Опыт использования суперкомпьютера "СКИФ Аврора" для решения научно-технических задач

А.А. Московский, М.П. Перминов, Л.Б. Соколинский, В.В. Черепенников, А.В. Шамакина

Вниманию читателей предлагается сравнительное исследование производительности ряда приложений численного моделирования на трех суперЭВМ: "СКИФ Аврора" и "СКИФ Урал", установленных в Южно-Уральском государственном университете (Челябинск), а также на кластере "*Endeavor*" компании *Intel* (*DuPont*, США). В качестве тестов были выбраны задача газовой динамики, задачи конечно-элементного анализа и задача конденсации наночастиц. Как показал анализ результатов, в большинстве случаев наилучшую производительность демонстрирует "СКИФ Аврора" – особенно в задачах, требовательных к пропускной способности подсистемы памяти.

Работа выполнена при финансовой поддержке Программы СКИФ-ГРИД, ФЦП "Научные и научно-педагогические кадры инновационной России" и РФФИ.

1. Введение

Сравнительное исследование производительности и масштабируемости различных приложений крайне важно для суперкомпьютерных центров, как с точки зрения оптимизации нагрузки на существующие машины, так и с точки зрения политики закупки новых платформ. В Южно-Уральском государственном университете (**ЮУрГУ**) установлены две машины семейства СКИФ: "СКИФ Урал" (2008 г.) и "СКИФ Аврора" (2010 г.) Для проведения исследований были выбраны не стандартные наборы тестов производительности, а несколько приложений, с которыми реально работают пользователи суперкомпьютерного центра ЮУрГУ. Такой подход позволяет получить более адекватную оценку возможностей вычислительных систем. Дополнительно нами был проведен, при помощи специализированных инструментальных средств, анализ особенностей приложений, обуславливающих характеристики их производительности.

2. Архитектура суперкомпьютера "СКИФ Аврора"

Платформа "СКИФ Аврора" изначально разрабатывалась как основа для высокопроизводительных систем большого масштаба. Целый ряд технических решений, использованных в СКИФ ряда 4, сдвигает баланс свойств в сторону специализации для применения именно в суперкомпьютерах. Подробно характеристики решения рассмотрены в работе [1]. Установка "СКИФ Аврора" в Южно-Уральском государственном университете является первым пилотным проектом по развертыванию системы такого класса. В данном разделе кратко описываются особенности системы с учетом её конфигурации системы в ЮУрГУ.

С учетом ресурсов систем охлаждения и бесперебойного электропитания проект позволяет установить до 8 вычислительных шасси "СКИФ Аврора". Каждое шасси содержит 64 двухпроцессорных узла с четырехъядерными процессорами *Intel Xeon X5570* (*Nehalem*), с рабочей частотой 2.93 *GHz*. Таким образом, в рамках одного монтажного шкафа удалось собрать 2048 процессорных ядер. Максимальная теоретическая производительность системы, состоящей из одного шкафа, равна 24 *TFLOPS*.

Рис. 1. Шкаф вычислителя "СКИФ Аврора"

2.1 Вычислительная часть

Высокая плотность упаковки процессоров в вычислителе диктует необходимость применения жидкостной системы охлаждения. Вычислительные узлы выполнены в виде печатных плат, с интегрированными на материнской плате коммуникационными, сервисными микросхемами, модулями памяти. Тестирование плат проводится на заводе-изготовителе, что уменьшает число отказов компонент при инсталляции и первичной настройке системы. Каждый узел-плата накрыт плотно прилегающей пластиной охлаждения. Пластины охлаждения оснащены быстроразъемными муфтами, что позволяет демонтировать отдельный вычислительный узел без демонтажа системы охлаждения корзины (шасси) в целом.

Каждый узел оснащен твердотельным накопителем информации объемом 80 *Gb*. Использование твердотельных накопителей имеет целью повысить надежность вычислителя, так как отказы шпиндельных дисковых накопителей дают львиную долю в списке причин отказов узлов кластерных установок и вычислительных ферм.

2.2 Коммуникационные сети

Ключевым компонентом любого суперкомпьютера является его коммуникационная среда. У "СКИФ Аврора" суммарная пропускная способность канала, учитывая как системную, так и вспомогательную сеть, достигает 100 *Gbit/s*.

