УСКОРЕНИЕ ИНЖЕНЕРНЫХ РАСЧЕТОВ В ANSYS MECHANICAL ПРИ ИСПОЛЬЗОВАНИИ ТЕХНОЛОГИИ NVIDIA MAXIMUS

ЗАО "КАДФЕМ Си-Ай-Эс":
Николай Николаевич Староверов,
инженер по внедрению и технической поддержке ANSYS Mechanical Solutions, к.т.н.
Александр Николаевич Чернов,
начальник инженерно-технического отдела

Ускорение расчетов с помощью GPU

Благодаря развитию систем инженерного анализа (CAE) сегодня инженеры во многих отраслях промышленности все чаще проводят виртуальные испытания разрабатываемых изделий. Специалисты стремятся максимально приблизить виртуальные эксперименты к реальности и получить наиболее адекватные результаты путем учета всех возможных технических деталей в расчетной модели. Растут расчетные возможности CAE-систем, в результате увеличиваются размерности задач, и возникает необходимость расширения вычислительной базы расчетных подразделений. В настоящее время в области высокопроизводительных вычислений (High Performance Computing - HPC) все более остро возникает проблема нехватки вычислительных ресурсов.

Традиционный подход к решению задач, состоящий в использовании центральных процессоров (CPU) и увеличении их производительности, уже не может справиться с необходимостью постоянного наращивания вычислительных мощностей. Технологический предел производительности для CPU оставляет единственную возможность масштабирования таких вычислительных систем - добавление десятков, сотен и даже тысяч отдельных вычислительных серверов и формирование вычислительного кластера. Этот подход требует серьезных финансовых затрат, и энергопотребление такой системы весьма существенно. Иной подход, зародившийся совсем недавно, приводит сферу HPC в эру гибридной модели вычислений, где на помощь CPU приходит графический процессор (GPU).

Предложение использовать графический процессор при расчете сложных инженерных задач явилось своего рода "глотком свежего воздуха" в сложившейся обстановке технологического тупика в производительности CPU. Возможность использования GPU в вычислениях позволила разделять сложные вычислительные задачи на тысячи небольших и решать их параллельно на ядрах графического процессора. Данная технология позволила инженерам и исследователям получать результаты численного анализа в разы быстрее. Кроме того, системы, использующие GPU, оказались более экономными с точки зрения энергопотребления, чем традиционные кластерные системы только на базе CPU.

Основное различие процессоров CPU и GPU состоит в их архитектуре. Являясь по природе массивно параллельным процессором, GPU значительно превосходит CPU в обработке большого объема однотипных данных. А CPU, являясь последовательным процессором, изначально не разрабатывался для подобного класса задач и предназначался именно для последовательных операций, таких как работа с операционной системой и организация потоков данных. Проведение вычислений с использованием GPU стало возможным благодаря созданию специфической архитектуры графических процессоров CUDA от NVIDIA, позволяющей задействовать сотни вычислительных ядер, работающих параллельно.

Передовая на сегодняшний день гибридная модель вычислений состоит в совместном использовании CPU и GPU, при этом последовательная часть кода приложения выполняется на CPU, а вся ресурсоемкая часть обработки больших объемов данных выполняется на GPU.

Технология NVIDIA Maximus

Технология CUDA для организации параллельных вычислений с использованием GPU была представлена в феврале 2007 г. компанией NVIDIA. Но прогресс не стоит на месте, и сегодня NVIDIA предлагает технологию NVIDIA Maximus, позволяющую задействовать весь потенциал процессоров CUDA на базе нескольких карт NVIDIA, работающих параллельно. Рабочие станции на основе технологии NVIDIA Maximus объединяют возможности визуализации и интерактивного проектирования графических процессоров NVIDIA Quadro с высокопроизводительной вычислительной мощностью графических процессоров NVIDIA Tesla на одной рабочей станции. Сопроцессоры Tesla при этом автоматически берут на себя выполнение ресурсоемких частей кода приложений, например, вычислений при численном моделировании или выполнение фотореалистичного рендеринга изображений. Это автоматически снимает нагрузку с CPU, позволяя ему работать в привычном режиме: ввод-вывод данных, запуск операционной системы и обеспечение многозадачности. При этом графические процессоры Quadro или Tesla производят операции, требующие высокой производительности. Конструкторы и инженеры получили возможность одновременно осуществлять проектирование в CAD-системах и проводить численный анализ в CAE-пакетах на той же рабочей станции.

