Автоматизированное распараллеливание программ для гетерогенных кластеров с помощью системы SAPFOR

Н.А. Катаев, А.С. Колганов

ИПМ им. М.В. Келдыша РАН

В статье будет рассмотрен подход к автоматизированному распараллеливанию программ для кластеров с помощью системы SAPFOR (System FOR Automated Parallelization). Главной целью системы SAPFOR является автоматизация процесса отображения последовательных программ на параллельные архитектуры в модели DVMH, которая является моделью программирования, основанной на директивах. Помимо этого система SAPFOR позволяет выполнять автоматически некоторый класс преобразований над исходным кодом программы по запросу пользователя через графический интерфейс. На определенных классах задач пользователь системы SAPFOR может рассчитывать на полностью автоматическое распараллеливание, если программа была написана или приведена к потенциально параллельному виду. Также в статье будут описаны подходы к построению схем распределения данных и вычислений на распределенную память в модели DVMH. Эффективность полученных алгоритмов построения схем распределения данных и вычислений будет продемонстрирована на примере некоторых приложений из пакета NAS Parallel Benchmarks.

Ключевые слова: SAPFOR, DVMH, автоматизация распараллеливания, распределение данных, распределение вычислений, гетерогенные кластеры

1. Введение

Высокопроизводительные вычислительные технологии получили широкое распространение среди большого количества областей науки: вычислительная гидродинамика, исследования климата и окружающей среды, нейронные сети и искусственный интеллект, и многие другие. Сложились разнообразные подходы к организации параллельных вычислений, предполагающие использование различных технологий параллельного программирования [1]. При этом при выборе и применении предпочтительных подходов для решения конкретной задачи часто приходится сталкиваться с различными трудностями [2].

Среди них можно выделить необходимость одновременного применения различных технологий программирования (MPI+OpenMP, MPI+OpenMP+CUDA и т.д.), чтобы задействовать все доступные вычислительные ресурсы. При этом прикладной программист должен обладать знаниями, позволяющими применять каждую из них. Множество используемых технологий может меняться по мере развития доступной вычислительной аппаратуры, что усложняет сопровождение и развитие уже написанных программных комплексов. Кроме того запуск вычислительной задачи на сложной, часто гибридной, вычислительной системе может сопровождаться необходимостью подбора многочисленных параметров для достижения максимальной производительности. В сложившейся ситуации крайне желательной становится автоматизация максимального количества этапов, составляющих процесс разработки и сопровождения параллельной программы.

Одним из направлений такой автоматизации является разработка единых подходов, охватывающих сразу несколько уровней параллелизма и позволяющих отображать программу на различные архитектуры вычислителей. Примером такого единого подхода является стандарт SYCL [3], который добавляет параллелизм в последние версии языка C++ для поддержки параллельного выполнения на GPU, CPU и FPGA. Параллельная программа должна явно описывать параллельное выполнение, при этом оставаясь программой на

стандартном языке C++, а ответственность за реализацию параллелизма ложится на используемые компиляторы. В исходном виде SYCL ориентирован на системы с общей памятью, но на его основе также разрабатывается открытый проект Celerity [4] для отображения программ на кластеры, оснащенные в узлах ускорителями.

Альтернативой такому подходу может быть применение директивных расширений стандартных последовательных языков программирования XcalabalACC [5], DVMH [6, 7]. Преимущество таких подходов заключается в том, что они позволяют сначала разрабатывать и отлаживать последовательную программу, а потом добавлять в нее спецификации параллелизма, в то время как подходы, аналогичные SYCL, требуют, чтобы программа изначально разрабатывалась с использованием конструкций, описывающих параллелизм.

Являясь достаточно универсальной для описания различных уровней параллелизма, доступных на гибридном вычислительном кластере, DVMH модель скрывает конкретные технологии программирования внутри реализации компиляторов с DVMH языков. Это в свою очередь позволяет при необходимости расширять множество поддерживаемых архитектур, избегая значительных изменений в DVMH модели, и обеспечивает переносимость существующих DVMH программ. Еще одним уровнем автоматизации, который обеспечивает DVM система, является возможность динамической настройки запускаемых параллельных программ на выделенные для их выполнения вычислительные ресурсы [8].

Несмотря на то, что модели программирования, опирающиеся на директивные расширения существующих языков, являются высокоуровневыми, их применение прикладным программистом все равно требует достаточных знаний в области параллельных вычислений и может быть сопряжено с трудностями. Способствовать решению данной проблемы может дальнейшая автоматизация процесса распараллеливания, связанная с созданием автоматизирующих систем, которые упрощают перевод последовательной программы в параллельную. Однако полностью автоматическое распараллеливание произвольных программ сталкивается со значительными трудностями, не позволяющими достичь приемлемой эффективности получаемых параллельных версий. Поэтому на рассматриваемые последовательные программы могут накладываться существенные ограничения, а пользователь получает возможность описывать свойства программ, которые невозможно проанализировать [9–12].

В связи с этим перспективным видится смешанный подход, который объединяет использование высокоуровневой модели параллельного программирования, системы автоматизации, ответственной за выполнение наиболее трудоемких этапов распараллеливания, и возможность для пользователя контролировать ход распараллеливания и принимать в нем активное участие [13].

В качестве инструмента автоматизации может выступать система SAPFOR (System FOR Automated Parallelization) [14, 15], ориентированная на использование DVMH языков как целевых. С одной стороны система включает автоматически распараллеливающий компилятор, при этом инкапсулированные в DVMH модели возможности по динамической настройке запускаемых параллельных программ упрощают его разработку. С другой стороны, система обладает широкими возможностями по статическому и динамическому анализу программ [18, 19], автоматизирует выполнение преобразований исходной программы, позволяя пользователю выбирать фрагменты программы, которые должны быть преобразованы. Графический интерфейс пользователя позволяет управлять процессом распараллеливания [17].

В предыдущих работах [13, 16] рассматривались возможности системы SAPFOR, направленные на распараллеливание программ для систем с общей памятью (мультипроцессор, GPU). Данная статья охватывает дальнейшее расширение возможностей системы SAPFOR в направлении отображения программ на многоядерные гибридные вычислительные кластеры. В статье основное внимание уделяется алгоритмам распределения данных и вычислений между узлами кластера в модели DVMH.

