ПРИМЕНЕНИЕ ПРОГРАММНОГО КОМПЛЕКСА «ПИРАМИДА» ДЛЯ SPMD-ВЫЧИСЛЕНИЙ НА ГЕТЕРОГЕННЫХ МАССОВО-ПАРАЛЛЕЛЬНЫХ ВС

А.В. Баранов, А.В. Киселёв, Е.А. Киселёв, В.В. Корнеев, Д.В. Семёнов

В докладе представлены результаты работ по совершенствованию программного комплекса (ПК) «Пирамида» организации массово-параллельнных вычислений с распаралелливанием по данным [1].

ПК «Пирамида» реализует модель вычислений вида SPMD (Single Program, Multiple Data – Единая Программа, Множество Данных) и предназначен для автоматизации запуска экземпляров прикладной последовательной программы на массово-параллельной вычислительной системе (МВС), распределения множества обрабатываемых данных между экземплярами программы, сбора результатов обработки и управления вычислениями.

Семантику вычислений пользователь определяет в последовательной программе с учетом следующих особенностей: экземпляры программы (ЭП) запускаются на выделенных пользователю вычислительных ресурсах и выполняется независимо и асинхронно. Выполняемая программа, ее параметры и требования к вычислительным ресурсам задаются пользователем в паспорте задания.

Множество D обрабатываемых программой данных представляет собой декартово произведение множеств $D = P_1 \times P_2 \times \ldots \times P_k$, задаваемых параметрами программы. Множество значений параметра может задаваться непосредственно, в виде списка значений, либо косвенно, в виде диапазона значений или ссылки на файл, содержащий множество значений. В текущей версии ПК «Пирамида» для задания входных данных можно использовать до 10 параметров. Тип параметра может быть как целым, так и строковым.

В основе организации распараллеливания по данным лежит принцип рекурсивного формирования множества подмножеств данных:

$$D = \left\{ D_{i_1} \,\middle|\, i_1 = \overline{1, n_1} \right., \; D_{i_1} = \left\{ D_{i_1 i_2} \,\middle|\, i_2 = \overline{1, n_2} \right., \; D_{i_1 i_2} = \ldots \right\} \ldots \left\{ D_{i_1 \ldots i_r} \right\} \ldots \Big\}$$

Элемент множества $D_{i_1 \ldots i_r}$ называется слайсом. Он задает элементарную работу, выполняемую экземпляром программы. Множество слайсов $\{ D_{i_1 \ldots i_r} \}$, выделяемых для обработки экземпляру программы, называется пулом работ. Размер пула работ – мощность множества – величина задаваемая при конфигурировании ПК. Причем, для каждого уровня рекурсии $l = \overline{1, r}$ может быть определена индивидуальная мощность множества n_l.

Способ распределения работ определил архитектуру вычислительной среды, создаваемой программным комплексом «Пирамида» на базе МВС. На основе ресурсов МВС ПК формирует r-уровневую иерархию вычислителей (рис. 1). Первый уровень иерархии соответствует вычислителю C, выполняющему обработку множества данных D. Каждый следующий уровень соответствует множеству вычислителей, обрабатывающих подмножества множества данных предыдущего уровня. Уровню r соответствуют n_r вычислителей (на рис. 1 обозначены как c^t), выполняющих обработку пулов работ $\{ D_{i_1 \ldots i_r} \}$.

Нижний уровень иерархии r может соответствовать различным видам вычислителей: вычислительная система, вычислительный модуль, микропроцессор, вычислительное ядро.

Распределение данных для обработки вычислителями осуществляют управляющие процессы – менеджеры. Из пула работ, выделенных вычислителю уровня i, менеджер M_i формирует пулы работ для вычислителей уровня i+1, контролирует процесс их выполнения, и, в случае необходимости (например, выхода из строя вычислителя), перераспределяет работы между вычислителями для обеспечения отказоустойчивости вычислений.

Основное отличие новой версии ПК «Пирамида» от предыдущей заключаются в возможности организации вычислений на гетерогенных МВС, вычислительные узлы которых имеют архитектурные различия (например, построены на программно-несовместимых микропроцессорах) или (и) состоят из вычислителей различного типа (например, вычислительные узлы содержат универсальные и графические процессоры).

