ОТЗЫВ

официального оппонента на диссертационную работу Лебедева Артема Сергеевича «Методы и средства распараллеливания программ линейного класса для выполнения на многопроцессорных вычислительных системах», представленную на соискание ученой степени кандидата технических наук по специальности
2.3.5. Математическое и программное обеспечение вычислительных систем, комплексов и компьютерных сетей

Высокопроизводительные параллельные вычислительные системы нашли широкое применение для решения практических задач в областях, требующих ресурсоемкого численного моделирования: вычислительная гидрогазодинамика, вычислительная биология и медицина, вычислительная химия, численный прогноз погоды. Развитие универсальных процессоров и специализированных вычислителей послужило причиной растущего многообразия вычислительного оборудования.

Многообразие вычислительных микроархитектур, сетевых топологий, неоднородность программного окружения затрудняют работу исследователя, для которого предпочтительнее фокусировка на особенностях решаемой прикладной задачи, а не на нюансах программирования вычислительного оборудования. Перенос программного кода не только с одной вычислительной архитектуры на другую, но и между сходными вычислительными системами может быть сопряжен со значительными трудозатратами, что может послужить причиной недостаточной проработки оптимизаций кода, и, как следствие, неэффективного использования ресурсов аппаратуры. Автоматическое распараллеливание программ является одним из подходов к решению обозначенных проблем.

Особый интерес для исследователей представляют участки со статическим потоком управления, удовлетворяющие критериям линейности программ (в западной литературе — affine programs, SCOP — static control parts), поскольку они представляют циклические конструкции, на исполнение которых может

быть затрачена большая доля времени выполнения программ для научных и инженерных расчетов.

Одним из наиболее распространенных видов суперкомпьютеров является однородный кластер, построенный на основе универсальных многоядерных процессоров x64, возможно с применением архитектуры NUMA в отдельных вычислительных узлах. Такие системы позволяют использовать многолетние наработки в виде отлаженных моделей и программных библиотек, не требуя значительных усилий для переноса программного кода, в отличие от систем с ускорителями типа GPU или Xeon Phi.

Работа посвящена распараллеливанию линейных программ для их выполнения как на рабочих станциях, так и на суперкомпьютерах, построенных на базе универсальных многоядерных процессоров, а именно решению актуальной **научной задачи** — разработке методов нахождения пространственных и временных отображений программ линейного класса, обеспечивающих локальность использования данных при их параллельном выполнении на многопроцессорных вычислительных системах.

Существующие подходы к решению этой научной задачи обладают как рядом достоинств, так и рядом недостатков, оказывающих существенное влияние на конечный результат — параллельную программу, на быстродействие которой существенно влияют выбранное расписание и размещение вычислений.

Учитывая вышесказанное, тема диссертационной работы является актуальной и имеющей практическое значение.

Обоснованность выводов и рекомендаций, сформулированных автором в диссертации, подтверждается достаточно полным анализом отечественной и зарубежной литературы по параллельным вычислениям, и, в частности, по модели многогранников.

Достоверность полученных автором научных результатов подтверждается корректным использованием в ходе исследования методов теории множеств, теории графов, линейной алгебры, выпуклого анализа,

дискретного программирования, описательной статистики. Результаты машинных экспериментов согласованы с теоретическими положениями.

Научная новизна исследований определена новым подходом к оценке аффинных отображений линейных программ, что позволило сформулировать новые критерии их оптимальности, **отличающиеся** возможностью ранжировать информационные зависимости и доступы к данным *для более гибкого количественного описания локальности использования данных*, а также разработать новый метод нахождения оптимальных пространственных и временных отображений программ линейного класса, **отличающийся** применением взвешенной суммы показателей качества решения.

Не менее значимым с точки зрения научной новизны является разработанный автором метод генерации параллельной MPI-программы, **не требующий** дублирования входных данных во всех исполняющихся процессах.

Оба разработанных автором метода применяются совместно при распараллеливании линейных программ для выполнения на системах с распределенной памятью.

