GPU Fermi – начало новой эпохи вычислительной визуализации

Александр Лакуша

Конкуренция – двигатель торговли. Торговля – локомотив экономики. Развитие экономики – основа прогресса. Благодаря конкуренции мы стали свидетелями стремительного совершенствования компьютерной техники. В многотомнике «Искусство схемотехники» американских авторов П. Хоровица и У. Хилла проведена блестящая аналогия: «Если бы «Боинг-747» прогрессировал с такой же скоростью, с какой прогрессирует твердотельная электроника, то он умещался бы в спичечном коробке и облетал бы без дозаправки земной шар 40 раз!» Ну, толку от такого маленького «Боинга» для пассажиров не так уж и много, а вот рост производительности компьютеров идет пользователям только на пользу! Благодаря постоянной борьбе за кошелек покупателя NVidia и AMD – основные игроки на поле графических ускорителей – вынуждены постоянно работать над усовершенствованием своих изделий.

2010 год для компании NVidia ознаменовался выходом целого семейства видеокарт, основанных на новой архитектуре NVidia Fermi. В компании буквально с нуля создали новый революционный графический процессор, который удостоился быть названным в честь выдающегося физика-ядерщика Энрико Ферми. Переход на новую архитектуру произошел во всех линейках изделий – игровых ускорителей GeForce, профессиональных видеоадаптеров Quadro и вычислительных сопроцессоров Tesla.

Новая видеокарта Tesla на базе архитектуры NVidia Fermi

Новая архитектура получила кодовое наименование GF100 и является прямым потомком архитектуры предыдущего поколения GT200, которую, в свою очередь, можно считать наследником заслуженного графического процессора G80/G92. Как и прежде, фундаментальным «строительным блоком» остается потоковый процессор (Stream Multiprocessor), теперь называющийся ядром CUDA. Базовая, полнофункциональная версия GF100 содержит уже 512 (максимум) ядер CUDA, тогда как GT200 имел на борту «всего лишь» 240, а G80/G92 – 112 ядер. При условии идентичности тактовых частот прирост «сырой» производительности может достигать почти 100%. Почему «сырой»? Итоговая производительность зависит от формы представления данных, например одинарной – 32 бит, или двойной – 64 бит, а также от точности вычислений, сложности алгоритма обработки и включенных в него опций, например активного сглаживания 8×AA. К примеру, производительность самой мощной видеокарты Quadro FX5800 составила 933 GFlops (GFlops – 10^9 операций/с) для 32-разрядных данных и «лишь» 233 GFlops для 64-разрядных данных. Ее «преемница» – новая Quadro 6000 – подняла планку до 1,8 TFlops (32 бит) (TFlops – 10^{12} операций в секунду) и 1,3 TFlops (64 бит). Компания утверждает, что смогла решить многие проблемы старой архитектуры, и это благотворно сказалось на общей производительности графической системы.

1,8 TFlops – это много или мало? Смотря с чем сравнивать. Например, 30 лет назад, в год московской Олимпиады, самым мощным суперкомпьютером считался Cray I. Изделие калифорнийской компании Cray Research занимало 150 м^2 площади, потребляло 150 кВт электрической мощности и имело пиковую производительность (на векторных операциях с разрядностью 128 бит) невероятные 0,5 GFlops. Теперь на рабочем столе у каждого компьютерного дизайнера «пасется целое стадо» Cray I!

Рассмотрим архитектуру Fermi – GF100 подробнее. Логика графического процессора разделена на кластеры обработки GPC (Graphics Processing Cluster), коих в полной конфигурации четыре. Каждый GPC оснащен собственным движком растеризации (Raster Engine), к которому подключено до четырех потоковых процессоров (SM). Каждый SM содержит 32 ядра CUDA и четыре отдельных текстурных блока (не считая логики – так называемых блоков планирования/диспетчеризации и 64 КБ буферной памяти). Кластеры GPC заполняются данными с помощью GigaThread Engine (планировщика потоков) из кадрового буфера. В отличие от GT200, где для доступа к чипам GDDR3 использовались восемь 64-разрядных контроллеров памяти (в итоге 512 бит), GF100 использует «всего» шесть 64-разрядных контроллеров (в итоге 384 бит), но для доступа уже к чипам следующего поколения GDDR5. Примененное решение привело как к удешевлению печатной платы видеоадаптера, так и к серьезному увеличению пропуск-

Суперкомпьютер начала 1980-х – Cray I

ной способности подсистемы памяти. Применение чипов с тактовой частотой 3000 МГц обеспечивает пропускную способность в 144 ГБ/с.