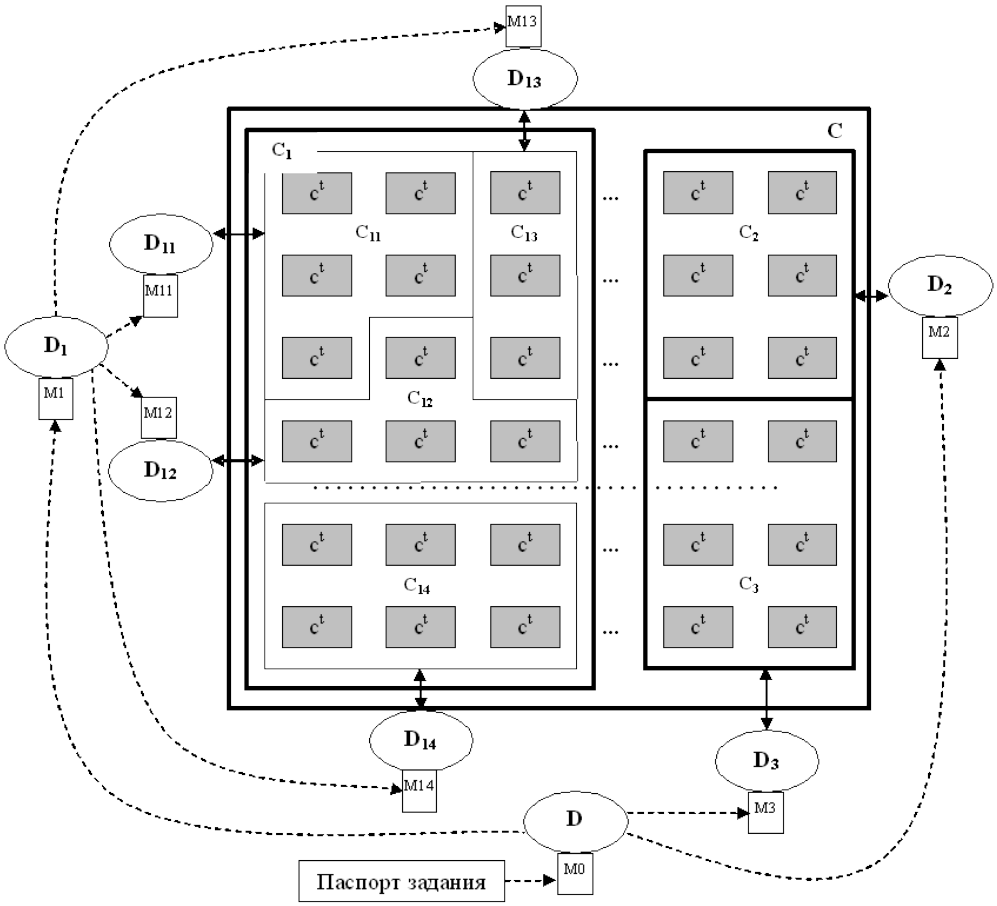

Рис. 1. Пример структуры вычислительной среды, организованной на базе ВС средствами ПК «Пирамида»

В обозначении c^t – вычислителя уровня r на рис. 1 – индекс t указывает его тип. $t \in \{ t \}$, где $\{ t \}$ – множество типов вычислителей МВС.

Пользователь должен предусмотреть подготовку вариантов исполняемых модулей программы для всех типов вычислителей, включаемых в состав вычислительной среды. Обозначив как $W^t(\{D_{i_1 \dots i_r}\})$ результат преобразования, выполняемого программой W^t над слайсом $D_{i_1 \dots i_r}$ на вычислителе c^t, можно записать требование, предъявляемое к различным вариантам исполняемых модулей программы:

$$\forall\ t_1, t_2 \in \{t\} \quad (W^{t_1}(D_{i_1 \dots i_r}) = (W^{t_2}(D_{i_1 \dots i_r})$$

Менеджеры, с учетом типа вычислителя t, осуществляют выбор варианта исполняемого модуля для запуска экземпляра программы, выполняющей выделенный ей пул работ $W^t(\{D_{i_1 \dots i_r}\})$.