Теоретическая значимость работы состоит в развитии подходов к исследованию влияния локальности использования данных на быстродействие программ линейного класса при их параллельном выполнении. Эти подходы могут быть применены при разработке методов и методик повышения быстродействия программ, затрачивающих большую часть времени выполнения на участки с циклическими конструкциями. Теоретические результаты, полученные автором, использованы в учебном процессе РТУ МИРЭА, что подтверждено соответствующим актом.

Практическая значимость работы связана с прикладной ориентацией на решение важной научной задачи, а именно заключается в разработанных компонентах, которые могут быть использованы в автоматически распараллеливающем трансляторе для поддержки распараллеливания программ, написанных на языке Си. Практическая значимость работы подтверждена актом

внедрения результатов работы в производственный процесс ООО «НПП САТЭК плюс» (г. Рыбинск), а также двумя зарегистрированными программами для ЭВМ.

Основные результаты диссертации отражены в 12 печатных работах общим объемом 11,5 п.л., авторский вклад 9,9 п.л., 7 из которых опубликованы в рецензируемых научных журналах, входящих в Перечень ВАК РФ, 2 — в сборниках трудов конференций, индексируемых Web of Science и Scopus, 3 — в иных сборниках тезисов докладов.

Автореферат соответствует основному содержанию диссертации.

Основное содержание диссертации изложено в четырех главах (помимо введения, заключения, списка литературы, приложений) и выдержано в традиционном стиле.

Введение содержит краткую характеристику работы в целом и обоснование необходимости решения задач, поставленных в диссертации.

В первой главе отмечается расширение области применения методов модели многогранников в системах трансляции для выполнения распараллеливающих преобразований программ, а также оптимизации локальности использования данных.

Рассмотрены все этапы распараллеливания программы в модели многогранников как для систем с общей, так и для систем с распределенной памятью. Автором проведен серьезный анализ публикаций, посвященных методам модели многогранников. Особое внимание уделено методам, нацеленным на оптимизацию локальности использования данных путем нахождения аффинных отображений программ линейного класса.

На основании рассмотренного материала выявлены достоинства и недостатки существующих методов, которые определили пути их совершенствования.

Во второй главе автором предложены критерии оптимальности аффинных отображений программ линейного класса, а также разработан метод

их нахождения как для систем с общей, так и для систем с распределенной памятью.

Для нахождения аффинных расписаний за основу взят жадный алгоритм П. Футриера, минимизирующий размерность многомерного расписания. Предложенное автором расширение этой схемы заключается в нахождении каждого отдельного компонента многомерного расписания в согласии с предложенными критериями оптимальности, а именно в согласии с идеей минимизации задержки использования данных.

Для нахождения аффинных размещений вычислений и данных продолжается процесс построения линейно независимых аффинных отображений в согласии с предложенными критериями оптимальности, а именно в согласии с идеей минимизации расстояния использования данных. Для систем с распределенной памятью нахождение размещений вычислений и данных выполняется совместно.

Автором доказаны два утверждения о верхней границе задержки и расстояния использования данных, что служит теоретическим обоснованием разработанного метода нахождения аффинных отображений.

В завершении главы приведен пример распараллеливания алгоритма LU-разложения, иллюстрирующий применимость предложенных критериев оптимальности аффинных отображений программ для количественного описания локальности использования данных, а также применимость разработанного метода нахождения оптимальных отображений.

В третьей главе автором разработан метод генерации параллельной MPI-программы, не требующий дублирования входных данных во всех исполняющихся процессах. Организация информационного обмена между параллельными процессами требуется для реализации вычислений на системах с распределенной памятью, и разработанный автором метод опирается на результат работы метода нахождения аффинных отображений (глава 2) и на результат кодогенерации по методу С. Бастуля. Автором определены множества

элементов данных, участвующих в актах информационного обмена — т. н. «многогранники коммуникаций». Приведен пример многогранников коммуникаций для параллельного MPI-варианта LU-разложения, иллюстрирующий применимость разработанного метода.