Если во вспомогательной сети используются стандартные решения, то системная сеть является оригинальной разработкой. Она имеет топологию трехмерного тора, маршрутизаторы сети реализованы на уровне адаптеров. Суммарная пропускная способность в пересчете на один узел составляет 60 *Gbit/s*. Важной особенностью сети является то, что можно обойтись без дополнительного оборудования (маршрутизаторов) и задействовать при монтаже кабели одинаковой длины, вне зависимости от размера системы. Соединения на уровне половины шасси (корзины) выполнены на соединительной плате. Трехмерная организация сети позволяет легче распределить задачи между узлами кластера при моделировании объектов реального мира (трехмерных) и распараллеливании методом декомпозиции области. Для системной сети реализован интерфейс *MPI* (*Message Passing Interface*) на основе *MPICH2*, удовлетворяющий спецификации версии *MPI 2.0*.

Вспомогательная сеть – *Infiniband QDR* (40 *Gbit/s*) с полной бисекционной пропускной способностью. Адаптеры сети интегрированы на платах-узлах. Маршрутизаторы первого уровня интегрированы на уровне корзин (шасси) на так называемых "корневых платах". Соединения между узлами и маршрутизатором первого уровня выполнены на соединительной плате (*backplane*), что существенно уменьшает количество кабелей *Infiniband*, подключаемых вручную при установке системы. Поскольку маршрутизаторы первого уровня в системе уже имеются, на втором уровне сети можно использовать относительно недорогие 36-портовые маршрутизаторы – количество кабелей *Infiniband* и их длина от этого не меняются.

2.3 Подсистема мониторинга и управления

Подсистема мониторинга и управления обеспечивает надежное выполнение всех функций по удаленному обслуживанию установки, за исключением функций, требующих физических манипуляций. Подсистема использует как возможности стандартных *IPMI*-средств мониторинга (*IPMI* – интерфейс интеллектуального управления платформой), так и оригинальную разработку – сеть *Servnet*.

Компоненты *Servnet* присутствуют во всех основных модулях "СКИФ Аврора":

- на уровне узлов интегрированы контроллер и датчики температуры и влажности;
- на уровне "корневой" платы интегрированы датчики и контроллер для управления вентиляторами;
- на плате блока питания интегрированы датчики напряжения и контроллер;
- соединительная плата обеспечивает связь сети *Servnet* на уровне половины шасси.

Отличительной особенностью подсистемы *Servnet* является возможность осуществления мониторинга даже в случае полного отключения электропитания всех основных систем, так как она имеет автономное питание.

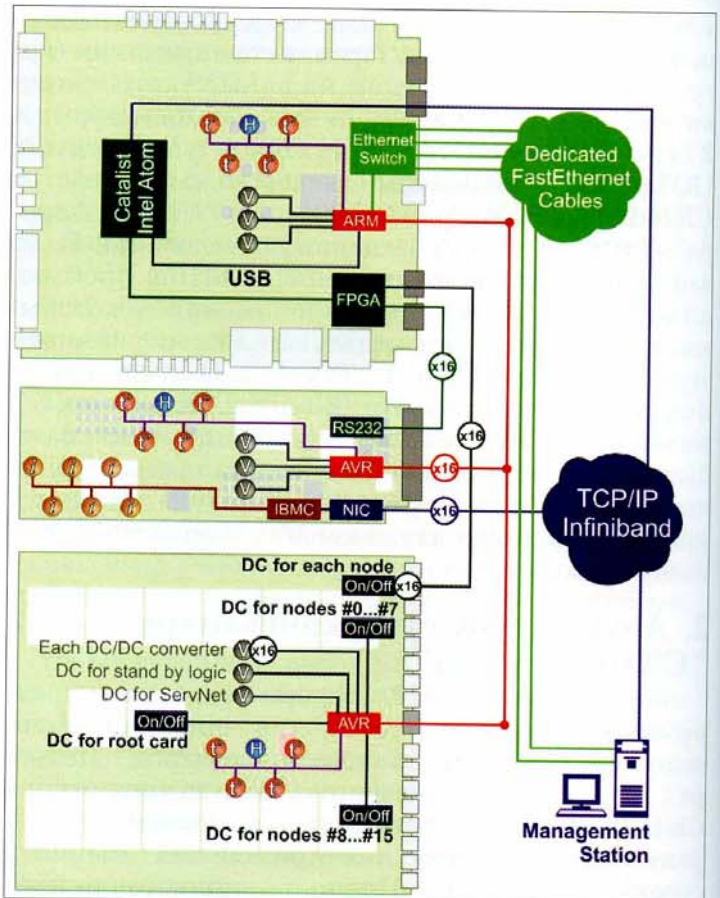

Рис. 2. Сети системы управления и мониторинга

"Корневые" платы играют важную роль в системе управления установкой. Программное обеспечение, которое работает на корневой плате, позволяет отключать и включать электропитание отдельных узлов, осуществлять мониторинг характеристик системы во время работы, а также вывод информации на сенсорные дисплеи, установленные в торцах шасси.