Отметим отличие подхода к распараллеливанию, используемого в ПК «Пирамида», от известных систем распределенных вычислений, таких как BOINC [2] и X-Com [3].

Названные программные продукты предназначены для работы в распределенных средах и грид-системах и ориентированы на обнаружение и утилизацию простаивающих вычислительных ресурсов из состава распределенной грид-среды. Предназначение ПК «Пирамида» несколько уже – работа на вычислительных кластерах с большим числом процессоров. При этом ПК «Пирамида», как с точки зрения прикладного пользователя, так и с точки зрения организации вычислений, является более «лёгкой» системой по сравнению с BOINC и X-Com.

ПК «Пирамида» не использует для учёта пользователей и заданий СУБД (в BOINC, например, используется MySQL) и имеет намного более простую структуру без веб-сайтов, планировщиков, генераторов заданий и подобных компонентов. Для прикладного пользователя ПК «Пирамида», прежде всего – это альтернатива MPI в части автоматизации распараллеливания по данным. Пользователь ПК «Пирамида» может вообще не задумываться об организации параллельных вычислений, сосредоточившись на реализации прикладной логики в последовательной программе, принимающей на вход параметры вариантного счёта. Прикладная последовательная программа может быть разработана программистом (пользователем ПК «Пирамида») с использованием любых технологий и языков программирования (в системе X-Com, например, требуется разделение прикладной задачи на серверную и клиентскую части, а также использование прикладных

интерфейсов (API) X-Com). Фактически ПК «Пирамида» представляет собой внешнюю управляющую оболочку над произвольной прикладной последовательной программой.

Непосредственными аналогами-альтернативами ПК «Пирамида» являются технологии MPI и MapReduce, однако ни та, ни другая система программирования не избавляет прикладного пользователя от необходимости организации параллельных вычислений при распараллеливании по данным. Следует отметить, что MapReduce позволяет эффективно организовать вариантный счёт в случае, если входные данные представляют набор строк в файле большого размера. В подобной ситуации использование MapReduce более предпочтительно, чем ПК «Пирамида». Однако, в случае, когда входные данные задаются в виде диапазонов чисел (номеров вариантов данных), для организации работы MapReduce-приложения пользователю придётся самостоятельно реализовывать процессы, управляющие распределением данных.

При использовании MPI пользователь-программист тратит значительное время на организацию вычислений (порождение процессов и распределение данных) в MPI-программе. При решении крупномасштабных задач, использующих большое число процессоров в течение длительного периода времени, на первый план выходит проблема обеспечения отказоустойчивости MPI-программы. Известно, что при отказе во время вычислений хотя бы одного процессора велика вероятность аварийного завершения всей MPI-программы. Для обеспечения надёжности вычислений пользователь должен предусмотреть периодическое сохранение контрольных точек и возможность рестарта программы, что требует дополнительных усилий при разработке программы. ПК «Пирамида» производит указанную работу за пользователя, позволяя значительно сократить время разработки прикладной программы.

В настоящее время ПК «Пирамида» установлен в Межведомственном суперкомпьютерном центре РАН на МВС cuda.jscc.ru, включающей 34 модуля Server Blade HP ProLiant BL2x220c G5 с двумя четырехядерными микропроцессорами Intel Xeon 3,0 ГГц и 8192 МБ оперативной памяти на каждом модуле.

ЛИТЕРАТУРА:

1. А.В. Баранов, А.В. Киселев, Е.А. Киселев, В.В. Корнеев, Д.В. Семенов. Программный комплекс «Пирамида» организации параллельных вычислений с распараллеливанием по данным // Труды международной суперкомпьютерной конференции «Научный сервис в сети Интернет: суперкомпьютерные центры и задачи» (20-25 сентября 2010 г., г. Новороссийск). М.: Изд-во МГУ, 2010. с. 299-302.
2. BOINC – Программное обеспечение с открытым исходным кодом для организации добровольных распределённых вычислений и распределённых вычислений в сети // http://boinc.berkeley.edu
3. Система метакомпьютинга X-Com // http://x-com.parallel.ru