Закон Гордона Мура и новый GPU NVIDIA Quadro M6000

Алекс Мастер

В этом году исполняется 50 лет первому публичному оглашению основанного на эмпирических наблюдениях так называемого закона Мура, утверждавшего, что (в современной формулировке) количество транзисторов, размещенных на кристалле микросхемы, удваивается каждые 18 месяцев. Это случилось в апреле 1965 года, примерно за три с половиной года до создания микропроцессорной империи Intel, отцом-основателем которой является Гордон Мур. Позже, в 1975 году, Мур «слегка» скорректировал текст своего закона, признав, что удвоение происходит каждые 24 месяца. Закон Мура очень быстро «подняли на щит» менеджеры и маркетологи многих компаний электронной промышленности, он очертил для них основную стратегию развития, позволявшую получать стабильный доход и технологические преимущества над конкурентами.

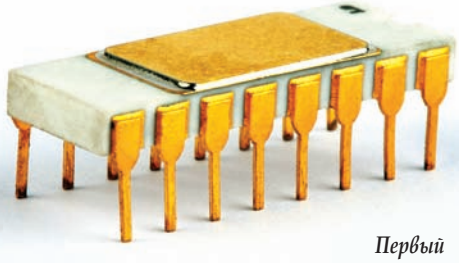
Удвоение каждые два года – много это или мало? Для ответа на этот вопрос можно провести параллели в других отраслях:

- каждые два года «аппетит» двигателей внутреннего сгорания сокращается вдвое;
- каждые два года скорость воздушных лайнеров возрастает вдвое, при этом цена на авиабилет двукратно снижается;
- каждые два года вместимость барабана стиральных машин увеличивается вдвое;
- каждые два года надои молока с одной коровы удваиваются.

Этот список абсурдов можно продолжать еще долго. Опыт развития в других отраслях подсказывает, что рано или поздно, но обязательно процесс развития достигает некоего предела – количественного или качественного. Наверное, многим знакомы такие известные математические задачи, как индийская притча о зернах на шахматной доске или римская о Цезаре и его храбром полководце. Эти и многие другие интересные математические задачи привел в своей книге «Занимательные задачи и опыты» популяризатор науки Я. И. Перельман. Упомянутые же выше примеры (и их вариации) наглядно демонстрируют высокую скорость роста экспоненциальных последовательностей. Существует такое понятие, введенное еще Рэем Курцвайлем – известным американским изобретателем и футурологом, как «вторая половина шахматной доски». Оно служит для обозначения некоторой области, после достижения которой экспоненциальный рост начинает оказывать существенное экономическое влияние, не считаться с которым уже невозможно. То есть какие-то качественные и количественные улучшения характеристик на «второй половине доски» все же происходят, но обходятся они все дороже.

Первым детищем новоиспеченной империи Мура стал процессор Intel 4004, увидевший свет в конце 1971 года. В его «недрах», созданных по нормам техпроцесса 10 мкм, трудились «в поте лица» (с тактовой частотой 0,1 МГц) целых 2250 транзисторов. Несмотря на столь скромные параметры, по одной из легенд Силиконовой долины, Intel 4004 смог «наследить» даже за пределами Солнечной системы – использовался на борту межпланетного зонда «Пионер-10».

За 44 года количество транзисторов удваивалось 22 раза, достигнув 10 млрд/кристалл. Цена же одного литографа 28-нм техпроцесса (вакуумной установки размером с письменный стол, в которой и сосредоточено большинство процессов изготовления чипа) перевалила за 5 млн долларов США и продолжает расти. Для развертывания полномасштабного производства таких установок нужна не одна сотня миллионов. В этом заключается первая проблема – технологии производства чипов не просто стремительно дорожают – они уже подошли к своему физическому пределу в 10 нм, когда размер одного транзистора соизмерим с размером атома. Но получение огромной массы «праздношатающихся по чипу» транзисторов не будет иметь смысла, если их оптимальным образом не «трудоустроить». А спроектировать новую архитектуру для чипов следующего поколения не менее накладно, чем изготавливать сами чипы.