Директор по стратегическому партнерству компании ANSYS, Inc., являющейся лидером рынка CAE-систем, Барбара Хатчингс (Barbara Hutchings) отмечает: *"GPU-вычисления способны значительно ускорить расчеты в программных продуктах ANSYS на рабочих станциях, а в некоторых случаях даже удвоить количество расчетов, что помогает нашим клиентам более широко использовать технологические возможности. С широкой доступностью платформы NVIDIA Maximus предприятиям теперь легче использовать программные продукты ANSYS в офисе для интерактивных и вычислительных задач"*.

Расчетные возможности продуктов ANSYS включили поддержку вычислений с участием GPU, начиная с 13-й версии программного обеспечения ANSYS в ноябре 2011 г. В бета-версии ANSYS 14.5, готовящейся к выходу на момент написания статьи, разработчики заявили о возможности проведения расчетов на базе нескольких GPU. Являясь официальным партнером ANSYS, Inc. в России, компания ЗАО "КАДФЕМ Си-Ай-Эс" протестировала работу технологии NVIDIA Maximus, выполнив серию расчетов в новой версии ANSYS Mechanical.

Maximus - это универсальная технология, предполагающая возможность балансирования нагрузки между графическими процессорами разных типов. Несмотря на то, что основное предназначение Maximus - разделение необходимых ресурсов для визуализации и CUDA-вычислений на различные процессоры (например, визуализацию на Quadro, вычисления - на Tesla), в ANSYS Mechanical из-за большой размерности задач все подключенные GPU использовались только для вычислений. Оценка еще одного преимущества технологии Maximus - возможности одновременной работы с задачами разных типов (визуализации сложных с графи-

ческой точки зрения моделей и реализации ресурсоемких вычислений) станет предметом наших дальнейших исследований. В рамках этого тестирования оценивалась работа решателей ANSYS Mechanical с участием нескольких GPU.

Тестовый стенд

Стенд для тестирования производительности расчетов с использованием технологии NVIDIA Maximus предоставлен инженерам ЗАО "КАДФЕМ Си-Ай-Эс" партнером NVIDIA в России, разработчиком и поставщиком профессиональных графических станций и высокопроизводительных решений, компанией ARBYTE. Характеристики тестового стенда приведены в таблице 1.

Таблица 1

Характеристики тестового стенда	
Модель рабочей станции	ARBYTE CADStation WS 479
CPU	Intel Core i7 3960X, 3,30 ГГц
RAM	64 Гб DDR3 1600 МГц (PC3-12800)
GPU #1	NVIDIA Quadro 6000
GPU #2	NVIDIA Tesla C2075
GPU #3	NVIDIA Tesla C2075
Твердотельный накопитель (SDD)	60 GB
Жесткий диск (HDD)	300 GB 10 K об/мин
Операционная система	Microsoft Windows 7 Профессиональная 64 bit, версия 6.1.7601 Service Pack 1
Программное обеспечение ANSYS, Inc.	ANSYS 14.5

В таблице 2 приведены характеристики использованных графических процессоров.

Таблица 2

Характеристики графических процессоров		
Характеристика	NVIDIA QUADRO 6000	NVIDIA TESLA C2075
Число ядер CUDA	448	448
Объем памяти	6 Гб GDDR5	6 Гб GDDR5
Интерфейс памяти	384 бит	384 бит
Пропускная способность памяти	144 Гб/с	144 Гб/с
Частота ядер	1,15 ГГц	1,15 ГГц
Одинарная точность	1030,4 Гфлоп	1030,4 Гфлоп
Двойная точность	515,2 ГФлоп	515,2 ГФлоп
Энергопотребление	204 Вт	225 Вт

Ускорение расчетов в ANSYS Mechanical 14.5 с помощью GPU

Сама технология использования GPU при проведении расчетов уже дает ощутимый прирост производительности. Опираясь на результаты тестирования решателей предыдущего поколения, проведенного инженерами ANSYS, Inc. и ЗАО "КАДФЕМ Си-Ай-Эс", можно сделать вывод о приросте производительности в среднем на 10...30 % и до 250 % при решении некоторого класса задач. Недостатком поддержки GPU в решателях ANSYS Mechanical всех предыдущих версий была необходимость того, чтобы задача целиком помещалась в память GPU.