Программные средства мониторинга интегрируют информацию из различных источников, включая подсистемы электропитания, охлаждения, хранения данных, отображает и хранит архив данных. Поскольку установка "СКИФ Аврора" носит экспериментальный характер, под нужды управления и мониторинга выделен отдельный сервер.

2.4 Подсистема электропитания

Питание вычислителя "СКИФ Аврора" осуществляется постоянным током с напряжением 48 вольт. За счет использования постоянного тока подсистема бесперебойного электроснабжения стала более простой: она содержит лишь выпрямитель и аккумуляторные батареи. Преобразователь постоянного тока в переменный оказывается здесь не нужен.

Для сервера мониторинга предусмотрено дополнительное резервирование питания, позволяющее обеспечить его автономную работу в течение полутора часов. Таким образом, система мониторинга вполне в состоянии исполнять роль "черного ящика" вычислительной системы.

2.5 Подсистема хранения данных

Подсистема хранения данных реализована на основе параллельной файловой системы *Lustre*. Общий объем запоминаемой информации – более 50 терабайт. Теоретически эта подсистема должна обеспечивать производительность более 4000 операций ввода-вывода (*IOPS*) и пропускную способность более 500 *Mb/s*. Узлы вычислителя имеют доступ к хранилищу данных через вспомогательную сеть – *Infiniband QDR*.

3. Описание задач

Для изучения эффективности выполнения приложений на вычислителе "СКИФ Аврора" группа сотрудников ЮУрГУ отобрала несколько задач, которые затем анализировались с позиции их масштабирования и оптимизации специалистами компании "Интел" (Нижний Новгород), работающими с *HPC*-проектами. Исследовались задачи инженерного моделирования и анализа, решаемые с помощью стандартных инженерных пакетов, а также программный комплекс для моделирования процессов формирования металлических наночастиц методом газофазной конденсации, реализованный на языке *FORTRAN* с использованием библиотеки *MPI*. Ниже представлены описания приложений.

3.1 Задача вычислительной гидродинамики для тонких турбулентных слоев в щелевых уплотнениях питательных насосов электрических станций

Надежность питательного насоса определяется его вибрационным состоянием. Основным источником вибрации является неуравновешенный ротор, динамика которого в значительной мере зависит от упругих, демпфирующих и инерционных свойств турбулентной жидкости, дросселируемой в щелевых уплотнениях [2].

Щелевые уплотнения характеризуются малым зазором (0.1÷0.5 мм) по сравнению с линейными размерами (для цилиндрических уплотнений: длина ~200 мм, диаметр ~200 мм; для радиальных – внутренний радиус ~140 мм, длина ~40 мм), а также наличием перекоса и эксцентриситета.

Традиционно при расчетах гидродинамики тонких турбулентных слоев в щелевых уплотнениях используются укороченные уравнения Навье-Стокса (уравнения тонкого слоя), которые принципиально не позволяют определить падение давления на входном участке тонкой щели.

Применение численных методов расчета полных уравнений Навье-Стокса с Рейнольдсовым осреднением позволяет в общем виде решить задачу формирования тонкого слоя на начальном участке и течения жидкости в щели при нестационарном движении твердой стенки. Определение гидродинамических сил в тонких слоях щелевых уплотнений мощных питательных насосов требует решения системы уравнений с числом неизвестных порядка 50÷100 млн. Решать подобные задачи можно только с помощью высокопроизводительных вычислительных систем и мощных пакетов ***CFD*** (*Computational fluid dynamics*).

3.2 Деформирование и разрушение тканевых бронежилетов при локальных ударах

Основной задачей при проектировании бронежилетов является минимизация их массы при сохранении заданного уровня защиты. Проверка качества бронежилета, не находящегося в контакте с защищаемым объектом, проводится с определением баллистического предела. Если же бронежилет контактирует с защищаемым объектом (тело человека), то в этом случае подход иной: существует критерий определения тупой травмы, который применяется для сравнения бронежилетов различных классов [3].

