<***title***>От G80 и до GK110</***title***>
<***hint***>

С точки зрения большинства пользователей, графические ускорители — это типа опционального свойства суперкомьютера. Они либо есть, и тогда можно запускать CUDA-программы, либо их нет. Жалко, конечно, но и головной боли будет поменьше. При этом что именно за ускорители там стоят — вопрос третий, если даже не пятый. Но если присмотреться, данный подход не совсем верен. Точнее, совсем не верен — этих самых укорителей уже аж 7 поколений, а по количеству архитектурных нововведений за последние 5 лет они спокойно обходят центральные процессоры.
</***hint***>

<***text***>

Действительно, много ли вы можете сходу вспомнить архитектурных изменений в тех же Intel Xeon'ax за произошедшие 5 лет? Причём таких, которые именно нововведения, а не улучшения? Если не очень много, то вот авторская версия: контроллер памяти и ряд других интерфейсов переехали в сам процессор, на смену SSE (128 бит) пришёл AVX (256 бит), ядра «научились» временно повышать свою тактовую частоту, добавился кэш L3. Ну, может ещё появились виртуальные ядра (по две штуки на одно реальное), но эта технология и раньше была. Итого четыре с половиной нововведения. В остальном — перед нами всё тот же Intel Xeon пятилетней давности, в котором пиковая производительность повысилась в основном за счёт увеличения частоты и количества ядер. Только обозначения поменялись, на смену «скучным» Intel Xeon W3565 пришли более модные Intel Xeon E5-2680.

В случае с GPU всё совсем иначе. Архитектура CUDA достаточно молодая, и количество «мини-революций» в ней крайне велико. Поэтому всегда стоит помнить, что, например, новая NVidia Tesla K20 — это не более «быстрый» вариант NVidia Tesla C2075, а ускоритель с сильно обновлённой архитектурой. Который на ряде задач может оказаться даже медленнее предшественника.

Собственно, цель данной статьи и заключается в обзоре всех этих «мини-революций». Причём не только последнего поколения Kepler (о чём лучше писать отдельную статью), а именно сразу всей линейки GPU от NVidia, в которых есть поддержка CUDA. Сразу стоит оговориться, что рассматриваться будут не только ускорители Tesla, но и «игровые» карты GeForce, так как в их основе зачастую лежат практически идентичные GPU. Итак, начнём!

<***subtitle***>GeForce 8800 GTX и Tesla C870<***/subtitle***>

Первой картой с поддержкой технологии CUDA стала GeForce 8800 GTX, выпущенная 8 ноября 2006 года. По тем временам это был большой шаг вперёд … да нет, настоящий прорыв! Если раньше для вычислений на GPU приходилось пользоваться не предназначенными для этого API типа OpenGL, то теперь появился собственная технология CUDA, в которой выкинуто всё лишние, связанное с графикой. И даже более того, в ней можно было напрямую, с помощью обычных указателей адресовать память GPU — роскошь, ранее недоступная.

И хотя перспективы использования новых ускорителей были очевидны, недостатков также хватало. К примеру, отсутствие поддержки операций с двойной точностью, крайне малый объём памяти (768 мегабайт), «сырость» соответствующих CUDA-библиотек и многое другое. Так, автор данных строк не раз сталкивался с ситуацией, когда вызов некой функции, подробно описанной в документации, завершался с кодом ошибки *cudaErrorNotYetImplemented*. Понятно, что для «серьёзных» вычислений подобная версия CUDA не годилась, оставаясь больше уделом энтузиастов.

Следующим шагом от NVidia стал выпуск «серверного» аналога под названием Tesla C870. И хотя отличие от игрового варианта заключалось лишь в двое большем объёме памяти (1.5 гигабайта) и более высокой надёжности, компания отчётливо заявляла о своём намерении закрепиться в HPC-сегменте.

<***subtitle***>GeForce 9800 GTX+</***subtitle***>

Очередной вехой в развитии GPU от NVidia стал выпуск карты GeForce 9800 GTX, в последствии заменённой на модификацию «с плюсиком» в конце. Каких либо кардинальных отличий от предшественника в лице GeForce 8800 GTX не наблюдалось — всё также было 128 ядер, а производительность возросла лишь на 36% за счёт увеличения тактовой частоты. Более того, данная карта даже не получила аватара в мире серверов.

Однако были у неё и существенные отличия, а именно поддержка интерфейса CUDA 1.1. Основная польза от которого, с точки зрения программиста, заключалась в появлении атомарных операций. Теперь несколько потоков вычислений могли независимо менять переменные в глобальной памяти, а отсутствие гонки данных обеспечивалось на аппаратном уровне. При этом скорость, мягко говоря, была незавидной — использование подобных атомарных операций в десятки раз замедляло

CUDA-программу. Но в любом случае, данное нововведение существенно сузило круг задач, решение которых было практически невозможно на GPU.

