ОТЗЫВ ОФИЦИАЛЬНОГО ОППОНЕНТА

доктора физико-математических наук, профессора Соколова Андрея Владимировича 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.

Актуальность темы исследований.

Сокращение времени выполнения тензорных операций имеет существенное значение для таких научных дисциплин как машинное обучение, спектральные методы и квантовая химия. Большинство методов, используемых для сокращения времени выполнения операций над тензорами, основаны на ручной настройке и автонастройке и, как следствие, неприменимы в процессе оптимизаций, выполняемых промышленными компиляторами, и в других условиях ограниченного времени. Оптимизации, использующие модель целевой архитектуры процессора, могут применяться в процессе компиляции ограниченной по времени для автоматического получения высокопроизводительных многопоточных реализаций тензорных операций для многоядерных процессоров. Таким образом, диссертационное исследование, выполненное Р.А. Гареевым в области методов сокращения времени выполнения тензорных операций на многоядерных процессорах и их автоматического распараллеливания без ручной настройки и автонастройки, является актуальным.

Цель, задачи и основные результаты диссертации.

Цель исследования Р.А. Гареева состояла в разработке эффективных методов обработки информации, получаемой со сложных систем, путем сокращения времени выполнения многопоточных реализаций тензорных операций на многоядерных процессорах общего назначения без ручной настройки и автонастройки. В рамках исследования для достижения этой цели соискателем были решены следующие задачи:

- Разработана модель целевой архитектуры процессора для сокращения времени выполнения матрично-векторных операций и их обобщений.
- Разработаны новые алгоритмы, уменьшающие время выполнения тензорных операций на многоядерных процессорах общего назначения.
- На основе созданной модели и алгоритмов разработана программная система для автоматической оптимизации времени выполнения тензорных операций и их автоматического распараллеливания при компиляции программ.
- Проведены вычислительные эксперименты, подтверждающие эффективность разработанной программной системы.

Текст диссертации оформлен в соответствии с требованиями Минобрнауки РФ, в изложении используется строгий язык и стиль научных публикаций. Текст автореферата достаточно полно отражает содержание диссертации.

Научная новизна исследований и полученных результатов.

В качестве научной новизны диссертационного исследования можно выделить следующее:

1. Разработано новое расширение модели целевой архитектуры процессора Лоу для сокращения времени выполнения матрично-векторных операций и их обобщений на замкнутые полукольца.

2. Разработаны оригинальные алгоритмы выполнения тензорных операций константной сложности относительно размерности тензоров для сокращения времени время выполнения таких операций. Выведены новые формулы, позволяющие получить значения параметров созданных алгоритмов в зависимости от характеристик многоядерных процессоров общего назначения.

3. Разработана оригинальная программная система, позволяющая автоматически оптимизировать время выполнения тензорных операций и осуществить их автоматическое распараллеливание в процессе компиляции программ для многоядерных процессоров общего назначения.

Основное содержание работы.

Диссертация состоит из введения, четырех глав, заключения и библиографии. В приложении 1 приведены основные обозначения, используемые в диссертации. