Сравнение вычислительных возможностей графических ускорителей NVidia при решении различных классов задач

Кривов М.А.[1], Казеннов А.М.[2],
[1] *Издательство SCR-Media, журнал «Суперкомпьютеры»,*
[2] *Московский Физико-Технический Университет*

Введение

Современные графические ускорители обладают не только колоссальными вычислительными возможностями, но и достаточно сложной архитектурой, затрудняющей их эффективное использование при решении задач из некоторых областей. При адаптации алгоритмов под графические ускорители, достаточно часто узким местом оказывается не производительность GPU, а недостаточный объём какого-либо типа памяти, затраты на подготовку данных центральным процессором или неэффективность использования потоковых процессоров.

В данной работе, являющейся расширенной версией статьи [1], проведено тестирование четырёх видеокарт от Nvidia из различных сегментов с целью оценки реально достижимой производительности при решении разных классов задач. В качестве тестов были выбраны GPGPU-бенчмарки SHOC и Rodinia, программы, использующие библиотеки MAGMA и CUBLAS, а также примеры из Nvidia CUDA SDK. В частности, были проведены сравнения производительности при использовании технологий OpenCL и NVidia CUDA, вычислениях с двойной и одинарной точностью и выполнении различных реализаций одного и того же алгоритма.

Тестовые системы

Тестирование проводилось на графическом кластере МФТИ, каждый узел которого оснащён двумя 6-ядерными процессорами AMD Opteron 2427 @2.2 GHz и 16 гигабайтами памяти типа DDR3. В качестве графических ускорителей использовались следующие видеокарты NVidia:

GPU	Пиковая производительность (float / double), GFlops	Объём памяти, GB	Количество CUDA-ядер	Частота CUDA-ядер, MHz
Tesla C2050	1030,4 / 515,2	3	448	1147
Tesla C1060	933 / 78	4	240	1296
GeForce 480 GTX	1344,96 / 168	1,5	480	1401
GeForce 295 GTX	2x894,2 / 2x74,5	1,75	2x240	1242

Таблицы 1. Описание тестируемых ускорителей

Графические ускорители Tesla C2050 и GeForce 480 являются представителями современной архитектуры GPU под названием Fermi, результатом чего является их более высокая производительность при операциях с двойной точностью и поддержка дополнительной функциональности (архитектурных расширений 2.0). Две другие видеокарты (Tesla C1060 и GeForce 295) являются модификациями более старой архитектуры GT200, представленной в 2008 году и предназначенной для проведения расчётов с одинарной точностью.

Стоит также отметить, что графический ускоритель GeForce 295 состоит из двух относительно независимых процессоров и программно распознаётся как две

видеокарты, поэтому при тестировании использовалась только половина доступных вычислительных возможностей данного ускорителя.

Тесты SHOC

Пакет тестов Scalable HeterOgeneous Computing (или SHOC) [2], разрабатывается известной американской лабораторией ORNL как независимый бенчмарк для гетерогенных систем, оснащённых графическими ускорителями. Все тесты логически разбиты на три уровня, определяющие их приближенность к реальности. Так, приложения из уровня 0 являются синтетическими тестами и нацелены на получение максимально возможной производительности, в то время как тесты из уровня 2 реализуют решения более реальных задач и, соответственно, показывают более скромные результаты.

Стоит отметить, что SHOC поддерживает как технологию NVidia CUDA, так и OpenCL, а начиная с версии 1.01 появилась возможность запуска бенчмарка на гетерогенных кластерах (с помощью библиотеки MPI). Результаты запуска некоторых тестов из CUDA-версии для одного узла приведены в следующей таблице:

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

Таблица 2. Результаты тестов SHOC (CUDA-версия)

В тестах *MaxFlops* и *Bandwidth* (принадлежащих уровню 0) подбирается такое вычислительное ядро, которое позволяет получить максимальные значения соответствующих характеристик. Если сравнить достигнутые значения с теоретической пиковой производительностью, то можно заметить, что они практически идентичны — в большинстве случаев разность не превосходит нескольких процентов.