Для имитации человеческого тела в экспериментах используют либо технический пластилин (степень травмирования тела при этом оценить довольно сложно), либо дорогостоящие экспериментальные модели грудной клетки. Экспериментально-аналитический путь оптимизации конструкции многослойных тканевых преград позволяет достаточно быстро определить оптимальное соотношение параметров для фиксированного воздействия – формы *индентора* (особо твердый наконечник прибора для измерения твердости) и

скорости нагружения; однако этот метод весьма затратный.

Чисто аналитических моделей, точно описывающих процесс динамического взаимодействия пули и бронежилета с учетом разрушения, на данный момент не существует. Получить такие модели, по всей видимости, невозможно из-за сложности физических явлений (большие перемещения, скольжение, фрикционные контакты, повышение температуры). Для того чтобы учесть их, надо учитывать структуру баллистической ткани.

Вычислительные возможности кластеров позволяют решать сложные контактные задачи, в которых нельзя использовать механику сплошной среды. Полученные результаты и методы исследования сопротивления тканевых преград ударам (выстрелам из огнестрельного оружия) находят применение при разработке новых средств защиты тела человека и значительно сокращают этап предварительной оценки служебных свойств такого рода изделий.

3.3 Моделирование поведения грудной клетки как механического объекта при локальных ударах

При разработке персональной защитной брони минимальной массы необходимо иметь представление о механизме повреждений, которые возникают в теле человека при локальных ударных воздействиях. Поэтому учеными разных стран ведется работа по созданию теоретических и экспериментальных моделей человеческого тела, которые в точности повторяют его форму и обладают такими же свойствами. Так, учеными Вашингтонского университета имени Джона Хопкинса (США) были созданы конечно-элементная и экспериментальная модели грудной клетки человека: были построены ребра, грудина, хрящи, позвоночник, сердце, легкие, печень, желудок, мышцы и кожа. Но если значения ускорений, полученные экспериментально и с помощью расчетов, близки, то отличие давлений остается существенным. Разработки теоретических и экспериментальных моделей грудной клетки человека активно продолжаются, однако каких-либо достоверных результатов на их основе получить пока не удалось. К тому же, были созданы только модели деформирования тела человека без учета степени травмирования [3].

Чтобы использовать при проектировании бронежилетов численную модель грудной клетки человека, надо знать механические свойства всех её элементов. Идентификацию параметров грудной клетки можно провести, сопоставив экспериментальные и расчетные перемещения при статическом нагружении, а также ускорения и спектры собственных частот колебаний при динамическом нагружении. При этом динамическое нагружение грудной клетки реального человека должно быть низкоскоростным, чтобы не нанести ему травм.

Экспериментальные модели грудной клетки имеют высокую стоимость, поэтому численное решение данной задачи является актуальной проблемой. При численном исследовании задачи возможно оценить степень травмирования грудной клетки.

3.4 Деформационные изменения структуры трикотажных полотен на различных участках фигуры человека

Сегодня изделия из трикотажного полотна широко распространены, поэтому актуальным является вопрос быстрого и качественного проектирования новых моделей одежды. Трикотажные изделия значительно растягиваются при эксплуатации, причем не одинаково на разных участках тела человека. Ситуация усложняется наличием швов различного вида (стачные, окантовочные, в подгибку с открытым срезом), которые имеют другие механические свойства. Поэтому при разработке трикотажных изделий некорректно использовать геометрический метод [4].

В настоящее время для изучения поведения тканей используют параллельные алгоритмы. Проектирование с помощью суперкомпьютеров позволяет значительно сократить материальные затраты и время на разработку нового изделия. В рамках виртуальной модели можно легко менять различные параметры: механические свойства ткани и швов, геометрию тела человека и изделия.

3.5 Моделирование процессов газофазной конденсации металлических наночастиц

В производстве микро- и наночастиц различных веществ часто используется метод самосборки частиц при их конденсации в пересыщенном паре в атмосфере инертного газа. При этом для дальнейшего использования полученного наноразмерного порошка надо соблюдать определенные требования в отношении размера частиц. Для этого необходимо задать определенный температурный режим в рабочей камере реактора, давление и вид инертного газа, геометрию установки, а также длительность производственного цикла. В настоящее время все эти параметры подбираются экспериментально, методом проб и ошибок. В таких условиях очень сложно осуществлять управление технологическим процессом и прогнозировать выходное распределение частиц по размерам. Разрабатываемые математическая модель и программный комплекс предназначены для решения данных задач [5].

