HPC-экономика: оценка стоимости решения задач гидродинамики на *GPU*-кластерах разных поколений

М.А. Кривов, М.Н. Притула (ООО "ТТГ Лабс"), А.А. Дектерёв (Институт теплофизики им. С.С. Кутателадзе)

29 марта 2016 года была опубликована последняя редакция списка самых высокопроизводительных систем СНГ. Не будем проводить глубокий анализ этого рейтинга, но акцентируем внимание на одной весьма любопытной детали – за 2015 год было установлено пять систем на базе *GPU*, в двух из которых используются ускорители, выпущенные в 2011 году и уже давно снятые с производства. Любопытной эту деталь делает тот факт, что обе системы установлены в коммерческих компаниях, название одной из которых скрыто, а второй является московское отделение *Schlumberger*. А предпочтение "новинкам" оказывают исключительно университеты и институты.

Логично предположить, что появление подобных "новых" кластеров на базе устаревших ускорителей является следствием поэтапного обновления вычислительных ресурсов, в рамках которого уже отжившие свой век серверы не списываются, а используются совместно с более новыми моделями. И интриги как таковой нет.

Однако имеется несколько "но", которые заставляют рассмотреть данный вопрос более внимательно. Во-первых, новые модели *GPU* по ряду показателей проигрывают предшественникам и, таким образом, не являются безусловными лидерами. Во-вторых, в популярном облаке *Amazon EC2* имеются узлы на базе обоих поколений ускорителей, причем стоимость аренды устаревших *GPU* даже немного выше – $1.18 против $0.72 за час. Так что можно предположить, что спрос всё-таки есть. В-третьих, до сих пор существует и активно используется множество кластеров с ускорителями 2010–2011 годов выпуска. Поэтому проблема выбора, на какой системе предпочтительнее проводить расчеты, возникает достаточно часто.

Целью данной статьи является ответить на вопрос, может ли быть экономически обосновано применение старых моделей *GPU*, или же все наблюдаемые факты объясняются просто инертностью отрасли. Естественно, ответ будет зависеть от задачи, алгоритмов и используемой программной реализации, поэтому на универсальность полученных выводов мы не претендуем, рассматривая вполне конкретный, но достаточно жизненный пример – обтекание трехмерного объекта заданной формы потоком вязкой несжимаемой жидкости.

Сравниваемые ускорители

На данный момент для промышленных расчетов в большинстве случаев используются ускорители компании *NVIDIA* двух поколений: ***Fermi*** (2010 г.) и ***Kepler*** (2012 г.), по традиции получившие свои названия в честь известных ученых. В отличие от центральных процессоров, в которых все изменения идут эволюционно, и переход на новые модели практически гарантированно означает повышение производительности, в случае графических ускорителей не всё так однозначно. Их архитектура часто претерпевает существенные изменения, и, как показывает опыт авторов, без особого труда можно найти программы, которые от подобного обновления "железа" лишь замедляются. Поэтому достаточно важно понимать, что же именно прячется за малоинформативными индексами типа *M2070* или *K40*.

В 2010 году появились ускорители *NVIDIA Tesla C2050* с новой архитектурой *Fermi*, которая позволяла решать задачи, ранее считавшиеся "неподходящими" для *GPU*. В частности, подобная универсальность обеспечивалась за счет появления системы кэшей, добавления "быстрых" атомарных операций, поддержки работы с указателями, но самое главное – новые ускорители полноценно работали с двойной точностью. Пиковая производительность в *float*- и *double*-операциях соотносилась как 2:1, что являлось сильным маркетинговым аргументом для компании *NVIDIA* – графические ускорители теперь не отличаются от обычных процессоров. И, что вполне ожидаемо, с появлением *Tesla C2050* начался настоящий бум по адаптации ПО под *GPU*, который затронул и инженерные пакеты. К примеру, именно в тот момент стали появляться первые сборки *OpenFOAM* для графических ускорителей, а компания *ANSYS* объявила о поддержке *GPU* серии *NVIDIA Tesla* в своих продуктах.

