УТВЕРЖДАЮ
Проректор по научной работе
Национальный исследовательский Нижегородский государственный
университет им. Н.И. Лобачевского
доктор физико-математических наук, доцент

Иванченко М.В.
«2» марта 2021 г.

ОТЗЫВ ВЕДУЩЕЙ ОРГАНИЗАЦИИ

на диссертационную работу ГАРЕЕВА Романа Альбертовича «Методы оптимизации выполнения тензорных операций на многоядерных процессорах», представленную на соискание ученой степени кандидата физико-математических наук по специальности 05.13.11 – Математическое и программное обеспечение вычислительных машин, комплексов и компьютерных сетей.

Машинное обучение, спектральные методы, квантовая химия и другие научные дисциплины широко применяют тензорные операции. Например, тензорные операции используются при решении дифференциальных и интегральных уравнений, описывающих многие прикладные задачи. Большая часть подходов, применяемых для сокращения времени выполнения операций над тензорами на многоядерных процессорах, требуют доступа к целевой аппаратной платформе и больших временных затрат. Важной проблемой является разработка методов сокращения времени выполнения тензорных операций и их автоматическое распараллеливание, применимое в процессе компиляции, ограниченной по времени.

В диссертационной работе Р.А. Гареева разработано расширение модели целевой архитектуры процессора Лоу, которое позволяет сократить время выполнения матрично-векторных операций и их обобщений на замкнутые полукольца с элементами из множества вещественных чисел. Р.А. Гареевым созданы новые алгоритмы выполнения тензорных операций константной сложности относительно размерности тензоров, использующие разработанное расширение целевой архитектуры процессора Лоу. В диссертации показано, как предложенное решение и алгоритмы могут быть использованы для создания программной системы, выполняющей автоматическую оптимизацию времени выполнения тензорных операций и их автоматическое распараллеливание при компиляции программ для многоядерных процессоров общего назначения.

Учитывая теоретическую и практическую значимость решения проблем сокращения времени выполнения тензорных операций на многоядерных процессорах и их автоматического распараллеливания, считаем, что выбранная соискателем тема исследования и предлагаемые им методы и подходы, несомненно, являются **актуальными**.

Целью проведенного диссертационного исследования было совершенствование методов обработки информации путем сокращения времени выполнения многопоточных реализаций тензорных операций на многоядерных процессорах общего назначения без ручной настройки и автонастройки. Для достижения этой цели соискатель ставит и решает следующие задачи.

1. Разработка модели целевой архитектуры процессора с целью сокращения времени выполнения матрично-векторных операций и их обобщений на замкнутые полукольца с элементами из множества вещественных чисел.

2. Разработка новых алгоритмов выполнения тензорных операций константной сложности относительно размерности тензоров, уменьшающих время выполнения таких операций.

3. Разработка программной системы для автоматической оптимизации времени выполнения тензорных операций и их автоматического распараллеливания при компиляции программ для многоядерных процессоров общего назначения.

4. Проведение вычислительных экспериментов, подтверждающих эффективность разработанной программной системы по сравнению с аналогами, использующими ручную настройку и автонастройку.

Все перечисленные задачи были успешно решены. Получен ряд новых научных результатов, основными из которых являются следующие:

1. Разработано расширение модели целевой архитектуры процессора Лоу, которое обеспечивает сокращение времени выполнения матрично-векторных операций и их обобщений.

2. Разработаны новые алгоритмы выполнения тензорных операций константной сложности относительно размерности тензоров, уменьшающие время выполнения таких операций.

3. Разработана новая программная система автоматической оптимизации тензорных операций для автоматической оптимизации времени выполнения тензорных операций и их автоматического распараллеливания при компиляции программ для многоядерных процессоров общего назначения. Экспериментально подтверждена применимость программной системы автоматической оптимизации тензорных операций для уменьшения времени их выполнения.