Обычная схема формирования металлических наночастиц конденсацией из газовой фазы выглядит следующим образом: в камеру с охлаждаемыми стенками накачивается инертный газ, вблизи дна камеры помещается нагреваемый сосуд с кипящим жидким металлом, который служит источником атомов металла – мономеров. Испаряясь с поверхности жидкости, металлический

некоторые ядра простаивают в ожидании необходимых данных из памяти. В то же время, если для "СКИФ Урал" насыщение масштабируемости происходит довольно быстро (10.5 *Gb/s* < 20 *Gb/s*), то на "СКИФ Аврора" (38 *Gb/s* > 20 *Gb/s*) и "*Endeavor*" (42.5 *Gb/s* > 20 *Gb/s*) производительность продолжает расти с ростом числа задействованных ядер.

4.2 Деформирование и разрушение тканевых бронежилетов при локальных ударах

Задача деформирования и разрушения тканевых бронежилетов при локальных ударах рассчитывалась для пакета размером 5×5 см из пяти слоев баллистических тканей.

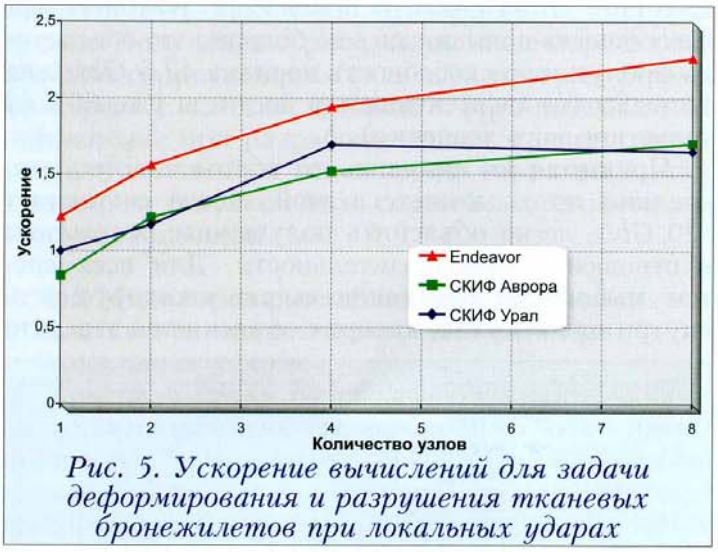

Рис. 5. Ускорение вычислений для задачи деформирования и разрушения тканевых бронежилетов при локальных ударах

Данная задача относится к другому типу – в ней превалируют вычисления, а взаимодействие с памятью не столь заметно. В силу этих обстоятельств, "СКИФ Урал" показывает результаты, сравнимые со "СКИФ Аврора", за счет более высокой тактовой частоты. Однако "*Endeavor*" показывает несколько лучшие результаты за счет большего количества ядер на сокет (6 против 4). Стóит отметить, что в этом случае имеют место вычисления с одинарной точностью. При расчетах с двойной точностью ситуация выглядела бы иначе – за счет возросшего в два раза объема взаимодействия с памятью.

Обратим внимание на еще один момент, связанный с насыщением масштабируемости. Дело в том, что исходный файл для решателя инженерного пакета *LS-Dyna* имеет небольшие размеры, в связи с чем время, затрачиваемое на коммуникации между узлами, быстро становится сравнимым со временем вычислений.

4.3 Задачи моделирования механического поведения грудной клетки при локальных ударах и деформационных изменений структуры трикотажных полотен на различных участках фигуры человека

Графики ускорения вычислений для обеих задач приведены на рис. 6, 7.