<***subtitle***>GeForce 285 GTX и Tesla C1060</***subtitle***>

Новая «мини-революция» произошла 17 июня 2009 года, когда начался выпуск карты GeForce 280 GTX, чуть позже обновлённой до версии GeForce 285 GTX. [Мини]революционным новый GPU оказалось сразу по нескольким направлениям — количество ядер увеличилось практически в два раза, объём памяти вырос ровно в два раза, существенно ускорился слабо-выровненный доступ к глобальный памяти, а самое главное — наконец-то добавили поддержку чисел с двойной точностью! Естественно, к этому времени специалистам NVidia удалось «дополировать» программную инфраструктуру, и всё необходимое для блицкрига в сегменте суперкомпьютеров было готово.

Для этой цели был выпущен ускоритель Tesla C1060, в который помимо всех прочих улучшений добавили поддержку контроля чётности памяти, а её объём увеличили до 3 или 6 гигабайт в зависимости от модели. И результат — на всех научных конференциях сразу же появилась GPU-секция, а в списках Top500 прописались суперкомпьютеры на базе Tesla. А что более важно, вычисления на GPU перестали считаться следствием избытка инициативности исследователя, став вполне уважаемой и серьёзной деятельностью.

<***subtitle***>GeForce 480 GTX и Tesla C2050</***subtitle***>

Чтобы закрепить успехи на поприще суперкомпьютеров, да и просто GPGPU-вычислений, 26 марта 2010 года миру была представлена новая архитектура с кодовым названием Fermi. На первый взгляд, ускоритель получился очень даже неоднозначным — пиковая производительность выросла «всего» на 30%, а количество ядер в одном мультипроцессоре увеличилось с 8 до 32. Это означает, что теперь программы должны были обладать в разы большим параллелизмом, а скорость, в принципе, осталась та же самая. Однако это исключительно на первый взгляд — столь небольшой прирост производительности объяснялся переходом к более «честным» гигафлопсам. Если раньше учитывалось, что при особых комбинациях операций ускоритель может выполнять по 3 команды за такт (что удавалось сделать крайне редко), то теперь эту возможность официально убрали, понизив порог до 2 команд за такт. И как следствие, на реальных задачах новые гигафлопсы оказывались намного «шустрее» предыдущих.

Изменения также коснулись и архитектуры GPU, получившей поддержку интерфейса CUDA 2.0. Нововведений в нём было не просто много, а очень много. Однако основным заметным улучшением стал кэш L1/L2. Теперь было необязательно мучаться с разделяемой памятью — можно просто адресовать глобальную, а «железо» само поместит нужные данные в кэш. Естественно, программа может получиться чуть-чуть медленнее, чем при ручном кэшировании, но уже не в разы (как это было ранее), а на проценты. Но главное, разрабатывать программы для GPU стало намного проще.

Для серверного сегмента был выпущен слегка «урезанный» ускоритель Tesla C2050, в который помимо традиционно большего объёма памяти добавили и лучшую поддержку чисел с двойной точностью. Так, если в игровом варианте ускорителя при работе с double скорость «проседает» в 8 раз, то при переходе на серверный аналог всего в 2 раза. То есть ровно также, как и на центральных процессорах. «Урезанность» же ускорителя проявилась в чуть меньшем количестве ядер — 448 вместо 480.

<***subtitle***>GeForce 580 GTX и Tesla M2090</***subtitle***>

А вот с данным поколением GPU всё менее радужно. Фактически, отличия от предыдущей связки GeForce 480 GTX / Tesla C2050 заключаются только в дополнительных 32 / 64 ядрах и чуть большей частоте. А во всём остальном — это тот же самый ускоритель. Видимо, решение выпуска столь нового GPU со столь небольшими изменениями было больше маркетинговым ходом, либо же все силы компании были брошены на разработку новой архитектуры под кодовым названием Kepler.

<***subtitle***>GeForce 680 GTX и Tesla K10</***subtitle***>

Именно в данном месте заканчивается экскурс в историю и начинается описание актуального на данный момент поколения ускорителей с кодовым названием Kepler. С которым нам ещё жить года два-три, пока на смену ему не придёт Maxwell. Если вы успели посмотреть на табличку с характеристиками, то могли заметить, что у нового ускорителя стало неприлично много ядер — ровно в три раза больше. А частота (и, как следствие, производительность отдельного ядра) в полтора раза меньше. Поэтому для данного GPU требуются задачи с намного большим параллелизмом, так как иначе он может оказаться даже медленнее предшественника. О чём и упоминалось в самом начале.

