<***title***>Сравниваем iGPU и APU | Тестируем GPU-заменители — APU и iGPU <***title***>

<***hint***>
Ахиллесовой пятой гетерогенных суперкомпьютеров до сих пор является разнородность памяти на уровне узла. CPU и GPU сидят в своих собственных адресных пространствах и взаимодействуют друг с другом не ахти как. Изящное решение этой проблемы предложила компания NVidia, пообещав в последующих версиях своих продуктов к тысячам CUDA-ядер добавить парочку универсальных ARM-ядер для операционной системы, что позволит полностью отказаться от, по заявлениям компании, столь ненужных центральных процессоров. Ну и, как следствие, обеспечить однородность памяти. Естественно, этот сценарий не очень понравился Intel и AMD, которые предложили поступить также, но с точностью до наоборот — GPU-ядра встраивать в CPU.
<***/hint***>

<***text***>
Причём если в случае NVidia пока всё сводится к рассказам про то, какими хорошими будут их новые ускорители, то Intel и AMD уже выпустили первые версии своих, как бы это ни странно звучало, гетерогенных центральных процессоров. А Intel даже пошла чуть дальше, начав продажи новой серверной линейки Intel Xeon E3-v2, в которых помимо 4 обычных ядер приютились по 64 графических.

В рамках «тестовой лаборатории» нашего журнала было решено провести тестирование самых первых представителей этих гетерогенных центральных процессоров, которые у AMD именуются как APU (Accelerated Processing Unit), а в случае Intel обычно носят более скромное наименование CPU с iGPU (integrated GPU). Причём в обоих случаях реализована одна и та же идея — на одном кристалле с x86-совместимыми ядрами располагаются ядра графического ускорителя, совместно использующие контроллеры памяти и, по возможности, L3-кэш. Как результат, пиковая производительность одного такого процессора возрастает в 4-5 раз, а процесс его программирования становится намного проще (по сравнению с «классическими» дискретными GPU).

<***subtitle***>Тестируемые модели</***subtitle***>

На момент написания статьи серверные модели рассматриваемых гетерогенных процессоров в свободном доступе отсутствовали, поэтому было решено сравнивать старших представителей десктопных линеек. Которые, если говорить честно, являются полными братьями-близнецами серверных моделей Xeon и FirePro, но без поддержки контроля чётности, регистровой памяти и ряда прочих «наворотов».

Компанию AMD в данном тестировании представлял ускоритель AMD APU A8-3850, который был выпущен в конце 2011 года. Данная модель содержит 4 CPU-ядра а также 400 достаточно слабых GPU-ядер, что обеспечивает пиковую производительность в 526 гигафлопс (одинарная точность) при энергопотреблении в 100 ватт. Если верить пресс-релизам AMD, в конце 2012 года ему на смену придёт новая линейка под кодовым названием Trinity, ну а пока это один из самых производительных APU от AMD.

От Intel в тестировании принимал участие процессор Intel Core i7-3770, в котором также 4 CPU-ядра, но гораздо более производительных за счёт поддержки инструкций AVX и повышенной частоты, компанию которым составляют 64 графических ядра, по сравнению с AMD также работающие на более высокой частоте. Суммарно это позволяет достичь пиковой производительности в 402 гигафлопса (одинарная точность) при 77 ваттах потребляемой энергии. Также стоит отметить, что появилась данная модель только весной 2012 года.

Если сравнивать технические характеристики решений от Intel и AMD, то легко заметить их немного разное позиционирование. В случае APU, производительность графической подсистемы повышалась явно в ущерб CPU-ядрам. Поэтому если APU «дорастут» до суперкомпьютеров, то вполне может возникнуть нелёгкий выбор — либо купить систему на Opteron'ах с большим количеством классических ядер, либо на базе APU, но с урезанными CPU-ядрами, зато встроенными графическими ускорителям. В случае же Intel подобной дилеммы пока нет — iGPU является скорее бесплатным дополнением, практически не затрагивающим «транзисторный бюджет» чипа.

Чтобы программировать подобные ускорители, как легко догадаться, необходимо использовать стандарт OpenCL, реализации которого имеется у обоих производителей. Однако, если быть честными, пока данные реализации очень и очень «сырые». У AMD часть функциональности поддерживается только под ОС семейства Microsoft Windows, а другая часть заявленной функциональности нигде не работает. В Intel решили принять более кардинальное решение и … полностью отказаться от поддержки ОС Linux для своей графической части. Поэтому если хочется гетерогенности, то только Windows. Естественно, что с массовым выходом серверных моделей ситуация изменится, но пока приходится мириться с данными временными неудобствами.