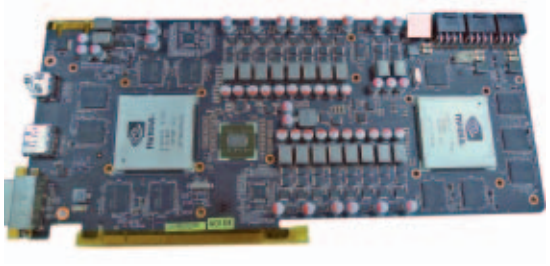
Требование поддержки DirectX 11 предписывает применение тесселяции (тесселяция (англ. tessellation) – способ в компьютерной графике, с помощью которого можно увеличить количество полигонов в полигональной трехмерной модели, используя кривые Безье. Ред.) и вынудило инженеров NVidia изменить компоновку вычислительных блоков, а именно ввести новый тип блока – ядро GPU Tesselation Engine. На каждый графический кластер GPC приходится по четыре таких блока – по одному на каждый мультипроцессор SM. GPU Tesselation Engine состоит из пяти ступеней логики с фиксированными функциями и совместно с SM выполняет последовательно этапы обработки – запрос вершин, тесселяцию, трансформацию области просмотра, настройку атрибутов и запись в память. Между этапами SM реализует просчет светотеней на вершинах и во впадинах (vertex/hull shading) и для общей геометрии (domain/geometry shading). Поддержка DirectX 11 потребовала введения в конвейер визуализации (rendering) трех новых этапов:

- hull shader – светотени корпуса: вычисление трансформаций контрольных точек;
- tessellator – тесселятор к просчету светотеней корпуса, который выдает доменные точки (domain points);
- domain shader – светотени домена, обрабатываются соответствующие точки.

Метод тесселяции, заложенный в DirectX 11, основан на применении карт смещения – монохромных текстур, используемых не для закрашивания полигона, а для изменения его геометрических свойств. Яркость (численное значение смещения) каждой точки в этой монохромной текстуре определяет высоту (отклонение) этой точки над исходной поверхностью. Ранее дизайнер должен был создавать несколько моделей с разным уровнем детализации, прописывая использование той или иной в зависимости от удаленности от переднего края сцены. Теперь используется одна модель, детализация которой определяется используемой картой смещения. Эффекты аппаратной тесселяции позволяют улучшить отображение водных поверхностей, обработку объектов в играх и самое главное – автоматически создавать множество мелких объектов кинематографического качества, не жертвуя производительностью.

По завершении всех пяти этапов обработки примитивы из движков GPU Tesselation Engine пересылаются в блоки растеризации ROP. В этих блоках отфильтровываются невидимые примитивы (так называемые обратные поверхности), затем геометрические данные преобразуются в экранные точки, которые, в свою очередь, сортируются и фильтруются по Z-координате. Конвейер GF100 состоит из шести блоков ROP, которые выводят по восемь 32-разрядных пикселов одновременно, тогда как целых восемь блоков ROP процессора GT200 выводили только по четыре пиксела за такт. Удвоение производительности позволило наряду со «старым» режимом сглаживания 8×AA ввести поддержку нового режима 32×CSAA (Coverage Sampling Anti-Aliasing), заметно улучшающего качество картинки.

Поддержка DisplayPort обеспечивает вывод высококачественного изображения на мониторы с разрешением до 2560×1600. Причем каждый графический процессор поддерживает до двух мониторов, а комбинацией устройств Quadro Plex 7000 можно получить возможность вывода изображения на восемь мониторов с общим разрешением до 36 мегапикселов.