Также автором разработано специальное программное обеспечение для распараллеливания программ линейного класса, написанных на языке Си. Приводится архитектура транслятора текст-в-текст, определяющая конвейер трансляции как для систем с общей, так и распределенной памятью.

В четвертой главе автор приводит результаты экспериментальных исследований производительности параллельных программ, полученных применением двух средств: разработанного автором программного обеспечения и современного распараллеливающего транслятора pluto, также выполняющего трансляцию текст-в-текст. Показано, что разница в производительности параллельных вариантов программ обусловлена различными аффинными отображениями, которые находит каждое из рассматриваемых средств. По результатам экспериментов видно, что, применяя разработанные автором критерии оптимальности аффинных отображений и метод их нахождения, можно получить параллельные программы с лучшим быстродействием, чем при использовании целевых функций на основе лексикографического упорядочивания. Также по результатам экспериментов видно, что в среде с распределенной памятью параллельные варианты программ, полученные применением разработанных автором методов и средств, генерируют более интенсивный информационный обмен, но это не имеет корреляционной взаимосвязи с итоговым быстродействием.

В заключении сформулированы основные результаты работы, которые показывают, что поставленная цель — достигнута, а задачи — решены.

В представленных материалах диссертационной работы имеются следующие недостатки:

1. Желательно было бы привести формулировки задач таймирования и распределения операций и данных между процессорами с использованием терминологии теории расписаний.
2. В работе не обсуждается существование множества наборов аффинных отображений для линейной программы, удовлетворяющих разработанным критериям оптимальности. В практической части применяется только одно найденное решение для каждой рассматриваемой программы.
3. В заключении к главе 2 приводится рекомендация по применению разработанного метода нахождения аффинных отображений программ в JIT-средах. При этом автор не приводит конкретных механизмов, благодаря которым может быть ослаблена параметризация программ. Желательно было бы привести примеры ситуаций, или обозначить классы программ, которые могут быть сведены к линейным в процессе JIT-трансляции.
4. Описание разработанного автором метода генерации параллельной MPI-программы в главе 3 опирается на блочную схему распределения процессоров. Автор не приводит обсуждения границ применимости предложенного формализма «многогранник коммуникаций»: неясно, применим ли он для циклической и блочно-циклической схемы распределения процессоров.
5. В главе 3 автор приводит методику оптимизации ветвлений. Несмотря на то, что приведен ассемблерный листинг, иллюстрирующий множество инструкций в накладных расходах при распараллеливании, действительный эффект от применения методики экспериментом не подкреплен.
6. Графики, иллюстрирующие разнородную нагрузку при параллельных MPI-запусках программ в главе 4, имеют не вполне понятные подписи сверху. Очевидно, что они обозначают схему распределения процессов по узлам, но в тексте работы расшифровки не приводится.

Однако, отмеченные недостатки не снижают общей научной и практической значимости работы.

В целом следует считать, что представленные результаты исследований Лебедева А.С. являются научной квалификационной работой, в которой содержится решение важной и актуальной научной задачи, позволяющее улучшить быстродействие параллельных программ, получаемых в результате применения средств автоматического распараллеливания, за счет нахождения аффинных отображений, улучшающих локальность использования данных.

Диссертация соответствует требованиям «Положения ВАК», предъявляемым к кандидатским диссертациям, соответствует специальности 2.3.5. Математическое и программное обеспечение вычислительных систем, комплексов и компьютерных сетей, а ее автор, Лебедев Артем Сергеевич, заслуживает присуждения ученой степени кандидата технических наук.

Официальный оппонент:

Доктор технических наук (05.13.06), профессор Института системной и программной инженерии и информационных технологий (Институт СПИНТех) Федерального государственного автономного образовательного учреждения высшего образования «Национальный исследовательский университет «Московский институт электронной техники» (Национальный исследовательский университет «МИЭТ»)

Портнов Евгений Михайлович

Подпись Е.М. Портнова удостоверяю:

20 мая 2024 г.