Первый микропроцессор Intel модели 4004

В отличие от AMD и Intel, самостоятельно выпускающих микросхемы, изготовлением графических чипов для компании NVIDIA занимаются несколько OEM-партнеров, расположенных в Юго-Восточной Азии. В частности, компания TSMC (Taiwan Semiconductor Manufacturing Company), в списке заказчиков которой, кроме NVIDIA, есть AMD и Apple. Еще больше компаний-партнеров, которые на базе графических процессоров NVIDIA выпускают видеокарты. Несмотря на все сложности подобных взаимоотношений, компания на сегодняшний день является разработчиком №1 GPU для бытового и профессионального рынков. И буквально недавно всех пользователей своих изделий и решений NVIDIA порадовала новыми устройствами – видеокартой для продвинутых геймеров GeForce GTX TITAN X и графическим ускорителем для рабочих станций Quadro M6000.

Оба ускорителя созданы на 28-нм GPU «Большой Maxwell» – чипе GM200, число транзисторов в котором достигло 8 млрд, что позволило сформировать 3072 CUDA-ядра. Это максимум на сегодняшний день и на 6% больше, чем у Quadro K6000. Иерархически это выглядит так: все 3072 ядра разбиты на шесть кластеров обработки графики (GPC), каждый кластер делится на четыре многопоточных процессора (MPU), содержащих по 128 ядер CUDA. По сравнению с архитектурой Kepler разработчики оставили в каждом MPU Maxwell всего по четыре ядра CUDA, совместимых с FP64, в результате чего суммарная производительность при работе с данными двойной точности упала до 1/32 части пиковой FP32-производительности. Для чипа GK110 Kepler этот показатель был намного выше – 1/3. Таким образом, архитектура чипа GM200 была существенно оптимизирована в первую очередь под API OpenGL, чтобы

увеличить вычислительную мощность для визуализации в профессиональных приложениях, но это не лучшим образом отразилось на производительности с данными двойной точности (FP64).

И здесь кроется вторая проблема – ускорители Quadro M6000 не станут адекватной заменой графических карт прошлого поколения Quadro K6000, основанных на чипе GK110. Теперь при выборе GPU для рабочей станции пользователю придется делать непростой выбор:

- для работы с графическим (CAD/CAM) и мультимедийным 2D/3D-контентом 4/8K и выше, 2D/3D-моделирования/визуализации объектов/сред/текстур, при необходимости подключения для отображения более двух мониторов оптимален Quadro M6000;
- для выполнения большого объема параллельных вычислений (анализ сигналов и изображений, кодирование, шифрование, задачи из области физики, астрономии, химии, биологии, медицины, геодезии, метеорологии) понадобятся специальные «числодробилки» – GPU Tesla, основанные на «старой» архитектуре Kepler. А переводить ускорители Tesla на новую архитектуру Maxwell в обозримом будущем компания не собирается – для этого готовится новый чип под кодовым названием Pascal.

Теперь снова к «новорожденному». В пресс-релизе компании делается акцент на том, что новые CUDA-ядра стали примерно в 1,4 раза производительней в абсолютном выражении и в два раза эффективней в пересчете на каждый ватт затрачиваемой энергии, чем ядра K6000. Другими словами, M6000 достигла производительности 7 TFLOPS (FP32, с данными одинарной точности), тогда как значение энергопотребления и TDP остались на том же уровне.

Новейший графический процессор NVIDIA – Quadro M6000*

384-разрядная высокоскоростная (F=1653 МГц) шина соединяет чип MG200 с 12 ГБ видеопамяти стандарта GDDR5 ECC (6600 МГц, с коррекцией ошибок), обеспечивая пропускную способность до 317 ГБ/с, что на 10% выше, чем у K6000. Процесс оптимизации не мог обойти стороной и функциональные блоки финальной обработки изображения – текстурирования и растеризации. В структуре чипа GM200 заложено 256 блоков текстурирования (TMU), но задействованы только 192, что, по-видимому, лучше сочетается с 96 блоками растеризации (ROP). Для сравнения – в структуре чипа GK110 имелось целых 240 активных блоков текстурирования и всего лишь 48 блоков растеризации.