Задача моделирования механического поведения грудной клетки человека при локальных

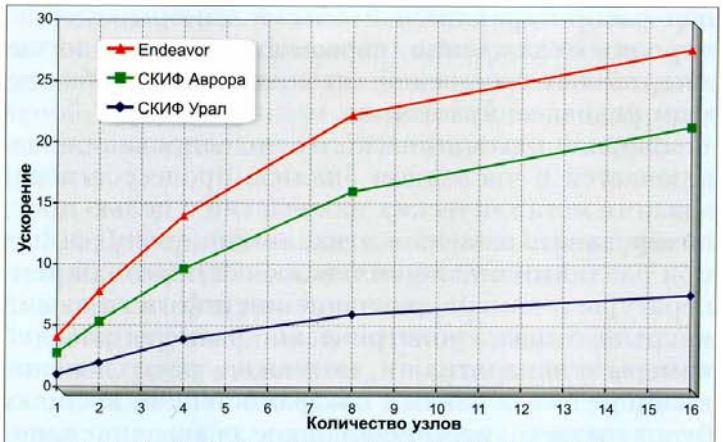

Рис. 6. Ускорение вычислений для задачи моделирования механического поведения грудной клетки

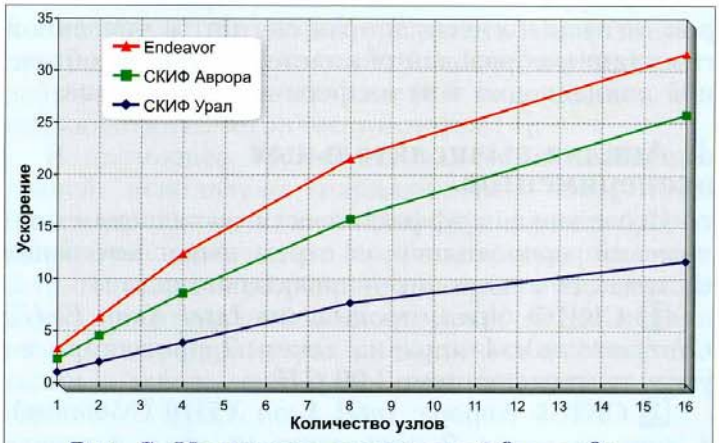

Рис. 7. Ускорение вычислений для задачи деформационных изменений структуры

ударах и задача деформационных изменений структуры демонстрируют сходное поведение, в целом типичное для *HPC*-приложений. Производительность систем определяется правильным балансом между их вычислительной способностью и скоростью взаимодействия с памятью. За счет улучшения этого баланса системы нового поколения показывают лучшую производительность, нежели "СКИФ Урал". Стóит также отметить, что при почти одинаковой производительности в расчете на одно ядро, "*Endeavor*" показывает более высокие результаты в расчете на один узел, так как имеет больше ядер – шесть.

4.4 Моделирование процессов газофазной конденсации металлических наночастиц

Зависимость скорости вычислений от количества задействованных ядер для задачи моделирования процессов газофазной конденсации металлических наночастиц приведена на графике (рис. 8).

Данная задача в целом очень сходна с задачей деформирования и разрушения тканевых бронежилетов при локальных ударах, поскольку в ней зависимость производительности от скорости общения с памятью не так велика. И всё же она сказывается, хотя и достаточно неочевидным образом. На

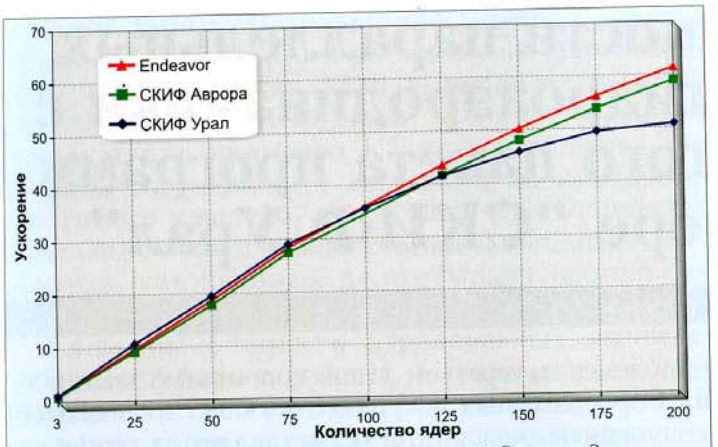

Рис. 8. Ускорение вычислений для задачи моделирования процессов газофазной конденсации металлических наночастиц

кластере "СКИФ Урал" производился запуск на меньшем количестве процессов на один узел с целью исключить конфликты при обращениях к памяти, исходящих от различных ядер внутри одного узла. Это приводит к росту числа задействованных узлов кластера, а соответственно и к увеличению объема коммуникаций между ними. Увеличение объема коммуникаций, в свою очередь, приводит к более быстрому насыщению шкалируемости, что и наблюдается в случае "СКИФ Урал". Таким образом, производительность подсистемы "процессор-память" внутри одного узла ограничивает кластерную масштабируемость задачи.