Тесты *FFT* (быстрое преобразование Фурье) и *GEMM* (перемножение матриц) оценивают производительность GPU на достаточно вычислительноёмких задачах, в то время как в тестах *MD* (задача N тел) и *Reduction* (свёртка массива) узким местом является память.

При запуске OpenCL-версий тестов были получены следующие результаты:

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

Таблица 3. Результаты тестов SHOC (OpenCL-версия)

Как видно, OpenCL-тесты в сравнении с CUDA-аналогами показывают либо схожую производительность, либо в несколько раз меньшую. Однозначно объяснить

данное явление достаточно сложно — можно лишь заметить о меньшей «зрелости» технологии и специфики бенчмарка, который, возможно, ещё не достаточно оптимизирован (отметим, что первая стабильная версия появилось в декабре 2010 года).

Тесты MAGMA и CUBLAS

Библиотеки MAGMA [3] и CUBLAS являются реализациями ставшего фактически стандартом интерфейса BLAS, содержащего основные операции линейной алгебры. Обе эти библиотеки используют технологию NVidia CUDA и хорошо оптимизированы под графические процессоры. Ниже приведены результаты сравнения для четырёх операций:

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

Таблица 4. Результаты тестирования библиотек MAGMA и CUBLAS

Стоит сразу отметить, что размер матриц и векторов выбирался с целью получения максимальной производительности. Так, в тесте *GEMM* проводилось перемножение двух матриц размером 1600 на 1600. Аналогично, тест *GEMV* заключался в умножении вектора размерности 9949 на соответствующую матрицу. В тесте *GESV* решалась СЛАУ с квадратной матрицей размерностью 10112, а в тесте *SYMV* проводилось перемножение симметричной матрицы размерности 8221 на вектор и последующее сложение с другим вектором.

Если сравнивать конкурирующие библиотеки MAGMA и CUBLAS, то однозначно определить более эффективную реализацию BLAS для технологии CUDA невозможно — в зависимости от размера данных, более быстрой является то одна библиотека, то другая. Стоит заметить, что обычно MAGMA более эффективно работает на больших данных (размером около 1 гигабайта), в то время как CUBLAS чаще выигрывает при работе с данными порядка 100 мегабайт, а также на ускорителях с архитектурой Fermi.

Тесты RODINIA

Бенчмарк Rodinia [4] был разработан группой исследователей из Вирджинского университета и является сборкой вычислительноёмких CUDA- и OpenMP-ядер, созданных раннее авторами бенчмарка в рамках различных исследований. В связи с его внутренней разнородностью, для данного тестирования была переписана система замеров времени для унификации результатов. Ниже приведено время работы некоторых ядер бенчмарка (затраты на копирование данных не учитываются):

	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

Таблица 5. Результаты тестов Rodinia

Подробное описание каждого теста может быть найдено в [5]. Как видно из результатов, время работы большинства ядер составляет миллисекунды (хотя тестирование проводилось на доступных данных максимального размера). Поэтому выигрыш от использования более новых графических ускорителей незаметен — узким местом уже становится копирование памяти и накладные расходы на запуск ядер. Возможно, это является следствием того, что бенчмарк разрабатывался в 2007-2009 годах и «затачивался» под менее производительные ускорители. В любом случае, самой производительной видеокартой оказалась GeForce 480 GTX, на одном тесте даже обходящая специализированную Tesla C1060 в 2.74 раза.

Примеры из NVidia CUDA SDK

Последним тестом, проведённом в данной работе, является запуск некоторых примеров из NVidia GPU Computing SDK. Будучи разработанными сотрудниками NVidia, они достаточно хорошо оптимизированы и большинстве случаев учитывают особенности используемого поколения архитектуры GPU.