Изменения архитектуры обусловлены возросшими требованиями сегодняшних пользователей – работа с контентом сверхвысокого разрешения и многомониторными конфигурациями рабочих станций. В результате на каждый MPU работают целых 8 блоков TMU, что в сумме дает скорость наложения текстур около 192 Гтекс/с. А 96 блоков ROP обеспечили скорость прорисовки пикселей до 96 ГПк/с, чего более чем достаточно для вывода одновременно четырех потоков видео 4K (60 кадр/с) или двух потоков 6K (60 кадр/с) и даже для подключения 8K-проектора. Наличие в рабочей станции двух видеокарт Quadro M6000 в SLI-конфигурации позволит, благодаря использованию фирменной технологии Mosaic, задействовать для вывода изображения уже восемь 4K-мониторов. Для вывода электрических сигналов видеокарта располагает четырьмя DisplayPort 1.2 и одним Dual-Link DVI-I, но одновременно могут работать только четыре из них в любом сочетании.

Возвращаясь к вопросу о низкой производительности новой видеокарты при работе с данными двойной точности, необходимо заметить, что это не является катастрофой. Существует еще великое множество математических и логических задач

не вычислительного характера, в которых с успехом может применяться Quadro M6000. Например, при создании всевозможных экспертных решений, в частности, систем ИИ (Искусственного Интеллекта). Еще одним достоинством новинки стали ее, по словам представителей NVIDIA, огромные возможности для просчета изображений методом трассировки лучей. Одновременно с новой видеокартой была представлена следующая версия программного обеспечения для просчета Iray 2015 от mental images (дочерняя компания NVIDIA). В отличие от mental ray, ПО iray вовлекает в процесс просчета не только ядра центральных процессоров, но и CUDA-ядра графических ускорителей NVIDIA. Визуализатор iray создает изображения физически точно, используя алгоритм трассировки световых путей (tracing light paths). Благодаря тому, что траектории лучей просчитываются полностью, со всеми преломлениями и отражениями (то есть без допущений), данный алгоритм является физически корректным, наиболее точно описывающим поведение света. А использование вычислительных мощностей GPU позволяет кардинально экономить время. Для сравнения: типичная производительность центрального процессора рабочей станции – около 130 GFLOPS, что более чем в 50 раз меньше показателя Quadro M6000. Другими словами, процесс просчета с применением новой видеокарты NVIDIA пройдет в 50 раз быстрее.

Как отмечено выше, энергопотребление и тепловыделение новой видеокарты остались на уровне «старой» Quadro K6000, то есть не более 250 Вт. Подтверждение достаточно скромных аппетитов – наличие всего одного 8-контактного разъема для подключения дополнительного питания. Расположен он, как принято для профессиональных карт, на задней кромке печатной платы, а не сверху, как у игровой карты.

В заключение имеет смысл остановиться на показателях производительности, а если конкретнее – показать прирост по сравнению с предшественницей Quadro K6000. Здесь одним из показательных является выполнение синтетических тестов из пакета SPECviewperf, моделирующих типовую нагрузку при работе через DirectX и OpenGL:

- при выполнении скриптов в профессиональном CAD-приложении CATIA V6 R2012 перевес оказался на стороне новинки – около 21%;
- в тестовом приложении Energy (обработка и визуализация данных геологоразведки) прибавка составила 123% (!);
- выполнение скриптов Maya-2013 ускорилось на 19,5%;
- тестовое задание на визуализацию данных рентгеноскопии в приложении Medical выполнено на 47% быстрее;
- обработка тестового задания в CAM/CAD-приложении Siemens NX 8.0 показало преимущество новинки более чем в 60%;
- скрипты в CAM-приложении Solidworks-2013 SP1 отработаны почти на 9% быстрее.