Статья состоит из введения, 3 разделов и заключения. В разделе 2 приведен краткий

обзор работ, направленных на автоматизацию разработки параллельных программ для систем с распределенной памятью. Раздел 3 посвящен алгоритмам распределения данных и вычислений, реализованным в системе SAPFOR. Результаты вычислительных экспериментов, охватывающие распараллеливание некоторых программ из набора NAS Parallel Benchmarks [20], приведены в разделе 4.

2. Обзор существующих работ

Распараллеливание для кластера обладает существенными особенностями, отличающими его от распараллеливания для систем с общей памятью. Разработчику приходится учитывать размещение данных на узлах системы наряду с распределением вычислений между отдельными вычислительными устройствами, чтобы обеспечить доступ к удаленным данным, максимально сократив при этом коммуникационные издержки. Таким образом, важную роль начинает играть требование локальности данных, используемых на каждом узле, и необходимость сбалансированного распределения данных, чтобы равномерно загрузить вычислительные устройства работой. Ситуация усугубляется тем, что приходится принимать глобальные решения в рамках всей программы в целом, так как отдельные ее фрагменты могут накладывать противоречивые требования, что в конечном счете приведет к дополнительным коммуникациям, направленным, на перераспределение данных.

В связи с этим рассматриваются различные подходы к распараллеливанию программ на вычислительный кластер. Так как процесс распараллеливания для систем с распределенной памятью можно разделить на три этапа: распределение данных, распределение вычислений, организация коммуникаций для доступа к удаленным данным или перераспределения данных, то автоматизации могут подвергаться только некоторые из них.

Например, в работе [21] рассматривается подход к построению оптимального распределения вычислений и организации коммуникаций, опирающийся на применение полиэдральной модели распараллеливаемой программы, для предварительно заданного фиксированного распределения данных. Подход, основанный на предварительно заданном распределении данных, также применялся в инструменте [22], при этом интересно, что инструмент обеспечивал пользователя диалоговой оболочкой, позволяющей управлять процессом распараллеливания, задавать распределение данных и управлять преобразованиями программ. Распределение вычислений выполнялось автоматически и основывалось на правиле собственных вычислений, когда запись значений выполняется на том процессоре, который владеет записываемыми данными. Подход, описанный в [21] позволяет ослабить это ограничение, в том числе обеспечивая размножение данных между узлами.

Инструмент Molly [11], расширяет возможности компилятора Polly [23] для систем с общей памятью, построенного на базе LLVM [24] и основанного на использовании полиэдральной модели. При этом вводится специальный тип данных, описывающий распределяемые массивы, что позволяет контролировать отсутствие операций адресной арифметики, применяемых к распределяемым данным. Распределение данных выполняется равными блоками распределяемых массивов между процессами, при этом выравнивание данных друг на друга не предусмотрено, а для распределения вычислений применяется правило собственных вычислений. Предполагается, что в дальнейшем пользователь сможет корректировать как распределение данных, так и вычислений, задавая соответствующие директивы. Кроме того накладываются дополнительные ограничения на структуру распараллеливаемых фрагментов (SCoP), вводится требование глобальности распределяемых данных, что позволяет избежать подробного межпроцедурного анализа, редукционные операции также не поддерживаются.

Подход с использованием директив, описывающих распределение данных, также предложен в работе [25]. Директивы содержат большое количество параметров, позволяющих гибко управлять требуемым распределением данных. Также предлагается нестандартное

распределение массивов с перекрытиями, что позволяет сократить частоту обменов данными. При этом использование специальных директив для описания такого распределения упрощает разработку программы и вероятность ошибок при задании сложных индексных выражений в обращениях к массивам.

Подход, основанный на оптимизации распределения вычислений и необходимых коммуникаций, также приведен в работе [27], расширяющей возможности полиэдрального компилятора Pluto [26] для систем с общей памятью. При этом распределение данных как таковое не требуется, а размещение данных на узлах в каждый конкретный момент определяется выполняемыми над ними вычислениями. Данный подход не предусматривает глобального принятия решений по распределению данных, которое максимально бы соответствовало распределению вычислений, что может привести к увеличению частоты и объема коммуникаций.

Построение распределения данных наряду с распределением вычислений и оптимизацией коммуникаций было реализовано в инструменте Paradigm [28]. Исследования были направлены на распараллеливание последовательных программ на языке Фортран 77. Инструмент предполагал поддержку достаточно широкого класса задач, в том числе за счет поддержки нерегулярных вычислений и распараллеливания циклов с зависимостями за счет организации конвейерного выполнения. При этом в работе отмечаются ограничения связанные с определением правила выравнивания измерения массивов, а экспериментальные результаты приводятся для небольших вычислительных ядер.

3. Построение схем распределения данных и вычислений в системе SAPFOR

3.1. Варианты отображения последовательной программы на кластер

Распределение данных – одна из основных проблем отображения последовательной программы на кластер в случае использования регулярных сеток. Для эффективного задействования всей мощности вычислительного кластера требуется учитывать специфику обработки данных в циклах последовательной программы, так как зачастую в них содержится основная вычислительная нагрузка. Рассмотрим следующие варианты отображения последовательной программы на кластер:

- выполняется только распределение вычислений, а данные остаются размноженными на всех узлах. При таком подходе существенно упрощается преобразование последовательной программы в параллельную для кластера, так как все данные дублируются на каждом узле кластера. В свою очередь, каждый цикл может «требовать» свое распределение, что легко реализуется данной моделью;

- выполняется распределение как данных, так и вычислений. При таком подходе необходимо учитывать интересы всех циклов, участвующих в распараллеливании. В этом случае в силу того, что на каждом узле присутствует только часть данных, необходимо выполнять их пересылки для корректного выполнения последовательных участков программы.

Система SAPFOR преобразует исходную программу в параллельную с использованием модели DVMH. Данная модель использует второй вариант отображения последовательной программы на кластер – отображение как данных, так и вычислений на узлы кластера. В связи с этим необходимо обеспечить такое распределение данных, чтобы количество коммуникаций между процессорами, а также их объем были как можно меньше. Можно отметить, что с помощью модели DVMH можно реализовать и первый вариант, но, во-первых, параллельная программа будет требовать столько же памяти на каждом узле, что и последовательная, и, во-вторых, эффективность выполнения такой программы может быть ниже

из-за большого объема коммуникаций, возникающих в момент перераспределения данных.