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

Таблица 5. Результаты тестирования примеров из NVidia CUDA SDK

Результаты данных тестов соответствуют синтетическим бенчмаркам — новые графические ускорители оказываются существенно быстрее предшественников. Исключением является пример Black-Scholes, при выполнении которого старые Tesla C1060 и GeForce 295 GTX обходят более мощные видеокарты, что вызвано «излишней» оптимизацией.

Результаты

Если обобщать полученные результаты, то в первую очередь стоит сравнить эффективность конкурирующих API для программирования под графические ускорители. Ниже приведено сравнение технологий NVidia CUDA и OpenCL, полученное усреднением схожих тестов из бенчмарка SHOC (одинарная точность):

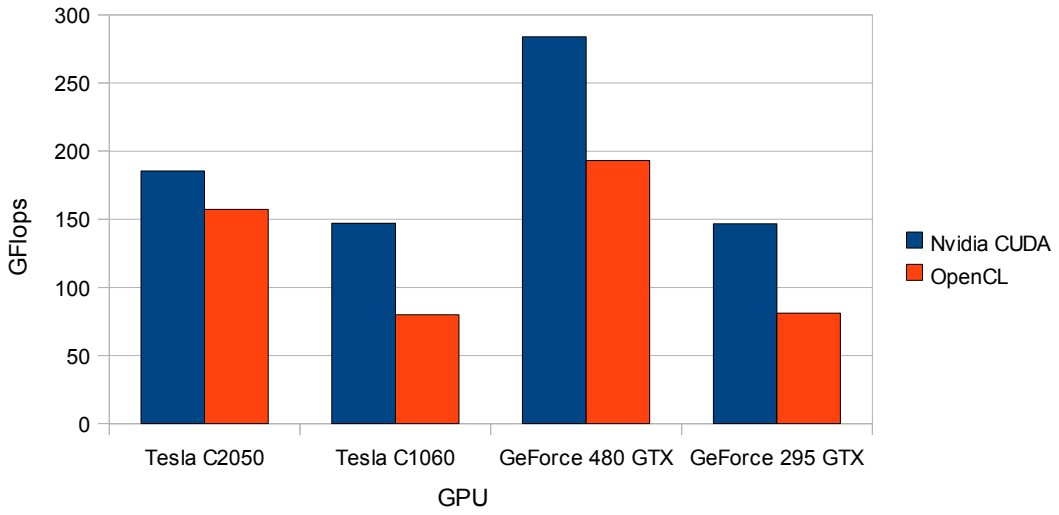

График 6. Сравнение технологий NVidia CUDA и OpenCL

Как видно из графика, NVidia CUDA стабильно обходит ещё молодой стандарт OpenCL, причём отрыв (в процентном соотношении) слабо зависит от используемой видеокарты.

Другим интересным результатом является сравнение производительности различных графических ускорителей при операциях с одинарной и двойной точностью, приведённое ниже:

