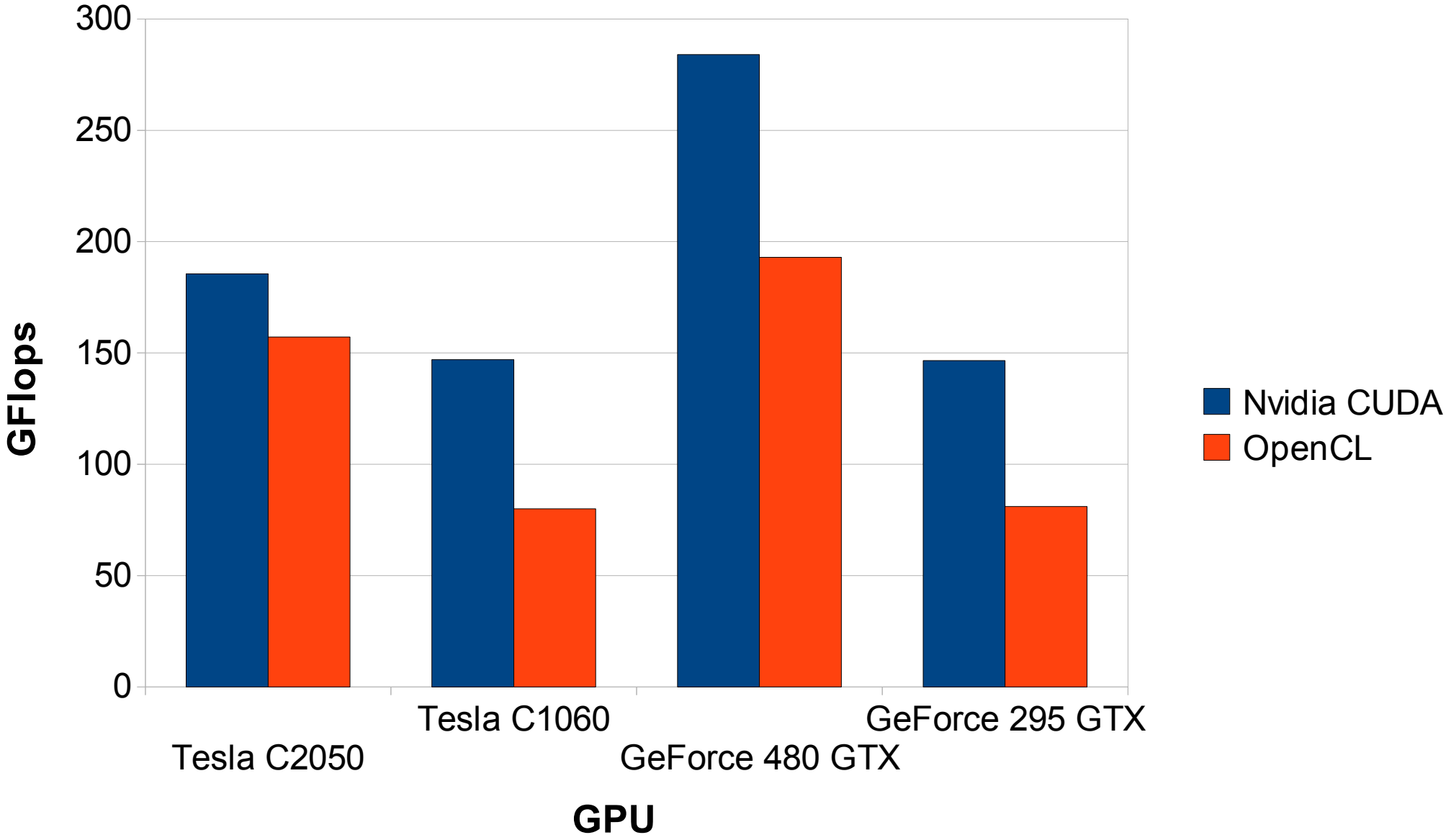
Бенчмарк SHOC

Тестирование MAGMA и CUBLAS