Алгоритм распределения данных состоит из двух этапов. На первом этапе выполняется межпроцедурный анализ всей программы: анализируются циклы и используемые в них массивы. Затем собранная информация обобщается и выполняется поиск распределения данных с как можно меньшими издержками по пересылке данных между узлами.

3.2. Определение распределяемых массивов

По умолчанию система SAPFOR считает все массивы в программе распределяемыми. Распределяемый массив – это такой массив, для которого необходимо построить распределение данных с помощью DVMH-директив и выполнить соответствующее выбранному распределению отображение вычислений в параллельных циклах и одиночных операторах. Но не все массивы могут быть распределенными. Чтобы снизить количество распределяемых данных, системе SAPFOR необходимо уметь распознавать такие ситуации либо в автоматическом, либо в полуавтоматическом режиме. Рассмотрим случаи, когда массив не требует распределения.

К первой категории массивов, которые не требуется распределять, относятся те массивы, которые были определены как приватизируемые или редукционные системой SAPFOR или были указаны в соответствующих спецификациях пользователем в исходном коде программы или через графический интерфейс. Данные массивы являются вспомогательными в местах использования (в основном в циклах), при этом использовать в циклах распределенный массив как приватизируемый или редукционный запрещено DVM-системой.

Ко второй категории относятся массивы, которые участвуют в операторах ввода-вывода, а также массивы, которые передаются как параметры во внешние процедуры (например, процедуры стандартных библиотек, либо процедуры, которые не доступны для анализа системе SAPFOR). В операторах ввода-вывода разрешается задавать только по одному массиву и только целиком из-за ограничений DVM-системы (то есть можно выводить целиком весь массив). Все остальные случаи отменяют распределение данного массива. Для того чтобы не отменять распределение массивов, которые участвуют в сложных операторах ввода-вывода, а также во внешних процедурах, требуется заводить массивы-копии и вставлять копирования до и после соответствующих операторов. В текущий момент в системе SAPFOR такое преобразование не автоматизировано.

К третьей категории относятся массивы, которые система SAPFOR автоматически отфильтровывает по тем или иным причинам. Процесс фильтрации состоит в том, чтобы запретить использовать распределяемые массивы, которые в дальнейшем будут отображены на разные деревья выравнивания и соответственно на разные DVMH-шаблоны в общем гнезде потенциально параллельных циклов. Разные деревья выравнивания образуются из несвязанных между собой графов массивов. Также отфильтровываются массивы, объявленные в глобальной области видимости, которые могли быть использованы в процедурах, вызываемых из цикла, без явного использования в самом цикле. Такое ограничение связано с тем, что в DVMH-модели во всех параллельных циклах необходимо «видеть» все используемые массивы. Те данные, которые используются в вызываемых из цикла процедурах, но не используются в циклах, считаются «невидимыми» для DVMH-компилятора, что приведет к ошибке конвертации программы и ее выполнения.

После выполнения фильтрации происходит распространение состояния потенциальной распределенности массивов в соответствии со связями фактических и формальных параметров процедур в программе. По массивам, которые не были отфильтрованы, будет построен граф массивов, которые в свою очередь будет отображен в дерево выравнивания в модели DVMH.

3.3. Анализ программы: построение связей циклов и массивов

На данном этапе системой SAPFOR для каждой функции в исходной программе выполняется анализ всех ее операторов. Для каждого обращения к массиву, находящемуся внутри гнезда циклов, для каждого из измерений (отдельно и независимо) выполняется следующий анализ:

- выполняется поиск косвенной адресации в обращении. Косвенная адресация не может быть эффективно распараллелена DVMH-моделью в случае структурированных сеток;
- выполняется поиск более, чем одной итерационной переменной цикла в индексном выражении в обращении к массиву. Для корректного отображения данных и связывания распределения вычислений с распределением данных, необходимо однозначное соответствие индексного выражения в обращении к массиву с итерационной переменной одного из циклов в гнезде. Если имеется связь с одним циклом, то выполняется попытка сопоставления индексного выражения с итерационной переменной цикла I с шаблоном $a * I + b$ и вычисления данных коэффициентов;
- выполняется проверка всех измерений массива. Если обращение к массиву используется на запись, проверяется факт того, что все индексные выражения в обращении к массиву имеют хотя бы одну итерационную переменную цикла. Если итерационная переменная цикла не встречается в индексных выражениях, то для рассматриваемого цикла отмечается факт наличия неопределенной запись в массив.

Важно отметить, что вся информация привязывается к циклам, то есть для каждого цикла формируется список всех обращений к массивам. Для того, чтобы цикл был оптимально распараллелен системой SAPFOR, требуется, чтобы каждое индексное выражение в обращении по конкретному измерению массива содержало только одну итерационную переменную цикла или иными словами должно быть однозначное отображение измерений массива на циклы.

После обработки функции будет построена общая информация о «хороших» обращениях к массивам с привязкой этих обращений к циклам с коэффициентами $a * I + b$, где $a, b > 0$ – вычисленные константы, $a \neq 0$, а I – итерационная переменная цикла. Остальные обращения к массивам будут порождать неизбежные обмены между узлами кластера.

3.4. Анализ программы: построение графа массивов

Граф массивов является основной структурой данных, на основе которой строятся распределение данных и вычислений. Построение графа массивов обусловлено выбором целевой модели при создании параллельной версии программы. DVMH-модель требует выполнения правила собственных вычислений: каждый процессор изменяет только собственные данные, то есть данные, которые распределены на этот процессор. Кроме того вычисление одной итерации цикла должно целиком выполняться на одном процессоре, и следовательно элементы массивов, вычисляемые на одной итерации оказываются связаны между собой и должны быть размещены на одном процессоре.

Для соблюдения данного правила необходимо использовать взаимное выравнивание массивов между собой с помощью спецификации ALIGN. После распределения массивов с помощью спецификаций DISTRIBUTE и ALIGN получается дерево выравнивания, которое описывает связи между всеми массивами. Правило собственных вычислений требует, чтобы все массивы, используемые в одном цикле, принадлежали одному дереву выравнивания.