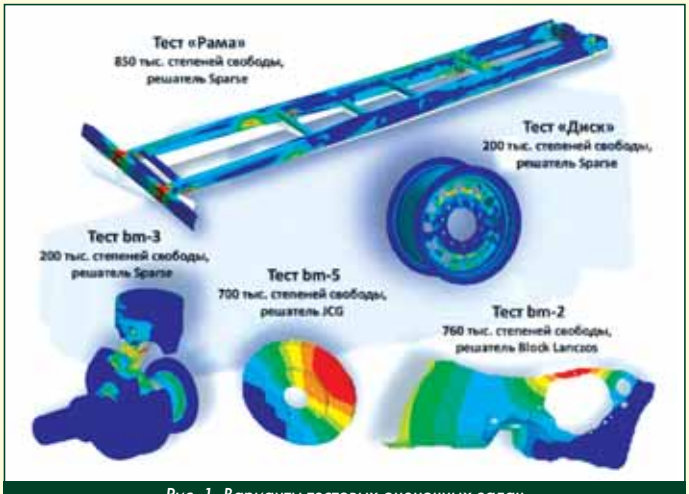

Рис. 1. Варианты тестовых оценочных задач

Поэтому в первую очередь тестирование ускорения проводилось на базе одиночного GPU для определения производительности системы в целом. Для тестирования производительности рабочей станции с одним GPU NVIDIA Quadro 6000 в ANSYS Mechanical выбраны несколько задач различной размерности: стандартные тесты производительности из набора ANSYS SP1 BENCH110 Benchmark Suite, в которых присутствуют линейные/нелинейные, стационарные/нестационарные задачи теории упругости, теории колебаний, а также отдельные задачи теории упругости и колебаний. Задачи представлены на рис. 1, а результаты тестирования приведены на рис. 2.

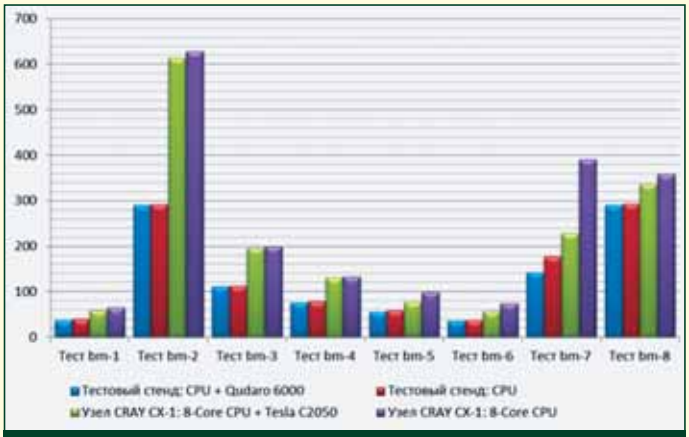

Рис. 2. Время расчета задач при тестировании GPU-ускорения ANSYS Mechanical 14.5 Preview 2, с

В целом, по результатам этих тестов получен ожидаемый результат - производительность системы с применением GPU-ускорения расчетов увеличивается на 10...30 %. Однако цель тестирования состояла в оценке работы решателей ANSYS Mechanical с технологией NVIDIA Maximus.

Подготовка оборудования

На этапе подготовки в первую очередь была настроена работа NVIDIA Maximus.

С целью обеспечения максимально быстрого обмена данными по шине PCIe графические процессоры были выставлены в следующей конфигурации:

1. GPU #1 (NVIDIA Quadro 6000) - вслотPCIe x16;
2. GPU #2 (NVIDIA Tesla C2075) - вслотPCIe x16;
3. GPU #3 (NVIDIA Tesla C2075) - вслотPCIe x8.

Посредством Maximus Configuration Utility всем 3 картам была определена возможность производить CUDA-вычисления, а NVIDIA Quadro использовалась и для вычислений, и для вывода графики.