Далее, в 2012 году, на базе очередной архитектуры *Kepler* была выпущена новая линейка ускорителей *NVIDIA Tesla K20/K20x*, которая в дальнейшем пополнилась моделями *K40* и *K80*. Основной её чертой должна была стать выросшая в 4 раза пиковая производительность, однако на деле именно эта особенность оказалась слабым местом, так как повышение производительности обеспечивалось благодаря 5-кратному увеличению числа ядер при одновременном понижении их тактовой частоты. В результате этого для задач, где параллелизма недостаточно или же узким местом является работа с памятью, выигрыш от перехода на новые модели если и был, то измерялся десятками процентов, а не разами. К сожалению, именно к данному классу задач относятся практически все методы конечных разностей и конечных элементов, записанные для неструктурированных сеток, поэтому в случае инженерных пакетов декларированное многократное ускорение работы наблюдалось редко. Если же учесть тот факт, что в поколении *Kepler* поддержка двойной точности была урезана с соотношения 2:1 до 3:1, то нежелание некоторых организаций обновлять свой парк серверов, с которым неоднократно сталкивались авторы, выглядит вполне логичным.

Полностью решить озвученные проблемы призвано новое поколение ускорителей с кодовым именем *Pascal*, выпуск которого в лице *NVIDIA Tesla P100* запланирован на лето 2016 года, и которое позиционируется как очередная веха в развитии вычислений на *GPU*. Однако до момента их массового распространения (а тем более адаптации ПО под них) пройдет достаточно времени, поэтому поднимаемые в статье вопросы пока еще актуальны.

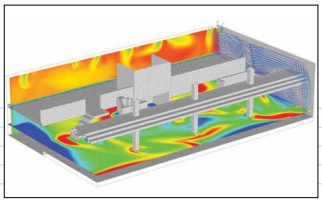

Рис. 1. Оптимизация режимов вентиляции помещений и зданий

Результаты тестирования

Прежде чем оценивать экономическую обоснованность использования того или иного поколения ускорителей, необходимо сравнить их производительность при решении приближенной к реальности расчетной задачи. В данной статье в качестве бенчмарка использовался пакет для гидродинамических расчетов ***SigmaFlow***, который с 1993 года разрабатывается коллективом Института теплофизики Сибирского отделения РАН, и который имеет версию для работы на кластерах с ускорителями *NVIDIA Tesla*. В отличие от академических проектов, этот пакет представляет собой готовый инструмент и активно используется для решения промышленных задач – например, с его помощью были выполнены работы по оптимизации процессов горения и теплообмена в топочных камерах энергетических котлов и в металлургических печах, а также осуществлен анализ нестационарных явлений в проточном тракте гидротурбин высоконапорных ГЭС. Результаты, взятые из некоторых выполненных проектов, приведены на рис. 1÷4.

В рамках тестовой задачи проводилось моделирование нестационарного течения вязкой несжимаемой жидкости при обтекании цилиндрического тела. Описывающая данный процесс система уравнений Навье-Стокса решалась методом конечного объема с использованием расщепления при помощи *SIMPLE*-подобной процедуры на совмещенных сетках. Системы линейных алгебраических уравнений, получаемые на каждом временно́м шаге, решались итерационно – трехслойным вариационным методом сопряженных невязок. Размер сетки, в зависимости от типа теста, задавался как 5, 10, 20 и 50 млн. узлов соответственно.

Тестирование проводилось на 30-ти узлах суперкомпьютеров “Ломоносов-1” (*Tesla M2070*) и “Ломоносов-2” (*Tesla K40*) Московского государственного университета в операциях одинарной и двойной точности. Чтобы оценить, как сильно влияет степень параллелизма задачи на полезную производительность, количество ускорителей варьировалось от 5-ти до 30-ти с шагом 5. Таким образом,

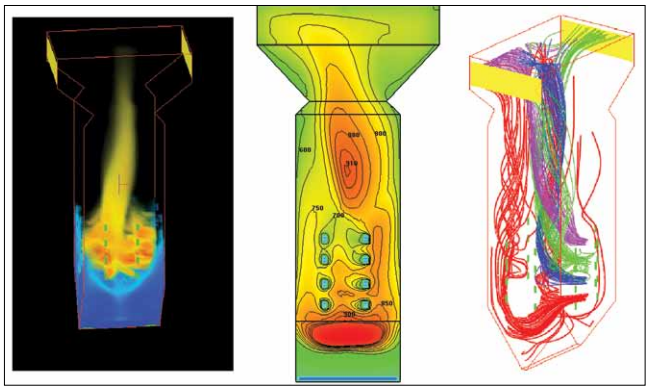

Рис. 2. Оптимизация режимов горения пылеугольного топлива в топке энергетического котла