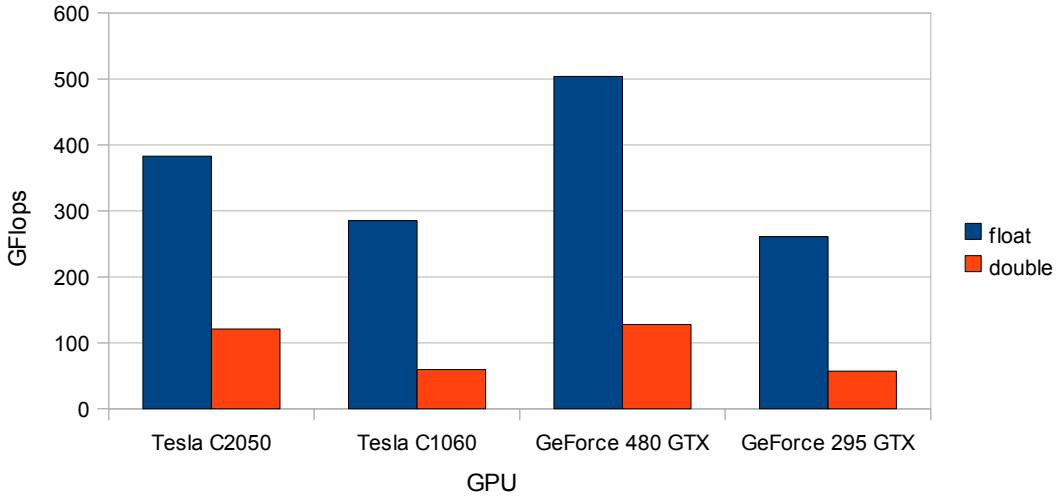

График 7. Сравнение производительности при операциях с двойной и одинарной точностью

Даже у графических ускорителей с архитектурой Fermi, для которой заявлена полноценная поддержка двойной точностью, производительность при выполнении double-операций падает более чем в 3-4 раза. Однако стоит заметить, что при этом они в разы превосходят представителей устаревающей архитектуры GT200.

Последнем сравнением является запуск различных реализаций одного и того же алгоритма (SGEMM из BLAS), содержащихся в разных библиотеках. Результаты работы при перемножении матриц сопоставимой размерности приведены ниже:

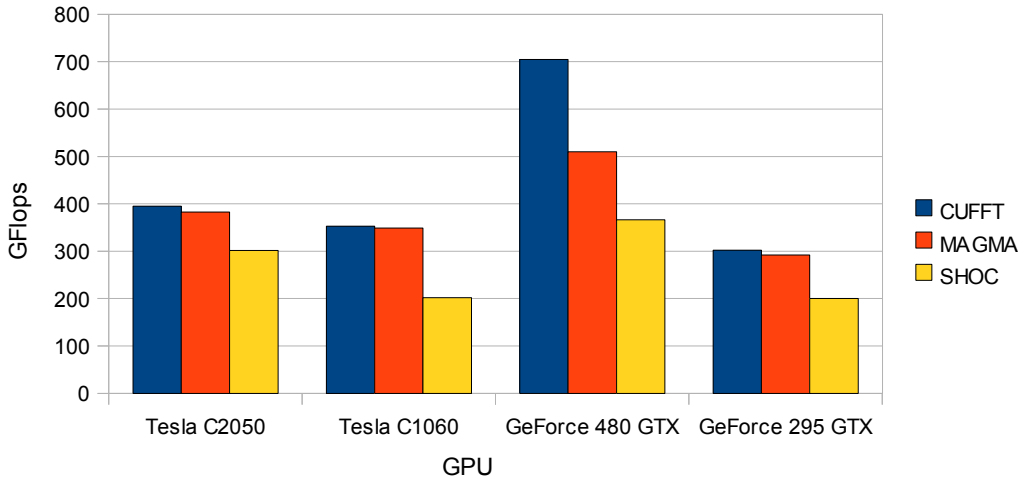

График 8. Сравнение производительности различных реализаций SGEMM

В данном случае оказалось, что производительность всех реализаций практически сопоставима, а разрыв между ними зависит больше от обрабатываемых данных.

Список литературы

1) Кривов М.А., Казеннов А.М., GeForce или Tesla? // Журнал «Суперкомпьютеры», Зима 2010, с. 42-43.
2) Danalis и др., The Scalable HeterOgeneous Computing (SHOC) Benchmark Suite // The Third Workshop on General-Purpose Computation on Graphics Processors (GPGPU 2010). Март 2010.
3) Электронный ресурс http://magma.maths.usyd.edu.au/magma.
4) S. Che и др., Rodinia: A Benchmark Suite for Heterogeneous Computing // IEEE International Symposium on Workload Characterization (IISWC), Октябрь 2009, С. 44-54,
5) Электронный ресурс https://www.cs.virginia.edu/~skadron/wiki/rodinia.