Основное тестирование решателей ANSYS Mechanical 14.5 с NVIDIA Maximus

Тестирование возможности ускорения вычислений проводилось на трех наиболее часто используемых на практике решателях ANSYS Mechanical 14.5: Sparse, PCG и BlockLanczos.

Решатель Sparse (с разреженной матрицей) применяется для наиболее быстрого поиска решения в нелинейных расчетах, а также в линейных расчетах, в которых итерационные решатели медленно достигают сходимости (особенно при низком качестве конечно-элементной модели). Решатель PCG (методом сопряженных градиентов с предобусловленной матрицей) имеет меньший объем операций ввода/вывода данных относительно решателя Sparse и больше подходит для задач большой размерности с Solid элементами и густой сеткой. Это наиболее надежный итерационный решатель ANSYS Mechanical. Решатель BlockLanczos (по блочному методу Ланцоша) используется в динамических расчетах, проводимых в ANSYS Mechanical, для поиска собственных частот и форм колебаний конструкции.

Все задачи решались в режиме INCORE, который определяет

размещение всех необходимых решателю данных в оперативную память и использует жесткий диск исключительно для чтения исходных данных и записи окончательных и промежуточных результатов. Это режим использования памяти отличается наибольшей производительностью. Запуск тестовых задач осуществлялся из командной строки.

Поддержка нескольких GPU решателями ANSYS Mechanical

Для корректной работы решателей ANSYS Mechanical с несколькими GPU требуется соблюдение следующих условий:

1. На машине должны быть установлены один или несколько графических процессоров NVIDIA TESLA (рекомендованы карты 20 серии) или/и один NVIDIA Quadro. Если установлены и Quadro, и TESLA, то решатель ANSYS Mechanical выберет в качестве основного GPU - TESLA.

2. На операционных системах семейств Windows x64 и Linux x64 должны быть установлены драйверы актуальной версии. Для операционных систем Windows рекомендуется использование режима работы драйвера TCC (Tesla Compute Cluster).

3. Согласно лицензионной политике ANSYS, Inc. для использования GPU в расчетах необходимо наличие лицензий ANSYS HPC Pack, используемых для организации доступа к параллельным вычислениям на CPU.

4. Поддержка нескольких GPU в расчетах возможна только в режиме распределенных вычислений (Distributed ANSYS) и только в том случае, когда число запущенных процессов ANSYS Mechanical превышает число используемых GPU.

Дополнительные опции

В операционной системе можно определить следующие переменные среды:

- ANSGPU_PRINTDEVICES = 1, в этом случае решатель ANSYS Mechanical при каждом запуске будет выводить в рабочую директорию файл AnsGPUdevices.lst, в котором будут перечислены все GPU с поддержкой CUDA, доступные в системе в том приоритетном порядке, в котором их будет использовать решатель.
- ANSGPU_DEVICE = N, где N - идентификатор (ID) того GPU из списка AnsGPUdevices.lst, который решатель должен использовать. Эта переменная позволяет избежать одновременного использования одного GPU двумя и более пользователями многопользовательской среды. Следует отметить особо, что определение этой переменной среды автоматически отключает возможность использования нескольких GPU в одном расчете.

Реализована также возможность отключения коррекции ошибок памяти (ECC), которая позволяет использовать больший объем памяти GPU. Однако для обеспечения точности результатов расчетов этой возможностью пользоваться не рекомендуется.

Тестирование решателя Sparse

Для проведения тестов решателя Sparse были подготовлены однотипные статические задачи теории упругости с 10 подшагами нагружения, занимающие от 4 до 50 Гб оперативной памяти (от 220 тыс. до 1,370 млн степеней свободы). Результаты тестирования приведены на рис. 3.