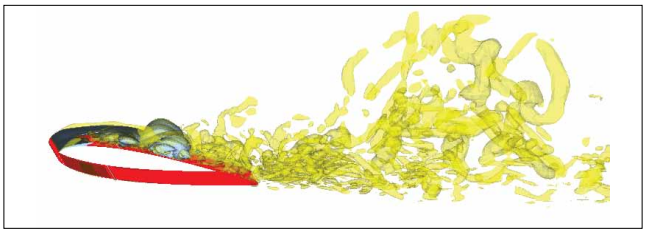

Рис. 3. Кавитационный режим течения при обтекании гидрокрыла

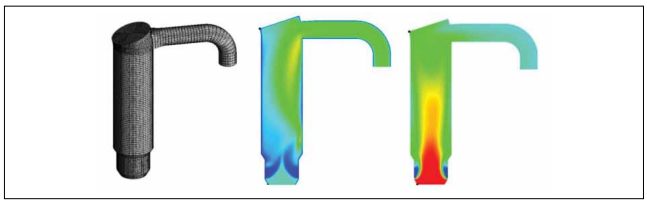

Рис. 4. Моделирование поля температур и концентрации CO в щелевой горелке

Табл. 1. Технические характеристики рассматриваемых поколений *GPU*

Архитектура	Модели серии *NVIDIA Tesla*	Пиковая производительность, *Gflops*		Число ядер	Эффективная частота, *MHz*	Объем памяти, *Gb*	Пропускная способность, *Gb/s*
		float	*Double*				
Fermi, 2010 г.	*C2050, C2070, M2070, M2090*	1030–1331	515–665	448–512	1150–1300	3–6	144–177
Kepler, 2012 г.	*K20, K20x, K40, K80*	2795–4291	932–1430	2496–2880	560–706	5–12	208–288
Pascal, 2016 г. (анонсирована)	*P100*	10 600	5300	3584	1328	16	720

ожидалось, что на 5-миллионной сетке при использовании 30-ти ускорителей будет наблюдаться недостаточность независимых операций, в то время как на 50-миллионной сетке важным станет уже объем памяти ускорителей.

Не будем проводить подробный анализ самих результатов, так как оценка масштабируемости (а тем более сравнение производительности ускорителей с центральными процессорами) – это тема для отдельной статьи. Отметим только, что на больши́х сетках один *GPU NVIDIA Tesla K40* по производительности соответствовал примерно 12÷20-ти ядрам процессоров класса *Intel Xeon E5*, а масштабируемость при увеличении числа узлов кластера с 1 до 30 была близка к линейной. Более важным для нас результатом является ускорение вычислений, достигаемое при переходе с моделей *GPU Tesla M2070* на новые *Tesla K40*, так как именно на этой основе можно оценить затраты на решение задачи (табл. 2).

Таким образом, в рамках рассматриваемой задачи изначальные опасения о низкой полезной производительности новых *GPU* по сравнению со старыми оказались беспочвенными – **выигрыш в среднем был двукратным**.

Сто́ит отметить, что всё же имеются области, в которых эффект от обновления *GPU* оказался практически незаметным. Например, при работе с двойной точностью на небольшом числе узлов кластера, расчеты ускорились всего на 30%. С другой стороны, наблюдался и всплеск производительности в операциях одинарной точности, когда *Tesla K40* оказалась быстрее своего предшественника более чем в шесть раз.

Оценка стоимости

Имея таблицу ускорения вычислений, перейдем к основному этапу, ради которого и затевалось тестирование этих двух систем, – к оценке стоимости решения рассматриваемой задачи. Для этого было предложено три схемы определения потенциальных затрат.

1 Аренда в облаке

Самым простым вариантом является именно аренда нужного количества ресурсов, когда покупателю предоставляется удаленный доступ к полностью настроенному кластеру, а оплата производится исходя из реально затраченного времени. Среди подобных сервисов наибольшую известность получило облако *Amazon EC2*, которое, согласно отчету компании *Gartner* по состоянию на октябрь 2015 года, являлось безусловным лидером в этом сегменте, набрав 4.8 баллов из 5 возможных. Однако следует отметить, что сдачей *GPU*-узлов в аренду занимаются и другие организации – например, *SoftLayer* (дочерняя компания *IBM*) или *Peer 1*. В России подобные услуги на базе своих кластеров оказывают, в частности, некоторые университеты, или же, если верить сайту, стартап *HPC HUB*.