Граф массивов представлен в системе SAPFOR в формате CSR (Compressed Sparse Rows) и является неориентированным. Каждое измерение массива становится узлом гра-

фа, а дуги показывают связь одного измерения массива с другим. Для заполнения графа массивов необходимо обработать информацию о вычисленных коэффициентах a, b в обращениях к массивам и их связях с циклом. Каждая дуга в графе массивов связывает одно измерение массива Arr_1 с измерением массива Arr_2. Дуги добавляются по следующему принципу (W или Write означает обращение на запись, R или Read означает обращение на чтение):

- связываются измерения массивов, обращения по которым присутствуют в левой части операторов присваивания в цикле с типом дуги запись-запись (связь $W - W$) и весом $Loop_w * SumB$;
- связываются измерения массивов Arr_1 и Arr_2, причем обращение к массиву Arr_1 содержится в левой части операторов присваивания, а обращение к массиву Arr_2 содержится в правой части операторов присваивания или в условиях IF, причем не обязательно, чтобы массивы Arr_1 и Arr_2 были в одном операторе. Связь создается с типом дуги запись-чтение (связь $W - R$) по данному циклу с весом $Loop_w * SumB$;
- в случае отсутствия операций записи в массивы связываются измерения массивов, обращения по которым присутствуют в правой части операторов присваивания или условиях IF в данном цикле, причем два обращения к разным массивам не обязаны быть в одном операторе. Связь создается с типом дуги чтение-чтение (связь $R - R$) по данному циклу с весом $Loop_w * SumB$.

Под $SumB$ понимается совокупное количество байт, которое потребуется передать другим процессорам в случае неудовлетворения конкретной связи с циклом. В худшем случае необходимо передать целиком все измерение массива, отображенное на соответствующий цикл. Количество байт вычисляется из размерности типа используемого массива и номера измерения массива. Данные о размерах массивов всегда известны системе SAPFOR и должны быть получены либо от статического, либо от динамического анализов, либо от пользователя, иначе невозможно построить дерево выравнивания в модели DVMH.

Под $Loop_w$ понимается вес цикла. В зависимости от количества арифметических операций и обращений к массивам в телах циклов тот или иной цикл с меньшим количеством витков может выполняться дольше аналогичного цикла, но с большим количеством витков. Вес цикла оценивается статическим образом, либо путем динамического профилирования (получение времени выполнения данного цикла). Вес цикла показывает сколько раз цикл был выполнен за все время работы программы. Например, если цикл выполняется всего один раз (цикл инициализации), то можно пожертвовать количеством коммуникаций в данном цикле в пользу итерационного цикла, который может выполняться сотни, а то и тысячи раз, где каждый лишний переданный байт будет серьезно сказываться на производительности программы в целом. В случае недостаточности информации для оценки веса цикла система SAPFOR полагает $Loop_w = 1.0$, что означает равенство всех циклов в программе.

В случае добавления дуги с одинаковыми вершинами происходит увеличение веса данной дуги путем суммирования текущего веса и веса добавляемой дуги. Данное правило позволит выделить наиболее важные связи, что позволит уменьшить количество пересылок между процессами при выборе определенной схемы распределения данных.

Для того чтобы отличать дуги по типу связи ($W-W, W-R, R-R$), необходимо расставить приоритеты. Тип дуги с типом связи $W - W$ имеет самый высший приоритет, однако все дуги с типом $W - W$ имеют равный приоритет между собой. Это объясняется тем, что неудовлетворение данной связи приведет к невозможности распараллеливания данного цикла, так как не будет выполнено правило собственных вычислений для связываемых массивов, что в итоге повлечет за собой бОльшие коммуникации (так как все обращения к распределенным массивам на чтение в таком цикле в худшем случае должны быть «покрыты» с помощью доступа к удаленным данным). Дуги с типом связи $W - R$ имеют приоритет

над дугами с типом связи $R-R$. Дуги каждого из типов $W-R$ и $R-R$ имеют также равный приоритет между собой.

Чтобы обеспечить соотношение приоритетов для разных типов дуг, был реализован следующий алгоритм выделения дуг по приоритетам. Сначала посчитаем общую сумму всех дуг с типом $R-R$, $S1 = \sum_{[R-R]}$. Затем прибавим число $S1$ к дугам с типом $W-R$. Тем самым мы «выделим» приоритет $W-R$ типа над дугами типа $R-R$, сохранив между тем равный приоритет между схожим типом. Затем вычислим общую сумму весов всех дуг с типами связи $W-R$ и $R-R$, $S2 = S1 + \sum_{[W-R]}$ и добавим теперь число $S2$ ко всем дугам с типом $W-W$. Таким образом, будет выполнено правило приоритета дуг: вес любой дуги с типом $W-W$ будет больше, чем $W-R$ и $R-R$, вес любой дуги с типом $W-R$ будет больше, чем $R-R$.

3.5. Поиск наилучшего связывания массивов

После того, как был построен общий граф массивов, необходимо выбрать то множество дуг, которое будет отражать наилучшие связи между массивами и минимизировать коммуникации между процессорами. Таким образом, необходимо построить усеченный графа массивов (такой граф массивов, который не содержит циклов, порождающих конфликтные ситуации) для последующего создания распределения данных. Конфликты могут быть двух типов. Рассмотрим данные типы конфликтов:

- наличие цикла в графе массивов, для вершин которого нельзя построить единственный вариант выравнивания измерений массивов между собой (где каждая вершина соответствует измерению массива). Такие конфликты будем называть конфликтами первого типа (1);
- присутствует явная или косвенная дуга (через другие дуги графа) между двумя измерениями одного и того же массива. Такие конфликты будем называть конфликтами второго типа (2).

Рассмотрим два подхода, которые позволяют сократить количество друг в графе массивов и найти совокупность наиболее важных дуг. Первый подход – перебор совокупности наиболее важных дуг в графе. В данном подходе мы выполняем перебор всевозможных наборов дуг и их оценку общего веса. Набор дуг с максимальным весов и будет решением данной задачи. Решение данной задачи можно представить в виде нескольких этапов.

Первый этап – получение всех простых циклов в графе массивов. Простой цикл в графе – это замкнутый цикл без повторного прохода по ребру или посещения вершины дважды, за исключением начальной и конечной вершин. В системе SAPFOR цикл представляет собой набор дуг графа массивов с сохранением веса.