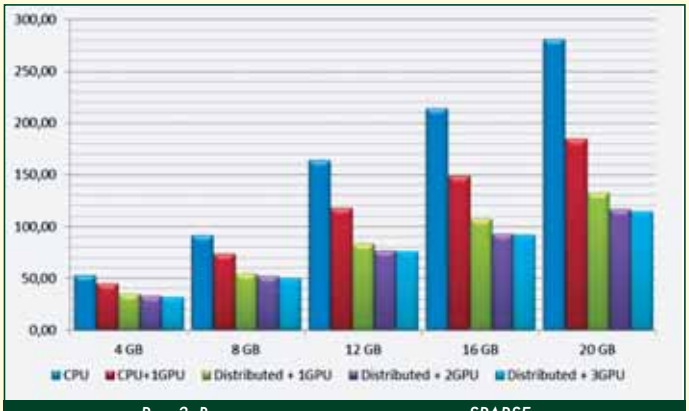

Рис. 3. Время расчета задач решателем SPARSE, с

Интересная особенность проявляется при дальнейшем росте размерности задачи. Поскольку режим распределения вычисления на несколько расчетных ядер сопряжен с дополнительными затратами вычислительной мощности на декомпозицию задачи и дальнейшее объединение данных с нескольких ядер в один результат, то для некоторого класса задач решение в режиме SMP (Shared Memory Parallel) оказывается значительно быстрее, чем решение в режиме распределенных вычислений DMP (Distributed Memory Parallel). В случае применения ускорения GPU проявляется аналогичная ситуация (рис. 4). Задача с размерностью 50 Гб в режиме DMP заняла в памяти суммарно 67 Гб, поэтому решение в режиме INCORE стало невозможным.

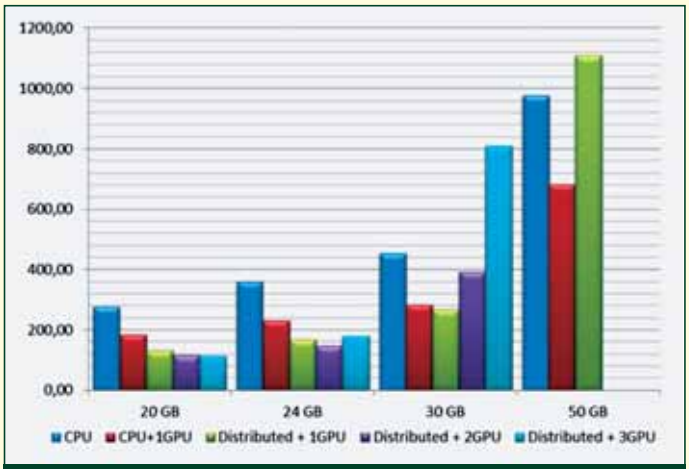

Рис. 4. Время расчета задач большой размерности решателем SPARSE, с

Загрузку GPU в процессе расчета можно наблюдать с помощью утилиты NVIDIA System Management Interfaceprogram (NVIDIA-smi.exe) из комплекта драйверов NVIDIA. С ее помощью можно записать в log-файл таблицу загрузки графических процессоров (рис. 5) с периодичностью примерно 5 с.

```
Mon Aug 06 10:50:18 2012
+------------------------------------------------------+
| NVIDIA-SMI 3.297.03   Driver Version: 297.03         |
|-------------------------------+----------------------+----------------------+
| Nb.  Name                TCC/WDDM | Bus Id        Disp.  | Volatile ECC SB / DB |
| Fan   Temp   Power Usage /Cap     | Memory Usage         | GPU Util. Compute M. |
|===============================+======================+======================|
| 0.  Tesla C2075           TCC     | 0000:02:00.0  off    |          0        0  |
|  30%   81 C  P0   125W / 225W     |  21% 1111MB / 5375MB |    67%   Default     |
|-------------------------------+----------------------+----------------------|
| 1.  Tesla C2075           TCC     | 0000:03:00.0  off    |          0        0  |
|  31%   82 C  P0    83W / 225W     |   2%   90MB / 5375MB |    74%   Default     |
|-------------------------------+----------------------+----------------------|
| 2.  Quadro 6000           WDDM    | 0000:01:00.0  On     |        Off           |
|  34%   84 C  P0    N/A /  N/A     | 100% 6115MB / 6143MB |    77%   Default     |
+-------------------------------+----------------------+----------------------+

+------------------------------------------------------+
| Compute processes:                                   |
|  GPU  PID     Process name                           |
|======================================================|
|  0.  2404     D:\ANSYS Inc\v145\ANSYS\bin\winx64\ANSYS.EXE |
|  0.  2760     D:\ANSYS Inc\v145\ANSYS\bin\winx64\ANSYS.EXE |
|  1.  1252     D:\ANSYS Inc\v145\ANSYS\bin\winx64\ANSYS.EXE |
|  2.  792      D:\ANSYS Inc\v145\ANSYS\bin\winx64\ANSYS.EXE |
+------------------------------------------------------+
```