В данной работе за ориентир были взяты расценки компании *Amazon*, так как авторы уже имеют успешный опыт запуска пакета *SigmaFlow* в их облаке, и, более того, только этот сервис позволяет выбрать поколение *GPU* – *Fermi* или *Kepler*. Однако имеющиеся у них модели ускорителей достаточно сильно отличаются от тех, на которых осуществлялось тестирование, поэтому была проведена дополнительная "нормировка" стоимости по пиковой производительности. Таким образом, один час пользования *Tesla M2070* был оценен как \$1.18, а *Tesla K40* – как \$1.35.

2 Оплата электроэнергии

В случае, если расчеты проводятся на собственном кластере в течение продолжительного времени, то основной статьей затрат становится оплата электроэнергии, которая расходуется и самими узлами, и системой охлаждения. Поскольку пиковая мощность (*TDP*) ускорителей *Tesla M2070* и *Tesla K40* отличается всего на 10 ватт (225 против 235-ти соответственно), то сравнивать суммарное энергопотребление двух рассмотренных кластеров было бы не совсем верно – вклад прочих компонент полностью нивелирует эту разницу, причем, на итоговую стоимость влияет в первую очередь КПД системы охлаждения, а не вычислительные модули. Поэтому было решено оценивать затраты только на электроэнергию, потребляемую самими *GPU*. Тем более, что если кластер изначально проектируется для работы с вполне конкретным ПО, то имеется возможность оптимизировать его, установив, например, по 4÷8 ускорителей на узел. И тогда одним из основных потребителей электроэнергии действительно могут стать ускорители.

Таким образом, если определить стоимость электроэнергии как \$0.1 за кВт·ч (цена, часто используемая как ориентир при майнинге криптовалют), а КПД блоков питания как 0.9 (стандарт *80 PLUS Platinum*), то один час работы ускорителей *Tesla M2070* и *K40* будет стоить \$0.025 и \$0.0261 соответственно.

3 Покупка ускорителей

Наконец, последний рассмотренный сценарий – приобретение ускорителей *Tesla M2070* и *Tesla K40* с целью их установки в уже имеющиеся серверы.

Табл. 2. Ускорение вычислений на *Tesla K40* относительно *Tesla M2070* (разы)

Количество *GPU*	Размер сетки, узлы							
	Одинарная точность				Двойная точность			
	5 млн.	10 млн.	20 млн.	50 млн.	5 млн.	10 млн.	20 млн.	50 млн.
5	6.3	2.0	2.0	×	1.7	1.3	1.3	×
10	5.6	2.0	2.0	2.0	2.4	2.4	1.7	×
15	5.3	2.0	1.9	2.0	2.4	2.3	1.8	2.3
20	4.9	1.9	1.9	2.0	2.6	1.8	1.9	2.4
25	4.7	1.9	1.9	2.0	2.7	1.9	1.9	2.2
30	4.4	1.9	1.9	2.0	2.9	2.0	2.0	2.4

Табл. 3. Оценка стоимости затрат на моделирование процесса (сетка из 20 млн. узлов, двойная точность)

Модель *GPU*	5 *GPU*	10 *GPU*	15 *GPU*	20 *GPU*	25 *GPU*	30 *GPU*
Стоимость аренды ресурсов для моделирования 10 секунд процесса, *USD*						
Tesla M2070	138.1	149.2	151.9	157.1	161.5	165.5
Tesla K40	118.5	98.8	95.3	96.4	95.4	94.1
Стоимость электроэнергии для моделирования 10 секунд процесса, *USD*						
Tesla M2070	2.9	3.2	3.2	3.3	3.4	3.5
Tesla K40	2.3	1.9	1.8	1.9	1.8	1.8
Стоимость покупки *GPU* в пересчете на полезную скорость моделирования, *USD* за мс/ч						
Tesla M2070	6.4	7.0	7.1	7.3	7.5	7.7
Tesla K40	24.6	20.5	19.8	20.0	19.8	19.5

Естественно, снятые с производства модели в официальной продаже отсутствуют, однако имеется множество предложений на вторичном рынке, когда уже бывшие в употреблении модули массово распродаются с большой скидкой. Найти официальные данные по сроку службы этих версий *GPU* нам не удалось, так как сайт *NVIDIA* лишь обещает "долгий срок эксплуатации продукта" и "гарантию в три года". Поэтому будем исходить из предположения, что ускорители серии *NVIDIA Tesla* устаревают быстрее, чем выходят из строя (что, в частности, подтверждается наблюдениями авторов).