Нахождение всех простых циклов в графе массивов является NP-трудной задачей, поэтому для ограничения поиска по времени и ресурсам в случае больших графов вводится два параметра, которые позволят ограничить этот поиск: максимальный размер цикла (размером цикла будем называть количество входящих в него дуг), который необходимо добавить в список простых циклов и максимальная длина просматриваемой цепочки при рекурсивном поиске таких циклов в графе.

Первый параметр используется для ограничения по памяти, второй – по времени поиска всех простых циклов. Можно отметить, что накладываемые ограничения не всегда позволяют найти все простые циклы в графе массивов, а это значит, что алгоритм не всегда может найти точное решение для больших графов массивов. С другой стороны – чем больше размер (длина) цикла, тем больше массивов он связывает между собой. Тем самым для получения наилучшего выравнивания массивов между собой не обязательно находить все простые циклы в графе.

Второй этап – обработка найденных простых циклов. После нахождения всех простых

циклов (далее просто циклов), происходит их сортировка по размеру (длине), а также разделение на независимые группы по длине. Затем для каждого цикла выполняется сортировка его дуг по весу, и все циклы в каждой группе упорядочиваются по суммарному весу. Данные сортировки позволят наиболее быстрым образом обеспечить поиск и удаление конфликтных дуг в графе массивов.

Для устранения конфликта (1) можно удалить любую из дуг цикла, например, с минимальным весом. Для устранения конфликта (2) необходимо удалить такую дугу, чтобы явная или косвенная связь между двумя измерениями массивов, которая выражена дугами графа, пропала. Если есть конфликтные циклы, то запускается процесс устранения конфликтов. Рассмотрим такой подход, который позволяет удалять конфликтные и неконфликтные дуги. После применения данного подхода мы получим усеченный граф массивов, который не содержит циклов.

Каждый цикл характеризуется суммарным весом всех его дуг или просто общим весом. Ранее все циклы были отсортированы с учетом их веса и размерности от самых маленьких до самых больших циклов по размерности (длине), а циклы с одинаковой размерностью были отсортированы по убыванию их веса.

Для устранения конфликтов запускается рекурсивная процедура. Цель данной процедуры – удалить дуги с минимальным суммарным общим весом, что позволит получить наиболее оптимальное с точки зрения обменов между узлами распределение данных на основе построения графа массивов и задания соответствующих весов. В начале процедуры имеем нулевой общий вес удаленных дуг и пустой список удаляемых дуг. Выбираем очередной цикл из списка полученных конфликтных циклов. Пытаемся по очереди удалить каждую из дуг данного цикла и смотрим, что получается:

- если это первая дуга цикла, то удаляем эту дугу, прибавляем вес этой дуги к общему весу всех удаленных дуг и заносим эту дугу в список удаляемых. После удаления данной дуги некоторые циклы из полученного списка перестанут быть циклами и, соответственно, не будут принимать участие в дальнейшем выборе. Далее рекурсивно вызываем эту процедуру. Рекурсивный вызов завершается в том случае, если нет больше циклов с конфликтами. В этом случае получен суммарный вес и список необходимых для удаления дуг;

- если эта дуга цикла не первая, то у нас уже имеется общий вес и список удаляемых дуг на каком-то конкретном уровне рекурсивной вложенности вызова рассматриваемой процедуры. Также мы имеем текущий суммарный вес на текущем уровне рекурсивной вложенности. Если сумма веса очередной удаляемой дуги и текущего общего суммарного веса больше, чем найденный наименьший общий вес удаляемых дуг, то рекурсивный вызов удаления этой дуги не выполняется, так как удаление этой дуги повлечет за собой увеличение общего веса всех удаляемых дуг (таким образом выполняется отсечение перебора всех вариантов наборов дуг для удаления). Если мы завершили рекурсивный вызов и получили меньший вес, чем был найден до этого, то корректируется общий вес и соответствующий список дуг для удаления.

После того, как мы получили список дуг для удаления, происходит формирование усеченного графа массивов, такого графа, который не содержит циклов. Затем необходимо проверить, не возникнут ли конфликты второго типа в усеченном графе, или нет ли таких путей в графе, которые косвенно (через другие массивы и соответствующие им дуги) связывают измерения одного и того же массива. Данная конфликтная ситуация не образует цикл, но требует разрешения.

Для этого необходимо для каждого массива для всех уникальных комбинаций пар его измерений добавить фиктивную дугу между этими измерениями с очень большим весом (например, большим, чем сумма всех весов в графе) и повторить весь алгоритм поиска простых циклов и удаления конфликтов. Если мы нашли конфликтный цикл, то мы удаляем

дугу. Из-за особенности алгоритма (применение сортировок и удаление дуг с минимальным совокупным весом) будет выбрана не фиктивная дуга, а существующая дуга в графе. Таким образом, будут разрешены все конфликты второго типа.

Стоит отметить, что после такой операции по данному графу массивов не всегда можно построить выравнивание всех массивов между собой, особенно если в графе было много конфликтов или были применены ограничения на поиск простых циклов, что является минусом данного алгоритма. Оценка сложности данного алгоритма является экспоненциальной в зависимости от количества вершин в графе, что накладывает ограничения на его применимость на больших программах. Данный алгоритм может быть использован на сравнительно небольших программах, где количество узлов графа массивов не более 10.

Альтернативным алгоритмом поиска наиболее важных дуг является модифицированный алгоритм поиска минимального остовного дерева. Минимальное остовное дерево – такое дерево, которое является максимальным по включению ребер подграфом, не имеющее циклов, и в котором сумма весов ребер минимальна. Если исходный граф связный, то будет построено остовное дерево, если же в исходном графе несколько несвязных компонент, то результатом будет остовный лес.

В нашей задаче требуется найти набор дуг с наибольшим весом. Таким образом, без изменения общности алгоритма поиска минимального остовного дерева можно искать максимальное остовное дерево – такое дерево, в котором сумма весов ребер максимальна. Была использована самая простая реализация – алгоритм Дейкстры-Прима. Недостатком данного алгоритма является тот факт, что решение получается не самое лучшее, как в случае полного перебора, так как не выполняется перебор всех цепочек дуг и их сравнение между собой. Главное достоинство такого подхода заключается в линейной оценке его сложности в зависимости от количества вершин в графе. Данное достоинство вместе с возможностью распараллелить данный алгоритм делает возможным его использование на любой программе с любым количеством массивов.