Рис. 5. Таблица контроля загрузки GPU в log-файле NVIDIA System Management Interface program

Анализируя загрузку GPU по log-файлам, был замечен тот факт, что в основном для хранения информации использовалась GPU Tesla, а в вычислениях непосредственно участвовали столько GPU, сколько указано в параметрах запуска. При этом GPU участвовали исключительно в процессе решения задачи. На этапе подготовки задачи и формирования матриц работал только CPU, а части сформированных матриц сразу передавались в память GPU в начальный момент решения.

Тестирование решателя PCG

Для тестирования решателя PCG был подготовлен ряд прочностных задач с контактным взаимодействием, занимающих объем оперативной памяти от 4 до 50 Гб (от 190 тыс. до 1,315 млн степеней свободы). Ощутимого прироста производительности за счет использования GPU, как для решателя SPARSE, в данном случае не наблюдается. Гистограмма времени проведения расчетов показана на рис. 6.

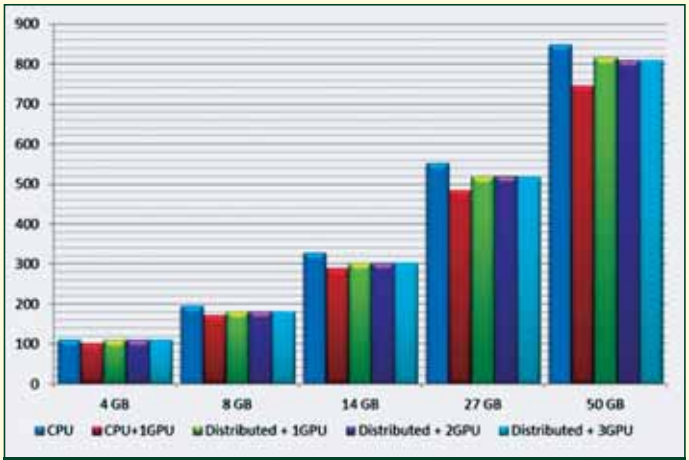

Рис. 6. Время расчета задач решателем PCG, с

Тестирование решателя BlockLanczos

Для тестирования производительности решателя BlockLanczos были подготовлены задачи поиска 20 собственных частот конструкции, занимающие от 4 до 44 Гб оперативной памяти. Для решателя BlockLanczos также ощутимого прироста производительности за счет использования GPU не наблюдалось. Гистограмма времени проведения расчетов показана на рис. 7.

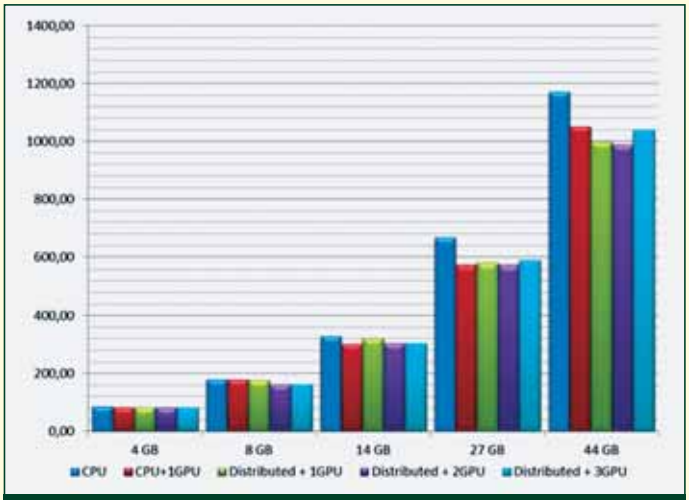

Рис. 7. Время расчета задач решателем BlockLanczos, с

Выводы

В ANSYS Mechanical 14.5 по сравнению с предыдущими версиями реализованы и доработаны следующие функциональные возможности, связанные с использованием GPU при расчетах. А именно:

1. Реализована возможность использования нескольких GPU при расчетах в режиме распределенных вычислений на локальной расчетной станции. Загрузка GPU во время тестирования была проверена с помощью специализированной утилиты NVIDIA.