Для определения конкретных диапазонов цен был использован интернет-аукцион *eBay.com*, на котором искались лоты с "почти новыми" ускорителями. Под это определение попадают ситуации, когда выставляемые на продажу платы по какой-либо причине практически не использовались и не имеют явных следов эксплуатации. Например, если верить описанию от одного из продавцов, ускорители шли в комплекте с серверами компании *Dell*, но, как оказалось, не поддерживались операционной системой, поэтому их сразу демонтировали. Если исходить из подобных критериев, то *GPU Tesla M2070* обойдется примерно в 550 *USD*, а *Tesla K40m* – порядка 2800 *USD*.

Определение себестоимости вычислений в каждом конкретном случае сводится к умножению реально затраченного времени на некую константу, зависящую от модели *GPU* и их количества. В целях уменьшения объема дублирующейся информации пересчет производительности в стоимость осуществлялся только для одного случая – операций двойной точности и сетки из 20 млн. узлов. Был выбран именно этот вариант, поскольку такого количества вершин уже достаточно для задания сложных объектов, а генерируемые матрицы еще могут уместиться в памяти пяти ускорителей. Полученные оценки для стоимости приведены в табл. 3.

Согласно результатам, использование устаревшего поколения *Fermi* может иметь смысл только в том случае, когда требуется прямо сейчас "с нуля" собрать небольшой *GPU*-кластер. Так как эти ускорители активно распродаются на вторичном рынке и еще обладают достаточной производительностью, то вариант с их покупкой позволит сэкономить до 60% бюджета. Естественно, при условии, что уже имеются серверы, в которые их можно установить. По остальным же критериям – стоимость аренды и электроэнергии – новое поколение хоть и не сильно, но стабильно обходит их.

Заключение

Подводя итоги, вернемся к упомянутому в самом начале статьи наблюдению о том, что в состав новых кластеров иногда помещают устаревшие узлы. Полученные результаты показывают, что такое решение оказывается вполне логичным. Во-первых, затраты на электроэнергию в пересчете на полезную скорость моделирования возрастают не так сильно – всего на 30÷90%. Во-вторых, соотношение производительности двух моделей ускорителей на удивление стабильно и практически всегда равно двум, в результате чего существенно облегчается балансировка нагрузки по всем узлам. А это важно для случаев, если вдруг потребуется задействовать сразу все узлы для решения одной большой задачи. В-третьих, подобный вариант позволяет заметно сэкономить, если исходить из предположения, что новая модель *Tesla P100* действительно окажется "прорывной", и на свалку отправятся уже не только узлы с *Tesla M2070*, но и с *Tesla K40*. Или, другими словами, если единственная цель использования такого разнородного кластера – это просто попытка "дожить" с минимальными затратами до нового поколения ускорителей, имея при этом необходимые для деятельности предприятия мощности.

Авторы выражают благодарность МГУ им. М.В. Ломоносова за возможность проведения тестов на вычислительных ресурсах университета, которые состоят из трех систем: "Чебышев" (60 *Tflops*), "Ломоносов-1" (1.7 П *Pflops*) и "Ломоносов-2" (2.57 *Pflops*).

Два последних кластера (занимающие, к слову, 95-е и 36-е место в мировом рейтинге суперкомпьютеров *Top500*) как раз оснащены графическими ускорителями: "Ломоносов-1" имеет 1692 платы *Tesla M2070*, а более новый "Ломоносов-2" – 1024 ускорителя *Tesla K40M*. Не лишним будет упомянуть, что на этих кластерах имеется большое количество предустановленных инженерных и научных пакетов, среди которых стóит назвать *OpenFOAM* и *FlowVision*.

Прочие детали и информацию о схеме получения доступа к ресурсам можно найти на сайте суперкомпьютерного центра МГУ (*http://parallel.ru/cluster*).