Масштабируемые мультипроцессорные вычислительные системы высокой производительности

Александр Буравлёв, Марк Чельдиев, Александр Барыбин, Валерий Костенко, Денис Тумакин, Галина Петрова

Рассматривается высокопроизводительная универсальная мультипроцессорная система, которая предназначена для решения прикладных вычислительных задач, требующих параллельных вычислений. Система построена на базе узлов с процессорами микроархитектуры Intel Core, объединённых высокоскоростным интерконнектом реального времени Infiniband.

Рост рынка вычислительных систем

В последнее время в прессе интенсивно обсуждается тема распараллеливания вычислений при решении различных задач как в области сложных расчётов при моделировании или проектировании объектов в промышленности, так и в телекоммуникационной отрасли. Причём новые разработки ведутся на обоих «фронтах»: производители процессоров предлагают многоядерные кристаллы, а разработчики систем проектируют мультипроцессорные конфигурации. Рынок вычислительных систем растёт с опережающей скоростью по отношению к рынку серверов и персональных компьютеров: 15–20% в год против 11%.

Универсальная мультипроцессорная система

ОАО «Научно-исследовательский институт вычислительных комплексов имени М.А. Карцева» (НИИВК) и компания ПРОСОФТ совместно со специалистами компании FASTWEL завершили работы по созданию высокопроизводительной масштабируемой универсальной мультипроцессорной системы (УМС), предназначенной для решения прикладных вычислительных задач. Система построена на базе 24 новейших двухъядерных процессоров микроархитектуры Intel Core, объединённых высокоскоростным интерконнектом реального времени Infiniband. Решение прикладных задач, требующих параллельных вычислений, является основным предназначением этого вычислителя.

Общие особенности системы

Отличительными особенностями разработанной системы являются рекордные значения достигнутой плотности вычислений (в терминах FLOPS/Вт) и использование высокоскоростной сети обмена данными между вычислительными узлами с низкими уровнями задержки сигнала. В качестве вычислительных узлов были использованы одноюнитовые серверы компании Intel с двумя двухъядерными процессорами Intel Xeon серии 5100, работающими при пониженном напряжении питания и, соответственно, с пониженными значениями рассеиваемой тепловой мощности. Это позволило создать компактное решение, где все компоненты системы: вычислительные узлы, коммутаторы Gigabit Ethernet, коммутаторы Infiniband и система бесперебойного питания – размещаются в одной стандартной закрытой 19-дюймовой стойке, в качестве которой был использован новый шкаф для электронного оборудования VARISTAR производства компании Schroff. Шкаф имеет пылевлагозащищённое исполнение (степень защиты IP54), что позволяет использовать мультипроцессорную систему не просто вне специально подготовленных кондиционированных помещений, а непосредственно в производственных помещениях без предъявления каких-либо требований к их обустройству (рис. 1).

Рис. 1. Шкаф мультипроцессорной системы

Другой отличительной особенностью выбранной архитектуры является возможность гибкого масштабирования и наращивания возможностей системы. Архитектура вычислительных узлов и характеристики используемых чипсетов позволяют добиться удвоения вычислительной мощности без существенного увеличения теплового бюджета системы и с сохранением объёмных характеристик.

Сетевые решения

Вычислительные узлы объединены двумя различными типами сетей, предназначенными для обмена данными в процессе вычислений (Gigabit Ethernet) и обмена служебной информацией (Infiniband), каждая из которых коммутируется соответствующим коммутатором сети. Разработчики приложений могут выбирать ту сеть обмена данными между узлами, которая наилучшим образом подходит для конкретного приложения. Использование Gigabit Ethernet в качестве сети обмена данными между узлами – наиболее простой и распространённый способ решения стандартных задач с использованием наработанных решений для систем специального назначения. Использование сети Infiniband позволяет практически в 10 раз повысить скорость обмена данными между узлами и в 20-30 раз снизить задержки при передаче данных. Результаты реальных тестов, проведённых на УМС, определили значения скорости обмена по сети Infiniband в диапазоне от 700 до 1000 Мбайт/с с латентностью в диапазоне 3-4 мкс. Кроме того, использование сети Infiniband позволяет снизить нагрузку на центральный процессор, связанную с обработкой достаточно большого массива служебных данных в протоколе TCP/IP. Протокол Infiniband является полностью открытым и поддерживается такими лидирующими производителями, как IBM, Cisco, Sun, Intel, Hitachi и многими другими. Сеть Infiniband хорошо масштабируется практически на любое количество узлов. На момент написания статьи были доступны коммутаторы Infiniband вплоть до 288 каналов по 10 Гбит/с каждый, что позволяет наращивать количество вычислительных узлов в системе, объединяя между собой стойки. Кроме того, использование Infiniband эффективно при конструировании более сложных топологий сети обмена данными – так называемой топологии переключаемой инфраструктуры (Switched Fabric), когда узлы соединены друг с другом через различные коммутаторы. Такая топология более надёжна, так как позволяет избежать потери данных при обрыве одной из связей с узлом.