Все результаты являются новыми. Научная новизна результатов, полученных в диссертации, заключается в следующем:

1. Разработано расширение модели целевой архитектуры процессора Лоу, которое позволяет сократить время выполнения матрично-векторных операций и их обобщений на замкнутые полукольца с элементами из множества вещественных чисел.

2. Разработаны новые алгоритмы выполнения тензорных операций константной сложности относительно размерности тензоров, уменьшающие время выполнения таких операций. Выведены

новые формулы, позволяющие получить значения параметров алгоритмов выполнения тензорных операций в зависимости от характеристик многоядерных процессоров общего назначения для архитектур x86-64, x86, ppc64le, aarch64. Впервые доказаны утверждения о существовании значений параметров, при которых отсутствует простой конвейера векторных инструкций модели целевой архитектуры процессора для представленных алгоритмов при возможности мгновенной загрузки данных из памяти на векторные регистры.

3. Разработана новая программная система для автоматической оптимизации времени выполнения тензорных операций и их автоматического распараллеливания при компиляции программ для многоядерных процессоров общего назначения. Получена новая оценка производительности многопоточной программы.

Теоретическая значимость. Теоретическая ценность диссертации состоит 1) в разработке расширения модели целевой архитектуры процессора Лоу, которое обеспечивает сокращение времени выполнения матрично-векторных операций и их обобщений на замкнутые полукольца с элементами из множества вещественных чисел; 2) в разработке новых алгоритмов выполнения тензорных операций константной сложности относительно размерности тензоров, уменьшающих время выполнения таких операций. С помощью моделирования выполнения представленных алгоритмов на расширенной модели целевой архитектуры процессора Лоу выведены новые аналитические зависимости, позволяющие получить значения параметров алгоритмов выполнения тензорных операций в зависимости от характеристик многоядерных процессоров общего назначения.

Практическая ценность. Практическую ценность работы определяет разработанная программная система автоматической оптимизации тензорных операций, которая может применяться для автоматической оптимизации времени выполнения тензорных операций и их автоматического распараллеливания при компиляции программ для многоядерных процессоров общего назначения.

Соответствие диссертации паспорту научной специальности. Полученные автором диссертации результаты имеют научную новизну для области создания программ и программных систем для параллельной обработки данных, а ее содержание полностью соответствует паспорту специальности 05.13.11 - «Математическое и программное обеспечение вычислительных машин, комплексов и компьютерных сетей», а именно п.8 «Модели и методы создания программ и программных систем для параллельной и распределенной обработки данных, языки и инструментальные средства параллельного программирования».

Рекомендации по использованию результатов диссертации. Результаты, изложенные в диссертации, могут быть эффективно использованы при реализации численных решений задач математической физики, математической геофизики, механики, квантовой химии на многоядерных процессорах общего назначения.

Обоснованность и достоверность. Результаты исследования подтверждаются данными экспериментов, выполненных в соответствии с общепринятыми стандартами.

В кратком изложении структура и содержание диссертации состоит в следующем.

Содержание диссертации

Диссертация состоит из введения, четырех глав, заключения, библиографии и приложения. Объем диссертации составляет 179 страниц, объем библиографии – 177 наименований.

Во введении дается общая характеристика работы, обосновывается актуальность темы проведенных исследований, формулируется цель работы, содержится обзор работ, близких к теме диссертации, и приводится краткое содержание по главам.

В первой главе рассматриваются основные подходы сокращения времени выполнения тензорных сверток. Описываются методы сокращения времени выполнения свертки тензоров, использующие ручную настройку, автонастройку и моделирование вычислений на целевой архитектуре процессора. Показано, как задача сокращения времени выполнения свертки тензоров сводится к задаче сокращения времени выполнения матрично-векторных операций.