3.6. Создание вариантов распределения данных

После того, как был получен усеченный граф (далее граф массивов), который связывает измерения массивов в соответствии с их использованием в циклах программы, можно построить варианты (схемы) распределения данных.

Так как граф может быть не связным, в нем ищутся все деревья, которые связывают массивы друг с другом. После такого поиска мы знаем, сколько поддеревьев у нас есть, и какие массивы в эти поддеревья входят. Для каждого дерева создается свой шаблон в DVMH-модели – DVMH TEMPLATE, который и будет распределяться с помощью директивы DISTRIBUTE и на который будут выровнены все массивы данного дерева. Шаблон представляет собой виртуальный массив, под который не отводится память в программе.

Шаблон создается по массиву с наибольшей размерностью, а среди одинаковых массивов одной размерности, выбирается тот, который занимает больше памяти. После того, как создан шаблон и найден массив, по которому строится этот шаблон, в граф добавляются дуги, связывающие измерения этого массива и измерения этого шаблона с весом 1.0 и типом связи $R-R$. Таким образом, в графе массивов появляется шаблон, до которого можно «добраться» по связям между массивами и узнать выравнивание на него.

Так как с шаблоном связан только один из массивов, необходимо уметь вычислять связь с шаблоном и для других массивов. Наибольшая сложность, которая может возникнуть при вычислении таких связей – не кратные коэффициенты при выравнивании на шаблон. Такие коэффициенты могут возникать в том случае, когда измерение одного массива требуется распределить, например, с раздвижкой $3*i$, а измерение второго массива требуется распределить на то же измерение шаблона с раздвижкой $2*i$. Таким образом, требуется вычислить наименьшее общее кратное и изменить соответствующие атрибуты в графе.

Варианты распределения данных создаются для каждого шаблона независимо, количе-

ство вариантов для каждого шаблона будет равно $2^{d(T_i)}$, а общее их количество – $\sum_{i=1}^{K} 2^{d(T_i)}$, где $d(T_i)$ – размерность шаблона (каждое измерение считается распределенным, либо размноженным), K – количество построенных шаблонов.

Правила выравнивания массивов на шаблон в модели DVMH не зависят от выбранного в дальнейшем распределения этих шаблонов. Для каждого массива из под-дерева, который не является шаблоном, выполняется поиск связи его измерений с шаблоном в этом поддереве. Поиск связи для конкретного измерения массива запускается в том случае, если это измерение присутствует в графе массивов. В результате поиска может быть не найдено связи с шаблоном по конкретному измерению, такое измерение будет размножено (указана * в спецификации ALIGN).

3.7. Создание директив распределения вычислений

Для создания директивы распределения вычислений необходимо найти массив из графа массивов, на который будет отображаться пространство витков потенциально параллельного цикла. Если в рассматриваемом цикле присутствует запись всего в один массив – он и будет выбран. Если массивов на запись несколько, то выбор происходит, прежде всего, среди массивов, которые участвуют в спецификации ACROSS, если таковая имеется. В модели DVMH требуется, чтобы цикл был отображен на массив, который участвует в спецификации ACROSS. Данная спецификация позволяет организовать конвейерное выполнение циклов с регулярными зависимостями по данным, необходимость ее задания для цикла определяется системой SAPFOR автоматически на основе результатов анализа программы.

В результате алгоритма поиска наилучшего выравнивания данных, который был описан выше, был получен усеченный граф массивов. Конфликтные ситуации первого и второго типов были разрешены с помощью удаления соответствующих ребер этого графа. В результате чего, интересы тех или иных циклов могли быть не учтены, что может привести к тому, что нельзя создать директиву распределения вычислений для этого цикла.

Если цикл не содержит ограничений на распараллеливание (считается потенциально параллельным), то директива «привязывается» к соответствующему циклу в дереве циклов. Для директивы заполняется информация о массиве, на который требуется отобразить вычисления, правила отображения, а также сохраняются свойства цикла, которые были получены в результаты его анализа системой SAPFOR и должны быть описаны в параллельной программе в виде спецификаций DVMH языков: приватные и редукционные переменные, информация о регулярных зависимостях по данным, теневые грани массивов и элементы массивов, которые должны быть отдельно получены с удаленного процессора с помощью спецификации REMOTE_ACCESS.

Далее выполняется объединение полученных директив для каждого из циклов, если имеет место тесная вложенность. Если цикл не тесно вложенный, то запускается преобразование, которое позволяет на основе информации, полученной от статического и динамического анализов, произвести объединение циклов и сделать их тесно вложенными (внесение инварианта цикла). Данное преобразование выполняется на лету и позволяет устранить не тесную вложенность, если это возможно, без потери результатов анализа и построенных структур.

В итоге, самый верхний цикл будет содержать объединенную информацию от всех тесно вложенных циклов, следующий по уровню вложенности цикл будет содержать объединенную информацию от всех тесно вложенных циклов, которые расположены ниже по уровню вложенности и т.д. Такой подход позволяет выбирать разные циклы для распараллеливания программы в зависимости от выбранного варианта распределения данных.

В зависимости от выбранного варианта распределения данных (шаблонов) будут выбраны соответствующие директивы распределения вычислений, а также созданы дополнительные директивы перераспределения данных и доступа к удаленным данным, если это потребуется.

4. Вычислительные эксперименты

В данном разделе исследуется эффективность программ в модели DVMH, получаемых в результате применения описанных алгоритмов распределения данных и вычислений. Нами было выполнено сравнение параллельных версий, полученных с помощью системы SAPFOR для вычислительных приложений BT, CG и EP из пакета NAS Parallel Benchmarks [20], с MPI-версиями данных программ, написанными их разработчиками вручную.

Исследование эффективности было выполнено на суперкомпьютере K10 [29], состоящем из процессоров Intel Xeon E5-2660 и графических ускорителей NVIDIA Tesla M2090. Каждый узел содержит два 8-ти ядерных процессора (CPU), связанных посредством общей памяти (архитектура NUMA), и три графических ускорителя (GPU). Для вычислительных экспериментов были использованы максимально возможные ресурсы только процессоров одного, двух и девяти узлов (графические ускорители не использовались). Время выполнения MPI программ и DVMH программ, полученных с помощью SAPFOR с использованием языков FDVMH и CDVMH, приведены в таблице 1.