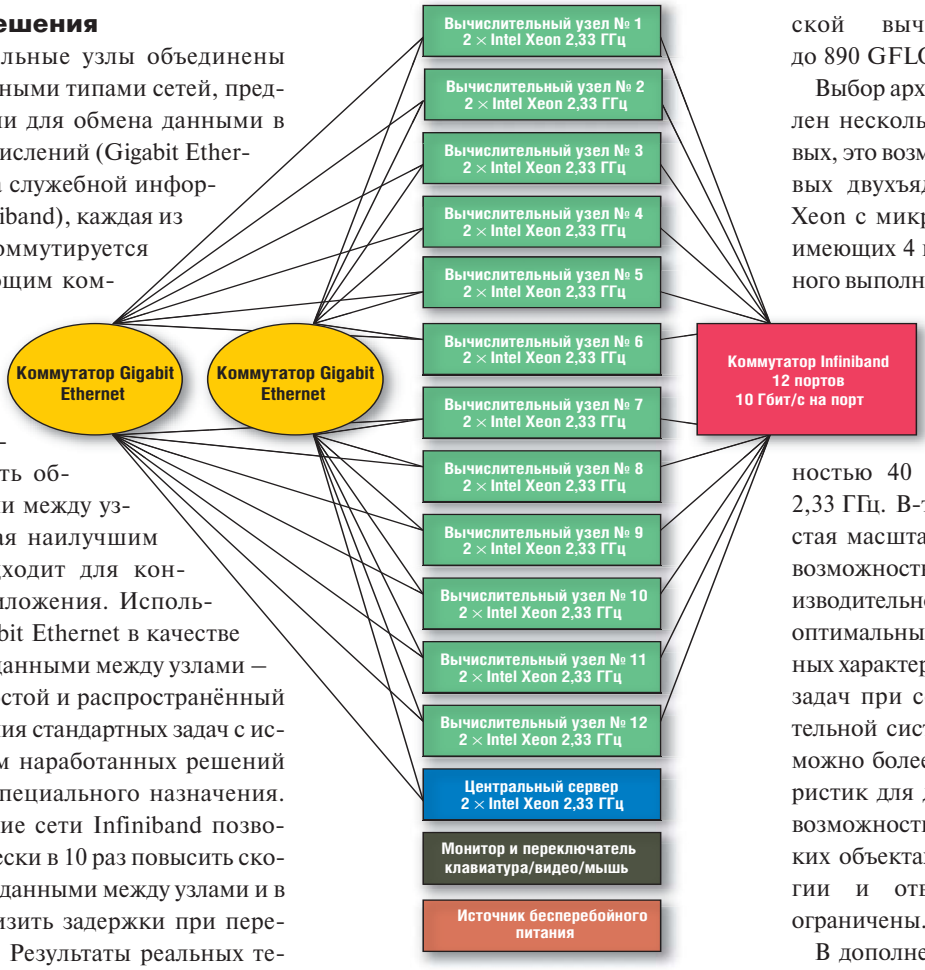

Рис. 2. Топология сети мультипроцессорной системы

В качестве топологии сети УМС было выбрано неблокирующее соединение узлов друг с другом через коммутаторы сетей Gigabit Ethernet и Infiniband, которое обеспечивает подвод двух каналов Gigabit Ethernet и одного канала Infiniband 10 Гбит/с к каждому узлу (рис. 2).

При такой архитектуре использование протокола Infiniband в неблокирующей топологии позволяет решать задачи, требующие высокоскоростного обмена данными между узлами в режиме реального времени.

Вычислительная мощность

Максимальная теоретическая вычислительная мощность текущей конфигурации системы составляет 447 GFLOPS при энергопотреблении 4,5 кВт. Таким образом, благодаря оптимизации архитектуры УМС по мощности и использованию низковольтных версий процессоров достигнуто высокое значение плотности вычислений, реально составляющее 75 GFLOPS/кВт. При этом общая архитектура построения УМС настолько гибка, что позволяет провести замену двухъядерных процессоров на четырёхъядерные с ростом теоретической вычислительной мощности до 890 GFLOPS.