Во второй главе представлено расширение модели целевой архитектуры процессора Лоу с инструкциями предвыборки и операциями из замкнутых полуколец с элементами из множества вещественных чисел. Представлены новые алгоритмы для вычисления матричных и матрично-векторных произведений, обобщенных на замкнутые полукольца. С помощью моделирования выполнения описанных алгоритмов на целевой архитектуре процессора выведены аналитические зависимости, позволяющие получить значения параметров описанных алгоритмов в зависимости от характеристик многоядерных процессоров общего назначения. Разработан новый алгоритм автоматического сокращения времени выполнения тензорных операций.

Третья глава посвящена разработке архитектуры программной системы автоматической оптимизации тензорных операций на основе моделей, алгоритмов и формул, предложенных во второй главе для оптимизации тензорных операций без выполнения автонастройки и ручной настройки. Приводится описание архитектуры программной системы автоматической оптимизации тензорных операций и рассматриваются ее возможные реализации.

В четвертой главе выполнена оценка эффективности программой система автоматической оптимизации тензорных операций при решении обратной задачи гравиметрии, общей задачи о путях, оптимизации тензорных операций. Приведены и проанализированы результаты масштабных вычислительных экспериментов на разных вычислительных архитектурах.

В **заключении** приводятся основные результаты диссертации и план дальнейшей работы в выбранном направлении.

Публикации и апробация результатов. Основные результаты, полученные соискателем в рамках диссертационного исследования, достаточно полно представлены в 7 печатных работах, из них 2 – в изданиях из перечня ВАК, 2 – в изданиях, проиндексированных в Web of Science, 2 – в изданиях, проиндексированных в SCOPUS. Кроме того, для случая матричного произведения,

обобщенного на замкнутые полукольца, созданная Р.А. Гареевым оптимизация была внедрена в основной код Polly проекта Low Level Virtual Machine.

Материалы диссертации успешно прошли апробацию на российских и международных конференциях. Автореферат диссертации правильно и полно отражает содержание работы.

Замечания по диссертации

1. В тексте присутствуют стилистические ошибки и опечатки. Например, на странице 5 слово «используется» должно быть употреблено во множественном числе. На странице 10 нужно убрать повторение имени «John Gunnels». На странице 14 вместо слова «метод» следовало бы использовать «методология». На странице 96 присутствует опечатка в слове «результаты». На странице 101 в таблице 4.8. дважды упомянута библиотека Intel MKL. На странице 155 аббревиатура «ГПУ» указана без расшифровки. На странице 179 не расшифрованы такие обозначения как «APP», «OpenMP» и «ГПУ».

2. В диссертационной работе не обсуждаются вопросы применимости разработанных способов декомпозиции данных при вычислениях с использованием графических процессоров - с общих позиций представляется, что разработанные подходы могут быть использованы и в случае ГПУ (по крайней мере, хотя бы частично).

3. В диссертационной работе утверждается, что выведенные аналитические зависимости позволяют получить значения параметров представленных алгоритмов выполнения тензорных операций в зависимости от характеристик многоядерных процессоров общего назначения для архитектур x86-64, x86, ppc64le, aarch64. Несмотря на это, применимость представленного подхода экспериментально показана только для небольшого количества процессоров с указанными архитектурами. Кроме этого, дополнительный интерес представляет исследование применимости представленного подхода для процессоров с архитектурами отличными от x86-64, x86, ppc64le, aarch64.

4. В диссертации отсутствуют технические параметры разработанной программной системы AOTO (размер программного кода, количество реализованных программных модулей и т.п.), которые позволили бы оценить трудоемкость выполненной разработки.

5. При описании выполненных вычислительных экспериментов в разделах 4.3.3-4.3.6 отсутствуют сведения об используемой точности вычислений для вещественной арифметики (одинарная или двойная точность) – понятно, что при использовании арифметики разной точности получаемые результаты могут значительно различаться.

