GeForce или Tesla?

Авторы Максим Кривов, Андрей Казеннов

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, представителями 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 бо**о**льшую 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 вероятность того, что для приемлемой точности памяти просто не хватит или в самый неподходящий момент что-то случится и потребуется еще один месяц вычислений.

При выборе графического ускорителя также имеет смысл обратить внимание и на видеокарты предыдущего поколения. Устаревающие ускорители можно приобрести с большой уценкой, а в некоторых задачах по вычислительным возможностям они не сильно проигрывают своим последователям. Для оценки, а стоит ли вообще переносить свои программы на графические ускорители, можно воспользоваться даже бюджетными решениями (которые есть почти в каждом компьютере), подобрав правильный «коэффициент пропорциональности». Для карты GeForce 210, используемой в данном 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 видеокарты:

MaxFlops – пытается подобрать такое вычислительное ядро, которое обеспечит максимальную производительность при выполнении операций с плавающей точкой.

FFT – выполняет одномерное Быстрое Преобразование Фурье для массива из 512 комплексных чисел.

GEMM – выполняет перемножение квадратных матриц, используя алгоритм из библиотеки BLAS.

S3D – производит моделирование процесса горения.

Memory Bandwidth – оценивает достижимую пропускную способность всех типов памяти видеокарты — глобальной, разделяемой и текстурной.

Reduction – производит суммирование элементов большого массива (свертку).

MD – решает задачу N-тел, используя алгоритм со списками соседей.

№	Модель	Пиковая производительность (float/double), GFlops	Объем памяти, GB	Количество ядер	Примерная цена, руб.	Цена за GFlops (float/double), руб.	MaxFlops 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	2x894.2/2x74.5	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/-*

Тест пропускной способности памяти, GB/s							
FFT (float/double), GB/s	GEMM (float/double), GFlops	S3D (float/double), GFlops	Global memory (coalesced)	Shared memory	Texture memory	Reduction (float/double), GB/s	MD (float/double), GB/s
192.6/65.5	318.5/85.1	45.5/25.0	151.8	270.2	94.0	85.7/89.0	93.5/68.2
102.5/52.3	**261.3/71.1**	**34.2/18.8**	**90.2**	**221.1**	**71.4**	**60.5/64.8**	**73.2/83.8**
96.2/26.9	210.5/15.6	27.7/14.7	64.0	96.7	54.6	50.2/42.2	52.2/26.1
92.7/36.4	**212.3/16.0**	**25.8/13.9**	**69.7**	**100.9**	**66.0**	**46.5/42.5**	**60.1/27.3**
101.0/29.7	233.8/17.7	31.2/16.6	70.4	109.4	60	56.5/47.6	57.1/29.5
15.3/-	*25.7*	*3.2*	*6.2*	*9.8*	*14.8*	*5.2/-*	*1.9/-*