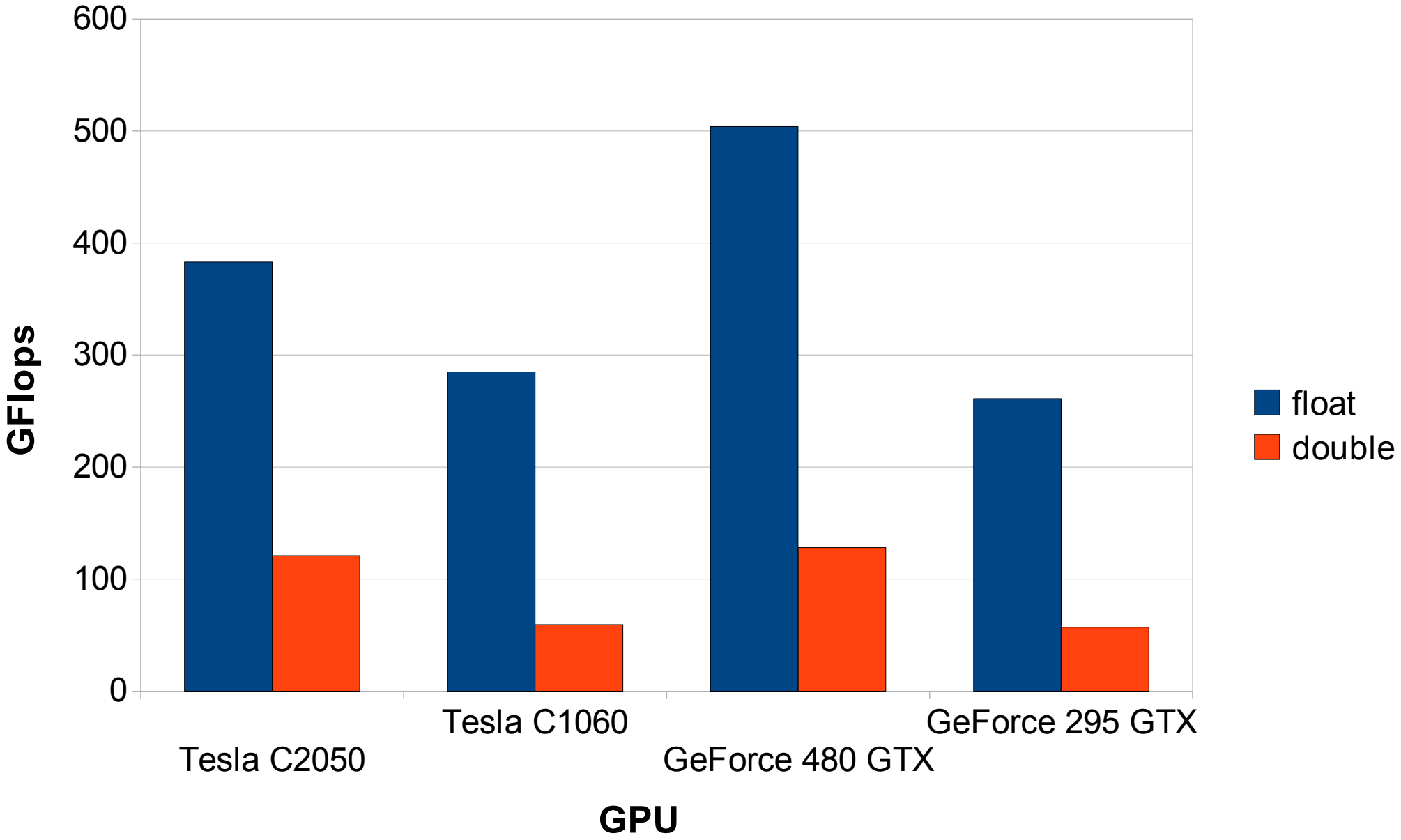
Тестирование RODINIA и примеров из CUDA-SDK