6. Эксперименты, выполненные на процессорах фирмы Intel, проведены с использованием технологии Turbo Boost, которая может автоматически приводить к увеличению тактовой частоты процессора свыше номинальной, но не гарантирует этого. Несмотря на то, что наибольшее увеличение частоты, которое может быть достигнуто с помощью технологии Turbo Boost, учтено в представленных верхних теоретических границах на производительность, неизвестны фактические частоты процессоров, которые могли быть меньше, чем использованные.

Высказанные замечания не снижают значимости полученных результатов и высокой оценки общего научного уровня диссертационной работы.

Заключение

Диссертационная работа Р. А. Гареева представляет собой самостоятельно выполненную и законченную научно-квалификационную работу, в которой автором разработаны новые модели, методы и алгоритмы для решения задач автоматического распараллеливания и сокращения времени выполнения тензорных операций на многоядерных процессорах. Решение указанных задач имеет существенное значение для развития области создания программ и программных систем для параллельной обработки данных. Диссертационная работа соответствует требованиям Положения о порядке присуждения ученых степеней, включая п. 9, а ее автор, Гареев Роман Альбертович, достоин присуждения ему ученой степени кандидата физико-математических наук по специальности 05.13.11 «Математическое и программное обеспечение вычислительных машин, комплексов и компьютерных сетей».

Диссертация и отзыв обсуждены и одобрены на заседании объединенного научного семинара кафедры математического обеспечения и суперкомпьютерных технологий и кафедры программной инженерии Федерального государственного автономного образовательного учреждения высшего образования «Национальный исследовательский Нижегородский государственный университет им. Н.И. Лобачевского» (Протокол № 7 от « 25 » февраля 2021 г.).

Мы, Гергель Виктор Павлович и Мееров Иосиф Борисович, даем согласие на включение наших персональных данных в документы, связанные с работой Диссертационного совета ЮУрГУ 212.298.18, и их дальнейшую обработку.

Заведующий кафедрой программной инженерии Национального исследовательского Нижегородского государственного университета им. Н.И. Лобачевского, доктор технических наук (05.13.12 – Системы автоматизации проектирования), профессор Гергель Виктор Павлович

Заместитель заведующего кафедрой математического обеспечения и суперкомпьютерных технологий Национального исследовательского Нижегородского государственного университета им. Н.И. Лобачевского, кандидат технических наук (05.13.18 – Математическое моделирование, численные методы и комплексы программ), доцент Мееров Иосиф Борисович

« 2 » марта 2021 г.

Национальный исследовательский Нижегородский государственный университет им. Н.И. Лобачевского,
603022, г. Нижний Новгород, пр. Гагарина, 23.
Тел. +7 (831) 462-30-03
e-mail: gergel@unn.ru, meerov@vmk.unn.ru.

Подпись Гергеля В. П. Мееров И. Б.
Канцелярия
Зав. канц. Ученый секретарь ННГУ
Л.Ю. Черноморская
Тел. 462-30-21

СВЕДЕНИЯ О ЛИЦАХ, УТВЕРДИВШИХ И ПОДГОТОВИВШИХ ОТЗЫВ ВЕДУЩЕЙ ОРГАНИЗАЦИИ

на диссертацию Гареева Р.А. «Методы оптимизации выполнения тензорных операций на многоядерных процессорах», представленную на соискание ученой степени кандидата физико-математических наук по специальности 05.13.11 – Математическое и программное обеспечение вычислительных машин, комплексов и компьютерных сетей