Процессоры NVidia GF100 на видеоплате ASUS Mars

Графический чип GF100 состоит из более чем 3 млрд транзисторов (1,4 млрд у GT200) и производится по нормам 40-нм техпроцесса на предприятии тайваньской компании TSMC. GF100 Fermi послужил основой для построения целой гаммы новых видеоускорителей. Это, конечно, «первопроходцы» – линейка игровых видеокарт серии GeForce – GTX480 (480 ядер CUDA, 1536 МБ), GTX470 (448 ядер CUDA, 1280 МБ), GTX465 (352 ядра CUDA, 1024 МБ), и последовавшие за ними GTX460 (версия GF104, 336 ядер CUDA, 1024/768 МБ) и GTS450 (версия GF106, 224 ядра CUDA, 1024/512 МБ).

В серии вычислительных видеокарт Tesla также произошло обновление – чип Fermi GF100 послужил основой для моделей C2050/M2050 (352 ядра CUDA и 3 ГБ памяти) и C2070/M2070 (448 ядер CUDA и 6 ГБ памяти). Особенность микросхем графических процессоров, применяемых в платах Tesla, состоит в отключенных блоках растеризации и возможности произволь-

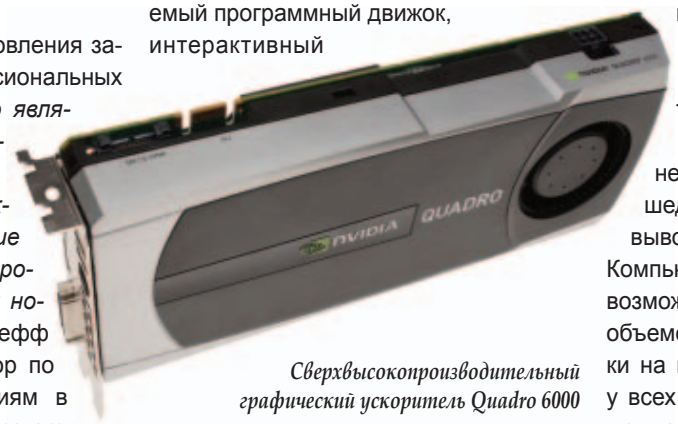
ного программирования блоков GPU Tesselation Engine, отвечающих собственно за алгоритм обработки. Данные видеокарты, как и их предшественницы, широко используются при построении настольных суперкомпьютеров для научных лабораторий. Например, с их помощью производится обработка результатов химических и биологических исследований, анализ данных, полученных в ходе обследования пациентов с помощью магниторезонансного томографа, анализ и классификация радиосигналов, принятых и записанных радиотелескопами.

И, наконец, процесс обновления затронул семейство профессиональных видеокарт Quadro. *«Quadro является выбором профессионалов №1 во всем мире уже более десяти лет. Мы также построили движки, такие как AXE, чтобы стимулировать создание приложений нового поколения,* – сказал Джефф Браун (Jeff Brown), директор по профессиональным решениям в NVidia. – *Если вы соедините эти технологии с нашей архитектурой Fermi, то получите новое семейство Quadro, которое значительно лучше всего того, что может сегодня предложить рынок».* Новая линейка профессиональных решений Quadro включает:

- Quadro Plex 7000 – настольная SLI-система из двух ускорителей Quadro 6000, подключаемая к рабочей станции посредством интерфейса PCI-E×16 и предоставляющая в распоряжение дизайнера в совокупности 896 ядер CUDA и 12 ГБ видеопамяти;
- Quadro 6000 – дискретная видеокарта высочайшего класса, пришедшая на смену модели Quadro FX5800 и располагающая 448 ядрами CUDA и 6 ГБ памяти;
- Quadro 5000 – дискретная видеокарта высокого класса, обладающая вычислительной мощностью 352 ядер CUDA и 2560 МБ памяти (достойная замена модели Quadro FX4800);
- Quadro 4000 – дискретная видеокарта среднего класса, оптимальная по соотношению цены и производительности, имеющая 256 ядер CUDA и 2048 МБ памяти, призвана заменить популярную модель Quadro FX3800;
- Quadro 5000M – графический процессор для мобильных рабочих станций, оптимизирован по уровню энергопотребления при питании в полевых условиях от аккумуляторных батарей и прочих автономных источников питания, с 320 ядрами CUDA и 2048 МБ памяти на борту.