Выбор архитектуры Intel был обусловлен несколькими причинами. Во-первых, это возможность использования новых двухъядерных процессоров Intel Xeon с микроархитектурой Intel Core, имеющих 4 конвейера для одновременного выполнения операций с плавающей запятой. Во-вторых, это возможность использования низковольтных версий таких процессоров с расчётной тепловой мощностью 40 Вт при тактовой частоте 2,33 ГГц. В-третьих, это гибкость, простая масштабируемость архитектуры и возможность дальнейшего подъёма производительности системы с сохранением оптимальных тепловых и массогабаритных характеристик. Одной из начальных задач при создании данной вычислительной системы было достижение как можно более низких тепловых характеристик для дальнейшего исследования возможности его использования на таких объектах, где подвод электроэнергии и отвод тепловой мощности ограничены.

В дополнение к перечисленным технологиям в серверных платформах, построенных на основе чипсета Intel 5000X и предназначенных для создания высокопроизводительных вычислительных систем, реализован фильтр когерентности кэш-памяти, который позволяет существенно увеличить производительность двухпроцессорных вычислительных систем при работе со многими типами сложных вычислительных задач, а также повысить эффективность использования полосы пропускания внешней шины процессора (FSB – Front Side Bus) при работе с ресурсоёмкими приложениями.

Основные технические характеристики унифицированной мультипроцессорной системы представлены в табл. 1.

Управляющий процессор, входящий в состав вычислительной системы, для передачи программ и данных в память, а также для параллельного запуска задачи использует стандартный интерфейс передачи сообщений MPI (Message Passing Interface) [1].

Области применения

Основной интерес для пользователя представляет реальная производительность вычислителя на «трудоёмких» задачах. Хорошо всем известно, что указанная в рекламе производительность ча-

Таблица 1

Основные технические характеристики УМС

Общие характеристики	
Число вычислительных узлов	12
Число процессоров/ядер	24/48
Тип процессора	Два низковольных двухъядерных процессора Intel® Xeon® 5138, 2,33 ГГц с двухканальной системной шиной частотой 1333 МГц, 40 Вт/процессор
Теоретическая пиковая производительность	447 GFLOPS
Производительность на тесте Linpack	340 GFLOPS, 76% от теоретической
Тип ведущей сети обмена данных и скорость	Infiniband 10 Гбит/с
Задержка при передаче пакетов данных по ведущей сети	3-4 мкс
Альтернативная сеть обмена данных	Gigabit Ethernet
Вспомогательная сеть	Gigabit Ethernet
Оперативная память узла	4 Гбайт FBDIMM DDR2 667 МГц, расширяемая до 64 Гбайт
Дисковая память узлов	2 диска SATA по 80 Гбайт
Операционная система	Red Hat Enterprise Linux 4
Среда параллельного программирования	Intel MPI Library 3.0
Компиляторы и библиотеки	GNU (gcc/g77), Intel C++ Compiler for Linux 10.0, Intel Fortran Compiler for Linux 10.0; Intel Math Kernel Library Cluster Edition 9.1
Конструкция и система питания	
Конструктив вычислительного узла	1U
Количество монтажных шкафов вычислительного кластера	1
Занимаемая шкафом площадь	0,6 м^2
Габариты шкафа	1000×600×2200 мм
Потребляемая УМС мощность	4,5 кВт
Система охлаждения	Воздушная, принудительная
Защита от внешних воздействий	IP54

сто не достигается из-за невозможности разделить задачу на параллельные процессы.

Для определения производительности вычислителя была принята методика австрийской фирмы AVL Advanced Simulation Technologies, которая проводит тестирование серверов различной комплектации с целью определения реального быстродействия. С этой целью разработаны пакеты программ (benchmark test), моделирующие различные технологические процессы. Пакеты моделирующих программ представляют собой расчёты параметров состояния объектов сложной геометрии, работающих в критических условиях. Таковыми, например, являются цилиндры автомобильных и авиационных двигателей, системы водяного и воздушного охлаждения, смесители различного назначения и др.

Моделирование складывается из двух процедур:

- задание геометрии объекта и построение его пространственной (трёхмерной) сетки;
- вычисление итераций искомых параметров состояния объекта (скорости, давления, температуры) по функции их значений, заданных на границе в каждой пространственной точке модели.

Исходными данными для запуска являются тип задачи, количество итераций, количество узлов и количество процессов, осуществляющих параллельную обработку. Тестирование проводилось на различных задачах, геометрия которых представлена 5×10^6 точками. При вычислении 10 итераций на 12 узлах заявлены библиотеки Lib (ia64-unknown-linux) с повышенной точностью.