<***subtitle***>Результаты бенчмарка LuxMark</***subtitle***>

Первый рассматриваемый в данной статье бенчмарк производит трассировку лучей с целью создания реалистичной картинки визуализируемого объекта. В нашем случае — комнаты. Сами расчёты ведутся как на CPU и GPU по отдельности, так и сразу на связке CPU+GPU. Что представляет особенный интерес, так как полностью соответствует идее рассматриваемых гетерогенных центральных процессоров. Однако стоит признать, что решаемая в бенчмарке задача несколько далека от реальных научных расчётов, да и плохо «ложится» на массивно-параллельные архитектуры. Если вас тоже посетила подобная мысль — ничего страшного, в следующем разделе будут приведены результаты «научного» бенчмарка.

Для тестирования использовались три сцены с разной сложностью (содержащие различное количество треугольников). При рендеринге шарика из 262 тыс. треугольников (что крайне немного — например, в каждом персонаже современного 3D-мультфильма десятки, а то и сотни миллионов треугольников) безусловным лидером оказался процессор Core i7 от Intel. Он обошёл конкурента в лице AMD APU в 1.5-2 раза как по производительности x86-совместимых ядер и графической подсистемы, так и по скорости вычислений сразу на связке CPU+GPU. Причём производительность CPU и GPU как у Intel, так и у AMD оказалась примерно одинаковой, а их совместное использование позволило получить двукратное ускорение.

В следующем тесте, где визуализируется комната из 488 тыс. треугольников, результаты не столь однозначны. На этой сцене производительность iGPU от Intel упала в разы, в результате чего его использование оказалось нецелесообразным — связка CPU+iGPU проигрывает просто CPU. У AMD же ситуация сильно не изменилась, оба типа вычислителей показали схожую производительность, но всё равно не дотянули до скорости CPU-ядер Intel (2735 против 4276 обработанных лучей в секунду).

Наконец, последний и самый «тяжёлый» тест из 2 миллионов треугольников оказался крайне «недружелюбным» для всех встроенных GPU. Графическое решение от Intel не только оказалось почти в 10 раз слабее своих x86-совместимых ядер-соседей, но ему даже не хватило памяти для завершения теста. У AMD наблюдалась похожая ситуация, но польза от встроенного GPU всё-таки была заметна — с ним вычисления шли в 1.4 раза быстрее, чем без него.

Подводя итоги по данным тестам, стоит отметить, что наблюдаемая картина полностью соответствует поведению обычных систем с дискретными GPU. Там, где количество ветвлений минимально (а именно в первом тесте — шарике), достигается максимальный выигрыш от использования графических ядер. Там, где слишком много предметов, от которых луч может отразиться и, как следствие, породить новые «ветки» вычислений (тесты с комнатами) предпочтительней оказывается центральный процессор. При этом совместное использование CPU и GPU ядер имеет смысл лишь тогда, когда разрыв между ними не является многократным. Иначе попытка всё ускорить лишь замедляет расчёты.

<***subtitle***>Результаты бенчмарка SHOC</***subtitle***>

В качестве другого бенчмарка был выбран пакет SHOC, уже не раз используемый в предыдущих статьях «тестовой лаборатории». Однако в данном случае возникла небольшая проблема, так как данный бенчмарк поддерживает только POSIX-совместимые системы, в то время как тестируемое «железо» работает исключительно под ОС семейства Microsoft Windows. Наиболее простым решением оказалось портирование данного бенчмарка на ОС Microsoft Windows 7, однако в результате данных действий часть тестов «сломалась» - из-за этого в некоторых местах стоят прочерки (тест не работает) и звёздочки (потребовалось вносить изменения).

Результаты тестирования можно разбить на три категории — (1) замеры накладных расходов, (2) решение вычислительноёмких задач и (3) проведение вычислений, узким местом в которых является работа с памятью. В первой категории присутствуют всего два теста — время компиляции OpenCL-ядер и задержка на их запуск. Наиболее интересным оказалось именно время компиляции ядер, так как в него закладывается оптимизация микропрограмм под конкретную архитектуру ускорителя. Так, у Intel подготовка кода для центрального процессора осуществляется в разы быстрее, чем для графических ядер. Что может говорить как о хорошем качестве проводимой оптимизации, так и, опять же, «сырости» реализации OpenCL для iGPU. Последующие тесты говорят скорее о втором. В случае же AMD времена, требуемые для подготовки CPU и GPU микропрограмм оказались идентичными.