Профессиональные видеокарты потому и профессиональные, что комплектуются дополнительным программным инструментарием – движками ускорения приложений AXE (Application Acceleration Engines): SceniX 6, CompleX, OptiX 2. Стоит немного остановиться на них.

SceniX – это свободно распространяемый программный движок, интерактивный



Сверхвысокопроизводительный графический ускоритель Quadro 6000

ускоритель ядра для множества профессиональных приложений, работающих с 3D-графикой в режиме реального времени. Он применим при автомоделировании, визуализации, имитации, в телевещании, интерактивном обучении и т. д. SceniX обеспечивает эффективное управление 3D-сценой и ее просчетом, его применение добавляет приложениям интерфейс SDI, 30-разрядное цветовое пространство, интерактивный процесс Ray Tracing.

CompleX – движок масштабирования, обеспечивает интерактивную работу с большими сценами целиком. Это достигается благодаря намного возросшему объему графической памяти на таких видеокартах, как Quadro 6000 и Quadro Plex 7000. Также CompleX поддерживает ультравысокие разрешения видео – вплоть до 8K (8192×4608) или 3D-4K (два канала по 4096×2304). В любом приложении OpenGL можно задействовать оба движка – SceniX и CompleX, получая при этом практически мгновенную реакцию на любые изменения обрабатываемого материала.

OptiX – движок ускорения трассировки лучей (Ray Tracing) с использованием архитектуры графических вычислений CUDA. То, что раньше продолжалось минуты, теперь выполняется за миллисекунды, позволяя дизайнерам интерактивно экспериментировать со светом, отражениями, преломлением и тенями в реальном масштабе времени и посредством привычного программного обеспечения. Гибкость механизма OptiX распространяется на процедурные определения и методы гибридного просчета, обеспечивая баланс скорости выполнения и качества итогового контента. Благодаря возросшему потенциалу инструмента ультрабыстрая трассировка лучей найдет свое применение в областях, где просчет ранее не использовался, например, в оптическом и акустическом дизайне, радиационных исследованиях, анализе результатов крэш-тестов и т.д.

Было бы большой ошибкой не упомянуть о стремительно вошедшей в нашу жизнь технологии вывода 3D-изображения на монитор. Компьютерный дизайнер получил все возможности NVidia 3D Vision в полном объеме, в том числе вывод 3D-картинки на несколько мониторов. Для этого у всех видеокарт Quadro имеется возможность подключения 3D-очков, правда у младшей модели данная возможность опциональна.

Поддержка работы с 30-разрядным цветом позволит вывести на экран целых 1,7 млрд цветов (по 10 бит на канал цветности, а не по 8 бит, как было доступно ранее). Именно поддержка 30-разрядного цвета обусловила применение на новых ускорителях двух разъемов DisplayPort. Правда, на сегодня существует только один монитор с поддержкой 30-разрядного цвета – HP DreamColor LP2480zx.

Наряду с высокой производительностью и мощными алгоритмами визуализации, новые графические карты Quadro поддерживают быстрые вычисления двойной точности и механизм коррекции ошибок. 64-разрядный модуль FPU (Flat Point Unit) операций с плавающей точкой дает возможность быстро выполнять любые расчеты в программах САПР или математического моделирования при создании 3D-объектов. Стало возможным применение ускорителей Quadro для выполнения приложений, требующих высокой точности обработки данных – для анализа и визуализации медицинских изображений, выполнения задач из вычислительной динамики сред, моделирования и визуализации метеорологических исследований.