Прогон через тестовую систему специализированных тестовых приложений (Benchmark) дало следующие результаты:

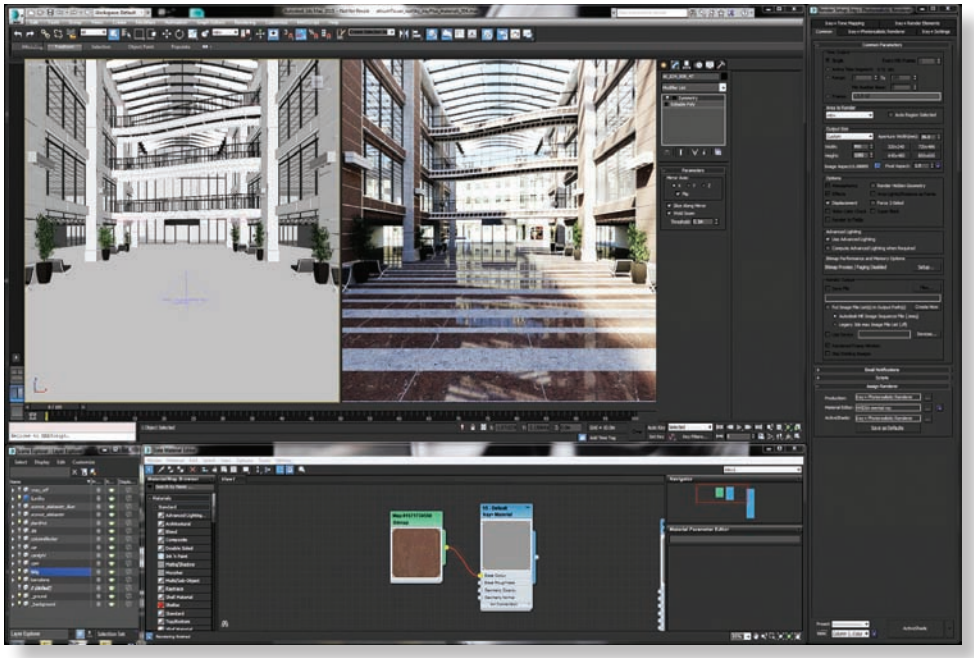
- в Autodesk 3DS Max 2015 SP3 (шейдеры DirectX 11, векторные карты) прирост общей производительности составил примерно 15…20% в зависимости от сложности визуализируемых сцен;
- в MAXON Cinema 4D R15 и R16 прибавка в скорости небольшая – около 3,5%;
- в 3D-редакторе Autodesk Maya-2015 (OpenGL) прирост составил 5…8% (чем сложнее модели, тем больше прирост).

В тесте Бенчмарк GPGPU из приложения SiSoftware Sandra-2014 активно используются наборы стандартных библиотек и инструментов OpenGL и CUDA для построения множества Мандельброта. Это нужно для тестирования производительности графической подсистемы на данных одинарной (FP32) и двойной (FP64) точности. Если в первом случае M6000 показала преимущество над K6000 около 70%, то во втором имело место отставание на 80%, что предсказуемо.

Но ситуация меняется при переходе от синтетических тестов к реальным. В частности, при прогоне тестового приложения LuxMark 3.0, созданного на основе программы просчета LuxRender, на сценах средней и высокой сложности графический ускоритель продемонстрировал рост производительности в 80…115%. Вполне вероятно, что и в фирменном визуализаторе Iray 2015 будет наблюдаться подобный результат, так как обе программы опираются на идентичные алгоритмы просчета – метод трассировки лучей.

Как показывает практика, результаты тестов по мере выхода новых версий драйверов улучшаются. В частности, в обновлении драйверов нуждаются распространенные приложения для видеомонтажа и композитинга от Adobe – Premiere Pro CC и After Effects CC. Так, финальный просчет видеоролика с наложением эффектов и с использованием движка визуализации Mercury Playback Engine в Premiere Pro CC 2014 продемонстрировал незначительное ускорение – около 3%. Есть надежда, что со временем этот результат улучшится.

Quadro M6000 вполне ожидаемо оказалась вдвое производительнее, чем Quadro K6000. Это особенно актуально в свете грядущего перехода студий на работу с изображением форматов 4K и 8K. А значит, в ближайшее время у новой видеокарты есть все шансы стать одной из самых востребованных для систем обработки медиаконтента.



Визуализация сцены в 3ds Max с применением iray

*Производителем и поставщиком профессиональных решений NVIDIA Quadro и Tesla в России является компания PNY Technologies