Другая категория тестов — проведение «тяжёлых» вычислений. Задача первого теста (MaxFlops) заключалась в попытке «выжать» такое количество реальных гигафлопс, которое будет соответствовать заявляемому пиковому. И это практически удалось выполнить на APU от AMD, и даже перевыполнить на CPU и iGPU от Intel. Что является причиной данного стахановского поведения — технология Intel Turbo Boost, автоматически повышающая частоту при нагрузке, или же «умная» реализация OpenCL, убравшая ненужные операции, сказать сложно. Но данная особенность

стабильно наблюдалась на двух системах с iGPU. При переходе к более реальным тестам типа перемножения матриц (GEMM), выполнения быстрого преобразования Фурье (FFT), моделирования процесса тепломассопереноса (Stencil2D) и горения (S3D) производительность CPU-ядер от AMD резко падала, в результате чего их GPU-собратья в одиночестве конкурировали с монстром от Intel. Что они успешно и делали — ускоритель от Intel оказался явно более сбалансированным, но в итоге понемногу проиграл в каждом тесте.

Последняя категория тестов — работа с памятью. При попытке оценить реальную пропускную способность для памяти различных типов результаты оказались сильно «прыгающими». Так, на GPU от Intel чтение из локальной памяти происходит быстрее, чем запись. А на GPU от AMD наоборот. С другой стороны, в APU при чтении из глобальной памяти CPU и GPU ядра демонстрируют одинаковую пропускную способность, в то время как в решение от Intel наблюдается явный приоритет у CPU ядер. Поэтому, чтобы абстрагироваться от архитектурных особенностей, стоит рассмотреть результаты более жизненных тестов. В частности, при решении задачи N-тел (тест MD), выполнении свёртки массива (Reduction) и скалярного произведения больших векторов (Triad) на этот раз во всех случаях более быстрым оказался ускоритель от Intel. Однако определить наиболее предпочтительную его составляющую не представляется возможным. В двух тестах лидируют CPU ядра, а в одном — GPU.

Если обобщать результаты, то стоит отметить что в случае AMD APU польза есть только от GPU ядер, в то время как в решении от Intel попеременно лидировали то x86-совместимые ядра, то ядра графического ускорителя. На задачах, где всё упирается в вычисления, лидером оказался AMD APU A8, а на тестах для работы с памятью уже Intel Core i7 опережал своего оппонента.

<***subtitle***>Итоги</***subtitle***>

Так как тестирование проводилось на самых первых моделях гетерогенных центральных процессоров и немного «сырых» реализациях OpenCL, то вполне ожидаемо, что результаты оказались далеко неоднозначными. Но выделить ряд закономерностей можно.

Во-первых, x86-совместимые ядра от Intel в разы, а то и в десяток раз производительнее, чем CPU-составляющая от AMD APU. Таким образом, решение от Intel является более универсальным — если у вас завалялся пакет программ, который ничего не знает про встроенные GPU, то при переходе на подобную систему вы толком ничего и не теряете. В случае же AMD игнорировать графические ядра крайне опасно, так как именно в них заключается вся вычислительная мощь.

Во-вторых, производительность встроенных графических ядер обоих производителей явно не дотягивает до мощностей дискретных ускорителей от NVidia, поэтому даже при «правильном» ПО ожидать десятикратных ускорений не стоит. С другой стороны, теперь благодаря однородности памяти появляется возможность с нулевой задержкой «перекидывать» вычисления между разными типами ядер, что может крайне упростить процесс портирования существующих программ на подобные системы.

Приживутся или нет рассмотренные решения в суперкомпьютерной отрасли покажет только время. Однако потенциал для этого явно есть, так как по сравнению с обычными GPU они не только упрощают программную модель, но и могут иметь намного большую энергоэффективность.

Автором выражается благодарность сотрудникам компании Intel Валерию Черепенникову и Сергею Хвостову за предоставленный доступ к системе на базе Intel Core i7-3770.
<***/text***>