Заключение

В данной работе рассмотрено поведение нескольких *HPC*-приложений, свойства которых различаются. Скорость вычислений для первой задачи главным образом зависит от пропускной способности системы "процессор-память", для второй – от вычислительной мощности системы; пятая задача чувствительна к объемам коммуникаций, а остальные сочетают в себе все вышеперечисленные свойства. Подводя общий итог, можно сделать следующий вывод. При работе с вышеперечисленными приложениями мы сталкиваемся с достаточно типичной для *HPC*-вычислений ситуацией: производительность зависит не только от частоты процессоров и количества ядер на чипе; не менее важными факторами являются производительность системы "процессор-память", коммуникации в распределенной системе и иногда скорость файлового ввода-вывода. Для обеспечения максимальной производительности требуется нахождение оптимального баланса этих факторов.

Система типа "СКИФ Аврора", построенная на процессорах с архитектурой *Nehalem*, делает значительный шаг вперед по сравнению со "СКИФ Урал" (процессор *Harpertown*), за счет улучшения баланса между вычислительной мощностью процессора и пропускной способностью подсистемы "процессор-память". Система "*Endeavor*" на процессорах с архитектурой *Westmere* является следующим шагом на пути развития: улучшены как вычислительная способность (6 ядер на чипе вместо 4-х), так и скорость взаимодействия с памятью (42.5 *Gb/s* против 38 *Gb/s*). Все эти технологические новшества позволяют исследователям расширять "область поиска" и производить более глубокий анализ интересующих явлений путем увеличения уровня детализации, а также принятия во внимание эффектов, которые прежде игнорировались.

Авторы выражают благодарность сотруднику корпорации *Intel* Николаю Местеру за организационную и методическую помощь при выполнении исследований, представленных в данной работе.

Авторы:
Московский А.А. – Институт программных систем (Переславль-Залесский)
Перминов М.П., Черепенников В.В. – Интел (Нижний Новгород)
Соколинский Л.Б., Шамакина А.В. – Южно-Уральский государственный университет (Челябинск)

Литература

1. Абрамов С.М. СуперЭВМ Ряда 4 семейства СКИФ: штурм вершины суперкомпьютерных технологий // Параллельные вычислительные технологии (ПаВТ'2009): Труды международной научной конференции (Нижний Новгород, 30 марта – 3 апреля 2009 г.).– Челябинск: Изд-во ЮУрГУ. 2009. С. 5–16.

2. Васильев В.А., Ницкий А.Ю. Сравнительный анализ области применения тестовых задач оценки вычислительной мощности HPC систем (мощных кластеров) // Параллельные вычислительные технологии (ПаВТ'2010): Труды международной научной конференции (Уфа, 29 марта – 2 апреля 2010 г.).– Уфа: Уфимский государственный авиационный технический университет, 2010, с. 422–430.

3. Долганина Н.Ю., Сапожников С.Б., Маричева А.А. Моделирование ударных процессов в тканевых бронежилетах и теле человека на вычислительном кластере СКИФ Урал // Параллельные вычислительные технологии (ПаВТ'2010): Труды международной научной конференции (Уфа, 29 марта – 2 апреля 2010 г.).– Уфа: Уфимский государственный авиационный технический университет, 2010, с. 141–152.

4. Усенко И.Н., Долганина Н.Ю., Персидская А.Ю. Суперкомпьютерное моделирование деформационных изменений трикотажных полотен на фигуре человека // Параллельные вычислительные технологии (ПаВТ'2010): Труды международной научной конференции (Уфа, 29 марта – 2 апреля 2010 г.).– Уфа: Уфимский государственный авиационный технический университет, 2010, с. 606–610.

5. Терзи Д.В. Моделирование процессов газофазной конденсации металлических наночастиц на вычислительном кластере "СКИФ Урал" // Параллельные вычислительные технологии (ПаВТ'2010): Труды международной научной конференции (Уфа, 29 марта – 2 апреля 2010 г.).– Уфа: Уфимский государственный авиационный технический университет, 2010, с. 600–605.