Таблица 1. Время выполнения в секундах программ на языке Фортран и Си, NPB 3.3 класс C.

	MPI Fortran			**FDVMH**			MPI C			**CDVMH**		
	BT	CG	EP	BT	CG	EP	BT	CG	EP	BT	CG	EP
1 узел	100	21.7	27.4	80	25.9	22.6	123.6	35.9	30.6	97.1	25.8	28
4 узла	34	11.04	7.09	24.8	47	5.67	34.7	10.2	7.85	30	47	7
9 узлов	16.26	7.51	3.25	14.9	60.1	2.52	17.10	7.05	3.43	16.57	60.7	3.13

Время выполнения оригинальных версий, написанных разработчиками пакета NAS Parallel Benchmark, приведено в первой и третьей группе столбцов (MPI программы).

Изначально рассматриваемые последовательные программы были написаны только на Фортране, поэтому потребовалось выполнить перевод рассматриваемых программ на язык Си вручную. Чтобы получить автоматически распараллеливаемые версии программ с помощью системы SAPFOR, были выполнены их предварительные преобразования (подстановка функций, объединение циклов, сужение размерности приватных массивов и др.) [16]. Вторая и четвертая группа столбцов показывают время выполнения этих параллельных версий программ с использованием FDVMH и CDVMH языков. Большинство преобразований было выполнено системой SAPFOR, другая часть преобразований, не реализованных на данный момент в системе или специфичных для конкретной программы и трудно формализуемых в виде отдельного преобразования, была выполнена вручную.

Программы BT и EP, полученные с помощью системы SAPFOR, показывают практически схожие ускорения, что и MPI программы, написанные вручную разработчиками тестов. Но программа CG, начиная с 16 и более процессов, начинает замедляться по отношению к MPI программе. Это связно прежде всего с тем, что основная доля времени приходится на умножение разреженной матрицы на вектор, что приводит к косвенной индексации. Данная особенность не может быть эффективно распараллелена в текущей модели на регулярных сеток. Для этого требуется модификация системы SAPFOR для использования нового расширения DVMH-модели для нерегулярных сеток. При распараллеливании в текущей модели, системе SAPFOR постоянно требуется выполнять одну коллективную операцию по пересылке удаленных данных между всеми процессами, что приводит к замедлению на их большом количестве. Если использовать графические ускорители, то 16 процессов будет достаточно для достижения высокой производительности, а пересылки не будут так сильно сказываться.

5. Заключение

В статье был рассмотрен подход к автоматизированному распараллеливанию программ для кластеров с помощью системы SAPFOR. Работа полученных параллельных программы была продемонстрирована на примере некоторых приложений из пакета NAS Parallel Benchmarks.

Предлагаемый нами подход к разработке параллельных программ сочетает модель программирования на основе директив (DVMH) и инструменты автоматизации и взаимодействия с пользователем (SAPFOR). Разрабатываемые системы взаимно дополняют друг друга, что, в свою очередь, позволяет получить существенное преимущество по сравнению с другими моделями параллельного программирования, такими как MPI+OpenMP+CUDA, при разработке параллельных программ для гетерогенных кластеров.

Основная сложность при распараллеливании программ на кластер заключается в минимизации накладных расходов, вызванных коммуникационными обменами между вычислительными узлами. В системе SAPFOR был реализован алгоритм распределения данных, учитывающий глобальные связи между массивами программы. Такие связи порождаются совместным использованием разных массивов в одном цикле и необходимостью соблюдения правила собственных вычислений при распараллеливании циклов в модели DVMH. Выбор между противоречивыми связями выполняется на основе оценки потенциально возможных коммуникаций, в случае не соблюдения одной из связей.

В систему SAPFOR входит автоматически распараллеливающий компилятор, который успешно справляется с потенциально параллельными программами без вмешательств со стороны пользователя. Если же реализованный в системе SAPFOR анализ кода не справляется, то пользователь может помочь системе с помощью соответствующих директив, указав недостающие свойства программы, либо выполнить с помощью системы SAPFOR необходимые преобразования, которые не приводят к снижению производительности и при этом повышают доступный уровень параллелизма. Можно отметить, что преобразования производятся на уровне исходного кода и не требуют детального изучения параллельных архитектур и языков параллельного программирования.

В сочетании с ранее выполненными работами [13,16] система SAPFOR может выполнять распараллеливание в модели DVMH на вычислительные системы в разной конфигурации, используя узлы кластера, а также графические ускорители и многоядерные процессоры внутри узла.

Таким образом, системы SAPFOR и DVM могут значительно сократить усилия, необходимые для распараллеливания программ, и позволяют задействовать все доступные внутри узла ресурсы (многоядерные процессоры и графические ускорители). Мы надеемся, что данный подход должен помочь эффективной разработке и оптимизации масштабируемых программ для суперкомпьютеров.

Литература

1. Штейнберг Б.Я., Штейнберг О.Б. Преобразования программ – фундаментальная основа создания оптимизирующих распараллеливающих компиляторов // Программные системы: теория и приложения, 2021, 12:1(48), с. 21–113. DOI: 10.25209/2079-3316-2021-12-1-21-113

2. Czarnul, P., Proficz, J., Drypczewski, K. Survey of methodologies, approaches, and challenges in parallel programming using high-performance computing systems // Scientific Programming, vol. 2020, P. 1058-–9244, 2020. DOI: 10.1155/2020/4176794

3. SYCL. URL: https://sycl.tech/

4. Celerity. High-level C++ for Accelerator Clusters. URL: /https://celerity.github.io/

5. Murai, H., Nakao, M., Shimosaka, T., Tabuchi, A., Boku, T., Sato, M. XcalableACC - a Directive-based Language Extension for Accelerated Parallel Computing // Supercomputing'14 poster, New Orleans, LA, USA, Nov. 2014

6. Konovalov, N.A., Krukov, V.A, Mikhajlov, S.N., Pogrebtsov, A.A. Fortan DVM: a Language for Portable Parallel Program Development // Programming and Computer Software. Vol. 21, No. 1. 1995. P. 35–38.