Наименование организации	Федеральное государственное автономное образовательное учреждение высшего образования «Национальный исследовательский Нижегородский государственный университет им. Н.И. Лобачевского»	
Ведомственная принадлежность	Министерство науки и высшего образования Российской Федерации	
Почтовый адрес, телефон, адрес электронной почты, адрес официального сайта в сети «Интернет».	603022, Россия, г. Нижний Новгород, пр. Гагарина, д. 23. +7 (831) 462-30-03 unn@unn.ru http://www.unn.ru/	
Сведения о лице, утвердившем отзыв	**ФИО**	Иванченко Михаил Васильевич
	Ученая степень (с указанием шифра специальности)	Д.ф.-м.н., 01.04.03
	Должность	Проректор по научной работе, заведующий кафедрой прикладной математики, профессор кафедры прикладной математики
Сведения о лице, подготовившем отзыв	**ФИО**	Гергель Виктор Павлович
	Ученая степень (с указанием шифра специальности)	Д.т.н., 05.13.12
	Должность	Заведующий кафедрой программной инженерии
Сведения о лице, подготовившем отзыв	**ФИО**	Мееров Иосиф Борисович
	Ученая степень (с указанием шифра специальности)	К.т.н., 05.13.18
	Должность	Заместитель заведующего кафедрой математического обеспечения и суперкомпьютерных технологий

Список основных работ сотрудников ведущей организации по теме диссертации в рецензируемых научных изданиях

1. Gergel V., Kozinov E. A Highly Parallel Approach for Solving Computationally Expensive Multicriteria Optimization Problems // Communications in Computer and Information Science. 2019. Vol. 1129. P. 3-14. DOI: 10.1007/978-3-030-36592-9_1.
2. Povelikin R., Lebedev S., Meyerov I. Multithreaded Multifrontal Sparse Cholesky Factorization Using Threading Building Blocks // Communications in Computer and Information Science. 2019. Vol. 1129. P. 75-86. DOI: 10.1007/978-3-030-36592-9_7.
3. Meyerov I., Bastrakov S., Sysoyev A., Gergel V. Comprehensive Collection of Time-Consuming Problems for Intensive Training on High Performance Computing // Russian Supercomputing Days. 2018. Vol. 965. P. 523–530. DOI: 10.1007/978-3-030-05807-4_44.

4. Pirova A., Meyerov I., Kozinov E., Lebedev S. PMORSy: parallel sparse matrix ordering software for fill-in minimization // Optimization Methods and Software. 2017. Vol. 32, no. 2. P. 274–289. DOI: 10.1080/10556788.2016.1193177.
5. Meyerov I., Bastrakov S., Barkalov K., Sysoyev A., Gergel V. Parallel Numerical Methods Course for Future Scientists and Engineers // Russian Supercomputing Days. 2017. Vol. 793. P. 3–13. DOI: 10.1007/978-3-319-71255-0_1.
6. Meyerov I., Bastrakov S., Surmin I., Bashinov A., Efimenko E., Korzhimanov A., Muraviev A., Gonoskov, A. Hybrid CPU + Xeon Phi implementation of the Particle-in-Cell method for plasma simulation // Supercomputing Frontiers And Innovations. 2016. Vol. 3, no. 3. P. 5–10. DOI: 10.14529/jsfi160301.
7. Gergel V., Kozinov E., Linev A., Shtanyk A. Educational and Research Systems for Evaluating the Efficiency of Parallel Computations // Lecture Notes in Computer Science. 2016. Vol. 10049. P. 278-290. DOI: 10.1007/978-3-319-49956-7_22.
8. Sidnev A. Hardware-Specific Selection the Most Fast-Running Software Components // Lecture Notes in Computer Science. 2016. Vol. 10049. P. 354–364. DOI: 10.1007/978-3-319-49956-7_28.
9. Gergel V., Kustikova V. Internet-Oriented Educational Course “Introduction to Parallel Computing”: A Simple Way to Start // Russian Supercomputing Days. 2016. Vol. 687. P. 291–303. DOI: 10.1007/978-3-319-55669-7_23.
10. Баркалов К.А., Лебедев И.Г., Соврасов В.В., Сысоев А.В. Реализация параллельного алгоритма поиска глобального экстремума функции на Intel Xeon Phi // Вычислительные методы и программирование. 2016. Т. 17, № 1. С. 101–110. DOI: 10.26089/NumMet.v17r110.

Проректор по научной работе

доктор физико-математических наук, доцент М.В. Иванченко