Операционная система Linux сопровождает полный мониторинг вычислений. Файл Log содержит протокол вычислений, включая время занятости вычислительного ядра (Linear Solver), время коммуникаций MPI и затраты на ввод-вывод.

Пример фрагмента протокола приведён на рис. 3.

Протокол показывает, что задача разделена интерфейсом MPI на 12 процессов (Linear Solver). При численном моделировании параметров на пространственной сетке в точках максимальное время вычислений, затраченное в совокупности всеми ядрами УМС при распараллеленных вычислениях (Linear Solver), равно 1953,90 с = 32,5 мин. Для сравнения приводится общее суммарное время вычислений УМС в однопроцессорном режиме, которое равно 22 823,97 с = 381 мин. Время, затрачиваемое для передачи данных и организации параллелизма вычислений пакетом MPI, составляет менее 6% от общих временных затрат.

Результаты проведённого тестирования УМС показали высокую степень распараллеливания согласно закону Амдала и, как следствие, целесообразность применения этой системы для задач численного моделирования и построения сложных виртуальных объектов.

На представленной масштабируемой мультипроцессорной системе было проведено решение ряда практических задач, требующих больших вычислительных мощностей. Рассмотрим некоторые из них.

Обучение нейронных сетей прямого распространения

Был реализован новый алгоритм мультистарта с отсечением для решения задачи обучения нейронной сети прямого распространения. Основная идея алгоритма мультистарта с отсечением заключается в проведении нескольких параллельных запусков (стартов) локально-оптимального алгоритма обучения с различными начальными приближениями. При этом после выполнения заданного числа шагов локально-оптимального алгоритма обучения выделяются «неперспективные» старты, которые исключаются из рассмотрения, и процесс обучения продолжается на более узком наборе стартов. При одинаковом времени работы алгоритма мультистарта и предложенного алгоритма мультистарта с отсечением предложенный алгоритм позволяет уменьшить ошибку аппроксимации до двух раз. Предложенный алгоритм хорошо подходит для параллельной реализации на рассматриваемой вычислительной системе, так как при этом требуется высокоскоростная среда обмена между вычислительными узлами. Это обусловлено тем, что для обеспечения равномерной загрузки узлов вычислительной системы в ходе работы алгоритма требуется перераспределение стартов между узлами, так как исключаемые старты заранее неизвестны и определятся в ходе работы алгоритма.

Имитация отжига для решения задач построения многопроцессорных расписаний

Реализован параллельный алгоритм имитации отжига для построения многопроцессорных расписаний [2], основанный на разбиении пространства корректных расписаний на непересекающиеся области и поиске решения в каждой из них отдельно. Предложенный алгоритм характеризуется низким тра-

```
*   Total Execution                  1    1944.32 [100.00%]   1953.88 [100.0%]
********************************************************************************
*
********************************************************************************
*   Processor  12                Count      CPU Time           Wallclock Time
********************************************************************************
*   ACCIF_EXCHANGE                  17       0.00 [  0.00%]      0.00 [  0.0%]
*   File IO                        184       7.79 [  0.40%]      8.27 [  0.4%]
*   Linear Solver                  379    1772.79 [ 91.20%]   1773.98 [ 90.7%]
*
********************************************************************************
*   Total MPI Timings            Count      Total CPU Time      Average WCT
********************************************************************************
*   ACCIF_EXCHANGE                 214       0.00                0.00
*   MPI Barrier                     15       0.82 [  0.04%]      0.82 [  0.0%]
*   MPI Communication            84782     109.91 [  5.65%]    109.92 [  5.6%]
*   Other                            1      46.90 [  2.41%]     55.15 [  2.8%]
*   Thermochemistry                 49       5.44 [  0.28%]      5.45 [  0.2%]
********************************************************************************
*   Total Execution                  1    1943.93 [100.00%]   1953.88 [100.0%]
********************************************************************************
*   File IO                       2208     141.88              14.26
*   Linear Solver                 4548   19904.51            1698.89
*   MPI Barrier                    180      17.55               1.54
*   MPI Communication           988192    2145.46             180.32
*   Other                           12     543.93              52.66
*   Thermochemistry                588      67.28               5.94
********************************************************************************
*   Total Execution                 12   22823.97            1953.90
********************************************************************************
```

Рис. 3. Фрагмент протокола вычислений

фиком обмена между узлами вычислительной системы, и его последовательное выполнение позволяет уменьшить время решения задачи до трёх раз по сравнению с классическим алгоритмом имитации отжига при сохранении, а во многих случаях даже при улучшении качества получаемых расписаний. Параллельный алгоритм является масштабируемым по отношению к числу вычислительных узлов в системе и при реализации на рассматриваемой вычислительной системе позволяет достичь ускорения, близкого к теоретически возможному, получаемому в соответствии с законом Амдала.

