Оценка ускорения от использования трёхуровневых гетерогенных систем при решении модельных задач

М.А. Кривов[1,2], М.Н. Притула[1]

[1]ООО «ТТГ Лабс», [2]МГУ им. М.В. Ломоносова

После появления гибридных процессоров, в которых имеются как классические x86-совместимые ядра, так и встроенный программируемый графический ускоритель, стал актуальным вопрос, какой или какие вычислители стоит использовать в построенных на их базе трёхуровневых гетерогенных системах. Действительно, ситуация, когда один фрагмент программы может быть выполнен сразу на одном из трёх идейно разных вычислителях (ядра центрального процессора, встроенный или дискретный графические ускорители), хоть и является нечастой, но вполне жизненной. К примеру, серверные процессоры начального уровня Intel Xeon E3, на базе которых достаточно часто строятся одиночные вычислительные станции с GPU NVidia Tesla, как раз обладают встроенным графическим ускорителем, в результате чего подобные системы и попадают под определение трёхуровневых гетерогенных систем.

В большинстве расчётных пакетов озвученная проблема выбора решается достаточно просто — априори всегда используется только дискретные ускорители, вне зависимости от характеристик других доступных вычислителей. Очевидно, что данных подход реализуется крайне легко и, что более важно, оказывается верным в большинстве случаев, так как производительность дискретного ускорителя на порядок выше. Однако в некоторых ситуациях он может привести и к существенному понижению скорости расчётов. Например, в задачах, где требуются частые барьерные синхронизации, а объём обрабатываемых данных достаточно мал, или же параллелизм алгоритма ограничен, центральный процессор оказывается предпочтительнее дискретного ускорителя. Аналогично, встроенный графический ускоритель за счёт использования общей памяти, и, как следствие, отсутствия необходимости копирования данных по шине PCI-E, в ряде ситуаций также оказывается быстрее, чем его дискретный аналог.

Целью данной работы является поиск ответа на два вопроса, возникающих при попытке более детального рассмотрения описанных выше ситуаций, а именно — (i) какое ускорение можно получить при «правильном» выборе вычислителя, и (ii) какое ускорение может быть достигнуто при одновременном использовании всех трёх типов вычислителей. Действительно, зная ответы на эти два достаточно неочевидных вопроса, разработчик программного пакета сможет решить, стоят ли ради подобного повышения скорости расчётов проводить достаточно серьёзную модификацию программной реализации, или же достигаемый эффект оказывается крайне малым.

Ответы на два рассматриваемых вопроса, очевидно, сильно зависят от алгоритма, программной реализации и обрабатываемых данных, а также характеристик вычислительной системы. В данной работе было решено рассмотреть частный случай, взяв две модельные задачи, в одной из которых узким местом является работа с памятью, а во второй — выполнение вычислительных операций, и проведя тестирование реализаций этих двух задач на данных существенно разного размера. Для того, чтобы полученные результаты соответствовали проводимым в реальной жизни расчётам, на роль первой задачи была выбрана операция SpMV по перемножению разреженной матрицы в формате CSR на вектор, которая, например, является узким местом в набирающем популярность бенчмарке HPCG, а на роль второй — решение задачи N-тел, вариации которой достаточно часто встречаются в совершенно разных областях. В качестве же аппаратной платформы была выбрана система на базе AMD APU A10-7850K и NVidia GeForce 680 как наиболее сбалансированная в плане производительности и пропускной способности памяти каждого компонента.

Полученные результаты показали, что использование подобных возможностей рассматриваемых трёхуровневых гетерогенных архитектур действительно позволяет обеспечить ускорение порядка 1.5-1.8 раз, однако только для вычислительноёмких задач и достаточно ограниченного диапазона данных. Также оказалось, что попытка задействовать сразу все типы вычислителей практически всегда приводит лишь к замедлению скорости расчётов, что, в частности, в подготовленных авторами программных реализациях было обусловлено дополнительными накладными расходами, возникающими при осуществлении работы программы в данном режиме.

Прогнозируемый объём статьи — 7-8 стр., 1 таблица (характеристики вычислителей), 2 иллюстрации (идеальный процент распределения нагрузки по устройствам в зависимости от размера данных), 2 графика (производительность четырёх вариантов запуска программы (CPU, iGPU, GPU, CPU+iGPU+GPU) в зависимости от размера данных).