7. Бахтин В.А., Клинов М.С., Крюков В.А., Поддерюгина Н.В., Притула М.Н., Сазанов Ю.Л. Расширение DVM-модели параллельного программирования для кластеров с гетерогенными узлами // Вестник ЮУрГУ. Серия: Вычислительная математика и информатика, 2012. № 18(277). С. 82-92.

8. Бахтин, В.А., Колганов, А.С., Крюков, В.А., Поддерюгина, Н.В., Притула, М.Н. Методы динамической настройки DVMH-программ на кластеры с ускорителями // Суперкомпьютерные дни в России : Труды международной конференции, Москва, 28–29 сентября 2015 года, М.: Изд-во МГУ, 2015, С. 257-268), 2015. С. 257–268.

9. Hwu, W.-m., Ryoo, S., Ueng, S.-Z., Kelm, J.H., Gelado, I., Stone, S.S., Kidd, R.E., Baghsorkhi S.S., Mahesri, A.A., Tsao, S.C., Navarro, N., Lumetta, S.S., Frank, M.I., Patel, S.J.Implicitly parallel programming models for thousand-core microprocessors. // Proceedings of the 44th annual Design Automation Conference (DAC '07), ACM, New York, NY, USA. 2007. P. 754–759. DOI 10.1145/1278480.1278669

10. Baghdadi, R. et al. PENCIL: A Platform-Neutral Compute Intermediate Language for Accelerator Programming // 2015 International Conference on Parallel Architecture and Compilation (PACT). 2015. P. 138–149 DOI: 10.1109/PACT.2015.17.

11. Kruse, M. Introducing Molly: Distributed Memory Parallelization with LLVM // CoRR, vol. abs/1409.2088. 2014 DOI: https://doi.org/10.48550/arXiv.1409.208810.48550/arXiv.1409.2088

12. Vandierendonck, H., Rul, S., De Bosschere, K. The Paralax Infrastructure: Automatic Parallelization with a Helping Hand // 2010 19th International Conference on Parallel Architectures and Compilation Techniques (PACT). 2010. P. 389–399.

13. Катаев, Н. А., Колганов, А. С. Дополнительное распараллеливание MPI программ с помощью системы SAPFOR // Вычислительные методы и программирование. 2021. 22. С. 239–251. DOI 10.26089/NumMet.v22r415

14. Клинов М.С., Крюков В.А. Автоматическое распараллеливание Фортран-программ. Отображение на кластер // Вестник ННГУ, 2009. №2. С. 128–134.

15. Бахтин В.А., Бородич И.Г., Катаев Н.А., Клинов М.С., Ковалева Н.В., Крюков В.А., Поддерюгина Н.В. Диалог с программистом в системе автоматизации распараллеливания САПФОР // Вестник ННГУ, 2012. № 5(2). С 252-245.

16. Kataev, N. LLVM Based Parallelization of C Programs for GPU //Voevodin V., Sobolev S. (eds) Supercomputing. RuSCDays 2020. Communications in Computer and Information Science, vol 1331. Springer, Cham, 2020. P. 436–448. DOI: 10.1007/978-3-030-64616-5_38

17. Kataev, N. Interactive Parallelization of C Programs in SAPFOR // Scientific Services & Internet 2020. CEUR Workshop Proceedings, Vol. 2784, 2020. P. 139–148.

18. Kataev, N. Application of the LLVM Compiler Infrastructure to the Program Analysis in SAPFOR // Voevodin V., Sobolev S. (eds) Supercomputing. RuSCDays 2018.

Communications in Computer and Information Science, Vol. 965,. Springer, Cham, 2018. P. 487–499. DOI: 10.1007/978-3-030-05807-4_41

19. Kataev, N., Smirnov, A., Zhukov A. Dynamic data-dependence analysis in SAPFOR // CEUR Workshop Proceedings, Vol. 2543, 2020, P. 199–208. DOI: 10.20948/abrau-2019-62

20. NAS Parallel Benchmarks. URL: https://www.nas.nasa.gov/publications/npb.html (Дата обращения: 8 апреля 2021).

21. Amarasingh, S. P., Lam, M. S. Communication Optimization and Code Generation for Distributed Memory Machines // PLDI '93: Proceedings of the ACM SIGPLAN 1993 conference on Programming language design and implementation. 1993 P. 126–138. DOI: 10.1145/155090.155102

22. Zima, H., Bast H., Gerndt, M. SUPERB: A tool for semi-automatic MIMD/SIMD parallelization // Parallel Comput. vol. 6. 1988. P. 1–18. DOI: 10.1016/0167-8191(88)90002-6

23. Grosser, T., Groesslinger, A., Lengauer. C. Polly — performing polyhedral optimizations on a low-level intermediate representation // Parallel Processing Letters, Vol 22(04), 2012. DOI: 10.1142/S0129626412500107

24. Lattner, C., Adve, V.: LLVM: A Compilation Framework for Lifelong Program Analysis & Transformation // Proc. of the 2004 International Symposium on Code Generation and Optimization (CGO'04). Palo Alto, California, 2004. DOI: 10.1109/CGO.2004.1281665

25. Гервич, Л.Р., Кравченко, Е.Н., Штейнберг, Б.Я., Юрушкин, М.В. Автоматизация распараллеливания программ с блочным размещением данных // Сиб. журн. вычисл. математики / РАН. Сиб. отд-ние. - Новосибирск, 2015. – Т.18, № 1. С–41-53

26. Bondhugula, U., Hartono, A., Ramanujam, J., Sadayappan, P. A practical automatic polyhedral parallelizer and locality optimizer // SIGPLAN Notices, 43(6), 2008. P. 101–113. DOI: 10.1145/1375581.1375595

27. Bondhugula, U. Compiling Affine Loop Nests for Distributed-Memory Parallel Architectures // SC '13: Proceedings of the International Conference on High Performance Computing, Networking, Storage and Analysis. 2013. P. 1–12. DOI: 10.1145/2503210.2503289

28. Banerjee, P., et al. The Paradigm Compiler for Distributed-Memory Multicomputers // Computer, Vol. 28, Issue 10, IEEE Oct 1995, P. 37-47. DOI: 10.1109/2.467577

29. Гетерогенный кластер K10. URL: https://www.kiam.ru/MVS/resourses/k10.html