Распознавание аномального поведения динамических систем

Распознавание аномального поведения динамических систем производилось согласно алгоритму, основой которого является разметка анализируемой фазовой траектории поведения системы аксиомами. Под аксиомой понимается бинарная функция, определённая в точке и некоторой её окрестности на фазовой траектории. Определение аномального поведения в работе наблюдаемой системы ведётся не путём поиска эталонных траекторий в наблюдаемой фазовой траектории, а путём поиска разметок эталонных траекторий в ряду разметки. Это позволяет повысить устойчивость алгоритма распознавания к искажениям эталонных траекторий по времени и амплитуде в анализируемой фазовой траектории системы по сравнению с другими известными методами. Задача построения алгоритма распознавания нештатных ситуаций может быть сформулирована как задача обучения по прецедентам. В работе [3] предложен алгоритм решения этой задачи. Однако для многих практических задач время обучения является неприемлемо большим. Численные эксперименты показали, что при параллельной реализации алгоритма обучения на рассматриваемой вычислительной системе удаётся добиться близкого к линейному ускорению обучения относительно числа используемых вычислительных узлов.

Реализация управляемого ковариационного адаптивного формирователя диаграмм

Рассматривается задача формирования диаграммы направленности, встречающаяся, например, в радиолокации или гидроакустике. Базовые шаги обработки информации включают временную сегментацию ряда, перекрытие и работу быстрого преобразования Фурье (БПФ), формирование управляющей ковариационной матрицы, инверсию ковариационной матрицы, использующей факторизацию Холецкого, оценку адаптивных векторов управления, формирование адаптивных лучей во временной области, удаление перекрытия и конкатенацию сегментов, чтобы сформировать непрерывную лучевую последовательность во времени.

Предположим, что M – число приёмников, N – число отсчётов, F_d – частота дискретизации. Для выполнения дискретного БПФ с двойным перекрытием необходимо выполнить $10 \cdot M \cdot N \cdot \log N$ операций за время $T = N/F_d$.

Необходимое быстродействие для БПФ составляет $10 \cdot M \cdot \log N \cdot F_d$.

Необходимое быстродействие для накопления матриц составляет $4 \cdot M \cdot M \cdot F_d$.

Необходимое быстродействие для обращения матриц один раз за M тактов составляет $4 \cdot M \cdot M \cdot F_d$.

Необходимое быстродействие для вычисления статистик составляет $4 \cdot M \cdot M \cdot F_d$.

Суммарное быстродействие составляет $M \cdot F_d \cdot (10 \cdot \log N + 12M)$.

Если принять $M = 2000$, $N = 16\,384$, $F_d = 5000$ Гц, то необходимое быстродействие вычислительной системы для этой задачи должно составить 241,4 GFLOPS.

Задача такой размерности решается предлагаемой вычислительной системой менее чем за 1 с, то есть с запасом для решения задач вторичной обработки.

Заключение

В заключение можно отметить, что рассматриваемая масштабируемая универсальная мультипроцессорная система будет эффективно применяться для решения наиболее трудных с вычислительной точки зрения задач, требующих больших объёмов памяти и производительности (моделирование естественных процессов в различных областях знаний, опережающее моделирование в производственных АСУ, задачи реального масштаба времени в таких сферах применения, как радиолокация, гидроакустика, мониторинг космической обстановки и т.д.).

Система УМС обладает пониженным тепловыделением и может использоваться в производственных помещениях без предъявления каких-либо требований к их обустройству. ●

Литература

1. Немнюгин С., Стесик О. Параллельное программирование для многопроцессорных вычислительных систем. – СПб. : БХВ-Петербург, 2002.
2. Калашников А.В., Костенко В.А. Параллельный алгоритм имитации отжига для построения многопроцессорных расписаний // Известия РАН. Теория и системы управления. – 2008. – № 3. – С. 101-110.
3. Костенко В.А., Коваленко Д.С. Метод построения алгоритмов распознавания, основанных на идеях аксиоматического подхода // Научная сессия МИФИ-2007. IX Всероссийская научно-техническая конференция «Нейроинформатика-2007»: сб. трудов. – М. : МИФИ, 2009.

Авторы – сотрудники компании ПРОСОФТ и ОАО «НИИВК им. М.А. Карцева»
Телефон: (495) 330-0929
E-mail: postoffice@niivk.ru