2. Устранено ограничение размерности задач, решаемых с участием GPU, связанное с нехваткой графической памяти для размещения задачи. Во время тестирования решались задачи размерностью до 50 Гб, что существенно превышает суммарный объем видео памяти предоставленных графических карт.

Подводя итог, можно с уверенностью сказать, что решатели ANSYS Mechanical постоянно модифицируются - увеличивается эффективность производимых ими расчетов, снижается нагрузка на файловую подсистему. Появившаяся в версии ANSYS Mechanical 14.5 поддержка ускорения расчетов с использованием нескольких GPU по технологии NVIDIA Maximus позволит инженерам значительно сократить время расчетов существующих классов задач и повысить размерность вновь создаваемых конечно-элементных моделей.

По результатам тестирования видно, что использование нескольких графических процессоров дает существенное ускорение расчета задач теории упругости методом Sparse в режиме распределенных вычислений. Сравнение полного времени расчета тестовых задач, включающего подготовку конечно-элементных моделей и формирование файлов результатов, показало эффективный прирост производительности решателя относительно расчетов без использования GPU в 2,5 раза. При этом, чем больше размерность задачи, тем ощутимее будет вклад от использования нескольких GPU.

Для решателей PCG и BlockLanczos прирост относительной производительности наблюдается, но его величина несколько ниже, чем при решении задач методом Sparse. Для решателя PCG с ростом размерности задачи становится более очевидным сокращение времени расчета задач с использованием одной, двух и трех GPU. Однако затраты времени на декомпозицию задачи оказываются существенными, поэтому максимальная производительность этого решателя проявляется в режиме SMP с ускорением с помощью одного GPU. Максимальный прирост относительной производительности системы с несколькими GPU составил 9 %, а в режиме SMP - 14,5 %. Для решателя BlockLanczos затраты времени на декомпозицию задачи аналогичным образом возрастают с увеличением размерности задачи. Поэтому, начиная с задач определенной размерности, время расчетов задач с ростом числа используемых GPU возрастает. И чем больше используется GPU, тем больше времени затрачивается на декомпозицию и сборку задачи. Максимальный прирост относительной производительности решателя BlockLanczos с использованием нескольких GPU составил приблизительно 18,5 %, а в режиме SMP - 16,5 %.

В целом, использование вычислительных средств с подобными конфигурациями оправдывает ожидания и экономические затраты на их приобретение. Рабочие станции данного класса позволят в короткие сроки получить точные результаты расчетов, сокращая процесс разработки новой продукции.

Специалисты ЗАО "КАДФЕМ Си-Ай-Эс" выражают благодарность руководству и техническому персоналу ЗАО "АРБАЙТ КОМПЬЮТЕРЗ" за предоставленное оборудование и техническую поддержку в процессе тестирования.

ИНФОРМАЦИЯ

В этом году вышла в свет книга "Координатно-измерительные машины для контроля тел вращения" (М. 2012г. - 206с; 101 илл.; 30 табл.). Бражкин Б.С. к.т.н., Исаев Н.И., Кудинов А.А., д.т.н., Миротворский В.С., к.т.н.

Данная книга посвящена проблеме разработки эффективных методов и производительных средств контроля сложнопрофильных деталей типа тел вращения. В книге обобщён многолетний научный, инженерный и практический опыт создания отечественных специализированных координатно-измерительных машин нового класса (КИМ-ТВ), производительность которых намного выше существующих средств контроля (30-35 мин. вместо 30-36 ч при измерении, например, распредвала и с более высокой точностью. В книге представлены оригинальные математические модели расчёта размеров контролируемых параметров, а также погрешностей измерения. Указанная книга восполняет пробел в отечественной технической литературе в области координатно-измерительных машин.

Книга предназначена для ИТР метрологических, сертификационных, технологических и конструкторских служб машиностроительных и приборостроительных предприятий, научно-исследовательских и проектных организаций.

Книга будет полезна студентам, аспирантам и выпускникам технических ВУЗов, профильных кафедр. Её можно использовать для повышения квалификации и переподготовки руководящих работников и специалистов машиностроительных предприятий.

Приобрести книгу можно в РИА "Стандарты и качество", Москва, ул. Мастеркова, д. 4. (WWW.RIA-STK.RU).