Что ещё интереснее, увеличение количества ядер было осуществлено исключительно за счёт перехода к новой структуре мультипроцессоров, в которых вместо 32 ядер сделали сразу 192. Ради чего их также переименовали из SM в SMX. Таким образом, теперь в алгоритмах должно быть не просто больше параллельных операций, а больше *однотипных* параллельных операций, которые нужно исхитриться «порезать» на блоки для выполнения на всех этих 192 ядрах. Что, естественно, немного усложняет процесс оптимизации CUDA-программ, но «взятый» рубеж в 3 терафлопса на одном GPU того стоит.

В качестве версии для суперкомпьютеров был выпущен ускоритель Tesla K10. Но немного нестандартный. Вместо добавления полноценной поддержки операций с двойной точностью в него добавили … ещё один GPU. Таким образом, одна Tesla K10 — это, фактически, два чуть более медленных GeForce 680 GTX. И хотя они размещёны на одной плате, с программной точки зрения там два абсолютно независимых GPU. Объяснение же оказалось вполне простым — более узкая специализация. Так, подобный ускоритель с 4.5 пиковыми терафлопсами крайне интересен для областей, где задачи ну очень параллельные, а двойная точность и не нужна. Например, для обработки множества видео-потоков. А если же без двойной точности никак, то для вас припасена и другая модель, описание которой в следующем разделе.

<***subtitle***>Tesla K20X</***subtitle***>

Легко заметить, что ранее наблюдалась чёткая закономерность — сначала выходит игровой вариант GPU, после чего на его базе делается серверная версия ускорителя. Что вполне легко объяснятся приоритетами — хоть Tesla и стоит раз в 5-6 дороже аналогичного GeForce, но последних удаётся продать явно побольше. Теперь же данная закономерность благополучно нарушена — на момент написания статьи новая Tesla K20x давно была в продаже, а про игровой (теперь уже) аналог пока ничего неизвестно. Поэтому если кто-то ещё считает вычисления на GPU «игрушками», то самое время переставать это делать, т. к. отныне графический ускоритель — это вполне себе серверный продукт, на базе которого даже выпускаются общедоступные видеокарты.

Вполне естественно, что очередная порция улучшений в первую очередь коснулась вычислительной части. При этом не только была повышена пиковая производительность как для чисел с одинарной, так и с двойной точностью, но и добавлена поддержка новых парадигм программирования. К примеру, раньше для каждого вызова некой функции на GPU нужно было сначала на центральном процессоре задать количество используемых в ней потоков вычислений и ряд параметров запуска. Теперь же эта необходимость полностью отпала. Каждый поток может самостоятельно породить другие, или даже вызвать функцию из сторонней библиотеки, в результате чего становится возможным создание GPU-программы, для работы которой центральный процессор, в принципе, уже и не нужен. Он только при запуске загрузит в память GPU нужные данные, а потом получит результат. Естественно, раньше такое было невозможно, так как на CPU лежала вся координационная работа, выполнение последовательных частей программы и вызовы всех библиотечных функций.

***

Дать подробное описание всех возможностей всех поколений GPU в рамках подобной статьи не представляется возможным, но основную идею, остаётся надеяться, проиллюстрировать удалось. Каждое поколение графических ускорителей имеет массу своих особенностей, и поэтому лучше оперировать понятиями не просто «3 GPU на узел», а хотя бы «3 GPU поколения Fermi на узел». Ну и помнить, что не все ускорители одинаково полезны.

</text>

<table>

Модель GeForce	Ближайший аналог Tesla	Год выпуска	Версия CUDA	Производительность (float / double)	Количество ядер	Частота	Объём памяти	Количество транзисторов
8800 GTX	C870	2006	1.0	518.4 / --- GFlops	128 (16 SM)	1350 MHz	768 MB	681 млн
9800 GTX+	---	2008	1.1	705.0 / --- GFlops	128 (16 SM)	1836 MHz	512/1024 MB	754 млн
285 GTX	C1060	2009	1.3	1062.7 / 88.5 GFlops	240 (30 SM)	1476 MHz	1/2 GB	1.4 млрд
480 GTX	C2050	2010	2.0	1344.9 / 168.1 GFlops	480 (15 SM)	1401 MHz	1.5 GB	3.0 млрд
580 GTX	M2090	2010	2.0	1581.0 / 197.6 GFlops	512 (16 SM)	1544 MHz	1.5/3 GB	3.0 млрд
680 GTX	K10	2012	3.0	3090.4 / 386.3 GFlops	1536 (8 SMX)	1006 MHz	2 GB	3.5 млрд
???	K20X	2012	3.5	3520.0 / 1170.0 GFlops	2688 (14 SMX)	650 MHz	6 GB	7.1 млрд

<***label***> Поколения GPU от NVidia </***label***>

Пожелание: хорошо бы в первых двух колонках цветным фоном «подсвечивать» модель GPU, для которой приводятся данные. В первых шести строчках это 1-ая колонка, в 7-ой строчке — 2-ая.

</***table***>