Отзыв официального оппонента
кандидата физико-математических наук, ведущего научного сотрудника Федерального государственного учреждения «Федеральный исследовательский центр Институт прикладной математики им. М.В. Келдыша Российской академии наук» Левченко Вадима Дмитриевича
на диссертацию Лебедева Артема Сергеевича
«Методы и средства распараллеливания программ линейного класса для выполнения на многопроцессорных вычислительных системах»,
представленную на соискание ученой степени кандидата технических наук по специальности 2.3.5. Математическое и программное обеспечение вычислительных систем, комплексов и компьютерных сетей

1. **Актуальность темы диссертационной работы**

Современные высокопроизводительные вычислительные системы, применяемые для решения как фундаментальных, так и прикладных задач в науке и экономике, обладают высокой степенью параллелизма вычислений и развитой иерархией подсистемы хранения данных. К тенденциям последних лет можно отнести распространение многоядерных микропроцессоров с многоуровневым кэшем, использование аппаратной векторизации. Современные узлы вычислительных кластеров содержат несколько таких процессоров, ускорители вычислений (в том числе графические), неоднородную архитектуру оперативной памяти, и объединяются высокоскоростными коммуникациями. Нетривиальной проблемой для таких систем является как использование без трудоемкой адаптации многолетних наработок в виде программных библиотек, так и ускорение решения прикладных задач специалистами предметных областей, часто не принимающих во внимание технические особенности программирования конкретной вычислительной системы. Это может повлечь неэффективное использование ресурсов вычислительного оборудования. Одним из подходов к решению обозначенных проблем является автоматическое распараллеливание программ. Тема диссертационной работы является востребованной и актуальной в силу следующих обстоятельств:

1). работа посвящена распараллеливанию программ линейного класса, часто встречающихся в виде циклических конструкций в коде приложений научных и инженерных расчетов, а это наиболее вычислительно емкие участки, представляющие интерес для оптимизации;
2). в качестве целевой параллельной архитектуры в работе рассматриваются универсальные многоядерные процессоры и построенные на их основе системы с распределенной памятью, что соответствует наиболее распространенным рабочим станциям и кластерным системам на основе многопроцессорных NUMA-серверов;

3). объектом исследования является распараллеливающий транслятор, и работа сближает перспективные методы модели многогранников (как существующие, так и разработанные лично автором в рамках диссертационного исследования) с инженерной практикой оптимизации программ перед итоговой компиляцией.

2. Научная новизна полученных результатов

К новым научным результатам, полученным в диссертационной работе, можно отнести следующее:

1). разработку новых критериев оптимальности пространственных и временных отображений программ линейного класса, отличающихся возможностью ранжировать информационные зависимости и доступы к данным для более гибкого количественного описания локальности использования данных;

2). разработку нового метода нахождения оптимальных пространственных и временных отображений программ линейного класса, отличающегося применением взвешенной суммы показателей качества решения;

3). разработку нового метода генерации параллельной MPI-программы, не требующего дублирования входных данных во всех исполняющихся процессах, что позволяет сократить накладные расходы памяти на поддержку распределенных вычислений.

3. Степень обоснованности и достоверности научных результатов, выводов и заключений, содержащихся в диссертации

Обоснованность научных положений диссертационной работы Лебедева А.С. подтверждается достаточным объемом проанализированных отечественных и зарубежных источников по тематике работы.

Как следует из введения, **целью** диссертационной работы является повышение быстродействия программ, получаемых в результате применения средств автоматического распараллеливания.

Для достижения поставленной цели автором был проведен анализ существующих методов и средств нахождения пространственных и временных отображений программ линейного класса, ориентированных на распараллеливание гнезд циклов, установлены ограничения и недостатки, обозначившие направления их совершенствования; разработаны новые критерии оптимальности пространственных и временных отображений программ линейного класса, отличающиеся возможностью ранжировать информационные зависимости и доступы к данным для более гибкого количественного описания локальности использования данных, чем целевые функции на основе лексикографического упорядочивания; разработан новый метод нахождения оптимальных пространственных и временных отображений программ линейного класса для распараллеливания гнезд циклов, устраняющий

необходимость полного перебора решений с их трудоемкой оценкой качества; разработан метод генерации параллельной MPI-программы, позволяющий организовать информационный обмен между параллельными процессами в случае явно заданного распределения данных между процессорами, при этом нет необходимости размещать входные данные на всех вычислительных устройствах; получены экспериментальные результаты исследования производительности параллельных вариантов программ (lu, atax, syr2k, floyd, gramschmidt), свидетельствующие о выигрыше в эффективности распараллеливания по сравнению с современным транслятором Pluto.