Результаты

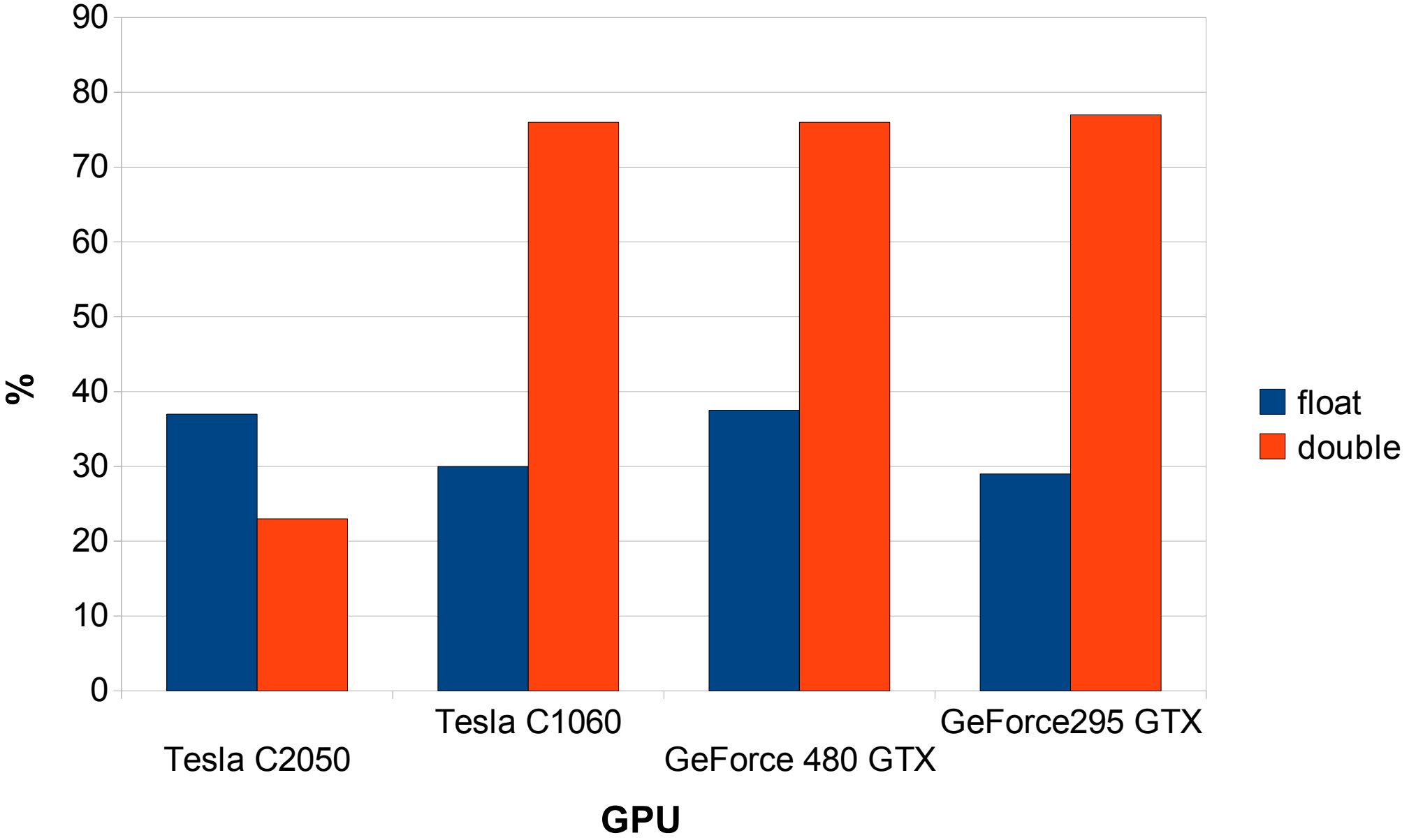
OpenCL vs CUDA



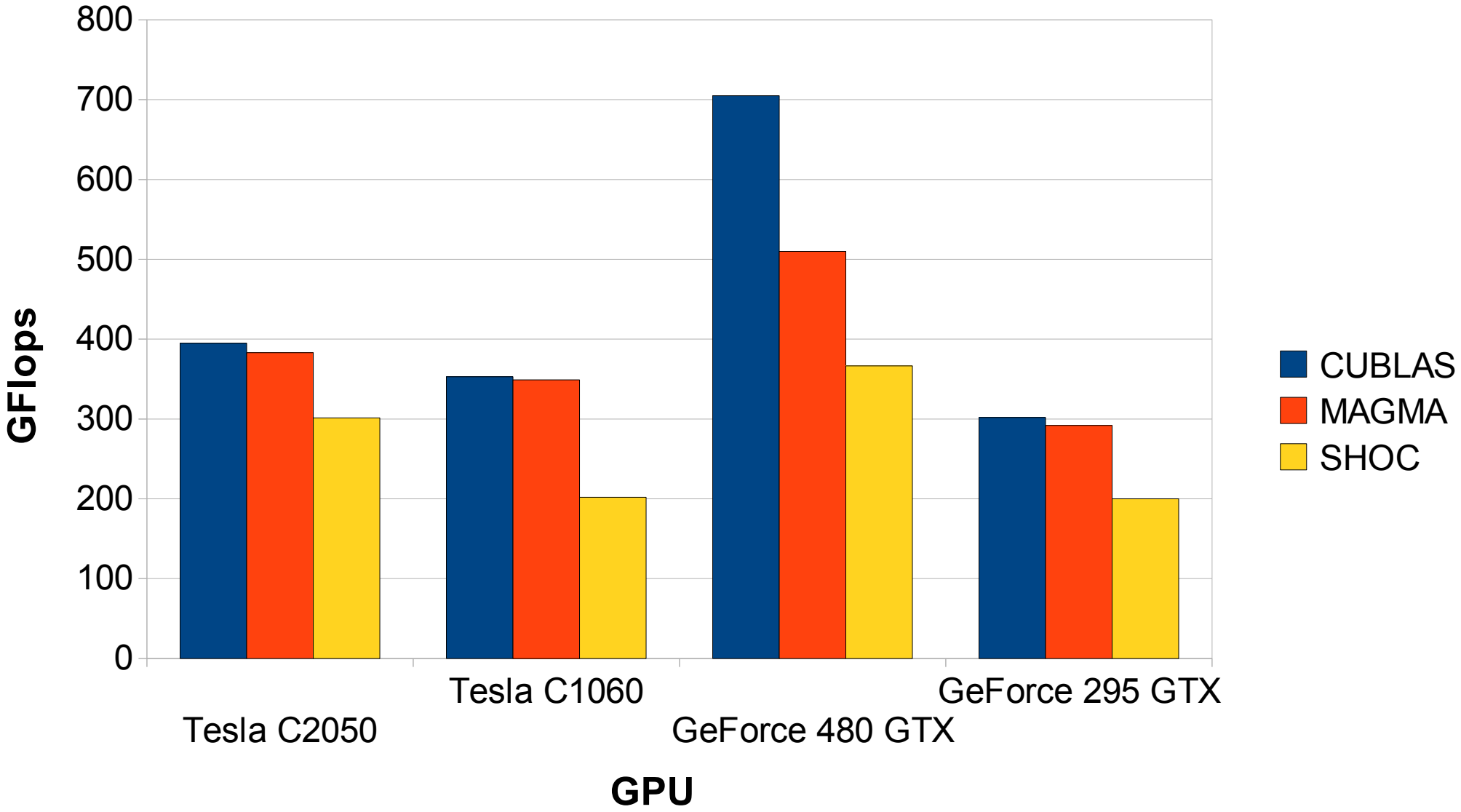
float vs double



float vs double (эффективность)



Различные реализации одного алгоритма (SGEMM)



Вопросы?

Сайт журнала:

http://www.supercomputers.ru

Контактная информация:

Кривов Максим, maxim.krivov@supercomputers.ru
Андрей Казеннов, kazennov@gmail.com