Достоверность полученных результатов работы подтверждена согласованностью теоретических и экспериментальных исследований, актами о внедрении, свидетельствами о регистрации программ для ЭВМ, участием в научных конференциях российского и международного уровня.

Выводы и заключения, представленные в работе, сделаны на основании фактического материала и его анализа при проведении теоретических и экспериментальных исследований.

4. Теоретическая и практическая значимость полученных в диссертации результатов

Теоретическая значимость результатов работы состоит в развитии подходов к исследованию влияния локальности использования данных на быстродействие программ линейного класса при их параллельном выполнении. Эти подходы могут быть применены при разработке новых методов повышения быстродействия линейных программ на высокопроизводительных рабочих станциях и кластерных системах.

Практическая значимость работы заключается в разработанных компонентах, которые могут быть использованы в автоматически распараллеливающем трансляторе для поддержки распараллеливания программ, написанных на языке Си: компонент нахождения пространственных и временных отображений программ линейного класса, скрипт расстановки директив OpenMP, библиотека макросов для организации информационного обмена и скрипт постобработки параллельных циклов для реализации блочной схемы распределения процессоров в MPI-программе.

Также практическую ценность представляют две методики, изложенные в главе 3:

1). методика расстановки конструкций информационного обмена, реализованных в библиотеке макросов (для языка Си);
2). методика оптимизации ветвлений для сокращения накладных расходов на распараллеливание.

Указанные методики изложены с достаточно высокой степенью формализации, что позволяет рекомендовать их программную реализацию.

Значимость результатов диссертационной работы для практики подтверждена актом о внедрении результатов в деятельность ООО «НПП САТЭК плюс» и двумя свидетельствами о регистрации программ для ЭВМ.

Полученные теоретические результаты использованы в учебном процессе Федерального государственного бюджетного образовательного учреждения высшего образования «МИРЭА — Российский технологический университет», что подтверждено соответствующим актом.

5. Структура работы

Диссертация состоит из введения, 4 глав, заключения и 7 приложений. Полный объём диссертации составляет 297 страниц, включая 39 рисунков и 44 таблицы. Список литературы содержит 124 наименования. Выводы приведены в конце каждой главы, основные теоретические и практические результаты сформулированы в заключении.

6. Соответствие работы требованиям, предъявляемым к диссертациям

Диссертация обладает внутренним единством, имеет научную новизну, теоретическую и практическую значимость, раскрывает сущность выполненного исследования, содержит обоснования полученных автором результатов и описание их практической реализации. Диссертация является логически завершенным изложением результатов научного исследования, выполненного автором.

Содержание диссертации соответствует научной специальности 2.3.5. Математическое и программное обеспечение вычислительных систем, комплексов и компьютерных сетей. Отраженные в диссертации научные положения соответствуют п.8. «Модели и методы создания программ и программных систем для параллельной и распределенной обработки данных, языки и инструментальные средства параллельного программирования».

Диссертация соответствует критериям Положения о присуждении ученых степеней.

7. Публикации и апробация результатов работы

Результаты диссертации отражены в 12 печатных работах общим объемом 11,5 п.л., авторский вклад 9,9 п.л., 7 из которых опубликованы в рецензируемых научных журналах, входящих в Перечень ВАК РФ, 2 — в сборниках трудов конференций, индексируемых Web of Science и Scopus, 3 — в иных сборниках тезисов докладов. Зарегистрированы 2 программы для ЭВМ.

Результаты работы апробированы на конференциях: Третий Национальный Суперкомпьютерный Форум (НСКФ-2014), Пятый Национальный Суперкомпьютерный Форум (НСКФ-2016), 11th IEEE International Conference on Application of Information and Communication Technologies (AICT-2017), 3rd International Conference «Futuristic Trends in Networks and Computing Technologies» (FTNCT-2020), Всероссийская научно-

техническая конференция «Многопроцессорные вычислительные и управляющие системы» (МВУС-2022).

8. Соответствие автореферата основным положениям диссертации

Автореферат и диссертация написаны технически грамотным языком, соответствуют требованиям ГОСТ Р. 7.0.11-2011 «Диссертация и автореферат диссертации». Структура и содержание автореферата полностью соответствуют основным положениям диссертации.

9. Замечания по работе.

1). В работе не обсуждается применимость разработанных методов совместно с техниками тайлинга (раздел 1.3.4).
2). В разделе 2.4 указано, что теоретические результаты могут использоваться для распараллеливания на графических ускорителях, но не обсуждаются подходящие для этого схемы сопоставления виртуальных процессоров нитям GPU.
3). Из рассмотренных в главе 4 примеров выделяется программа atax, при распараллеливании которой изучалось несколько дополнительных вариантов, включая модификацию исходного кода. К сожалению, влияние свойств полученных параллельных вариантов atax на итоговую производительность подробно не обсуждается.
4). В распараллеливании на основе OpenMP используется механизм планирования по умолчанию. Неясны причины, по которым автор не использует, например, динамическое планирование.
5). При проведении экспериментальных исследований автор применяет достаточно старую версию компилятора gcc 4.8.5.
6). Числа в таблицах со статистикой приведены с избыточной для возможностей человеческого восприятия точностью.
7). Из замеченных опечаток, на стр. 11 прилагательное «вычислительный» должно быть во множественном числе, на стр. 24 пропущена буква «д» в наречии «между», слово «кэш» один раз написано через «е» и несколько раз через «э».

Перечисленные замечания не снижают научной и практической ценности диссертационной работы и не являются определяющими при ее оценке.

10. Заключение о соответствии диссертации критериям, установленным Положением о присуждении ученых степеней

Диссертационная работа Лебедева Артема Сергеевича, представленная на соискание ученой степени кандидата технических наук, является научно-квалификационной работой, в которой решена актуальная **научная задача** — разработка методов нахождения пространственных и временных отображений программ линейного класса, обеспечивающих локальность использования данных при их параллельном выполнении на многопроцессорных

вычислительных системах, что соответствует п.9 «Положения о присуждении ученых степеней» ВАК РФ.

Диссертация имеет прикладной характер, в ней приведены сведения о практическом использовании полученных научных результатов, предложенные решения аргументированы (п.10 «Положения о присуждении ученых степеней» ВАК РФ).

Работа отвечает пп. 13, 14 «Положения о присуждении ученых степеней» ВАК РФ, предъявляемым к диссертациям на соискание ученой степени кандидата технических наук.

В целом, диссертация «Методы и средства распараллеливания программ линейного класса для выполнения на многопроцессорных вычислительных системах» на соискание ученой степени кандидата технических наук в полной мере соответствует требованиям действующего «Положения о присуждении ученых степеней», а ее автор Лебедев Артем Сергеевич заслуживает присуждения ученой степени кандидата технических наук по специальности 2.3.5. Математическое и программное обеспечение вычислительных систем, комплексов и компьютерных сетей.

Официальный оппонент:

Кандидат физико-математических наук (05.13.16), ведущий научный сотрудник Федерального государственного учреждения «Федеральный исследовательский центр Институт прикладной математики им. М.В. Келдыша Российской академии наук» (ИПМ им. М.В. Келдыша РАН)

Левченко Вадим Дмитриевич

08.05.24

Подпись В.Д. Левченко удостоверяю:

Ученый секретарь
ИПМ им. М.В. Келдыша РАН,
к.ф.-м.н.
(должность)

(подпись)

Давыдов
Александр Александрович
(Фамилия И.О.)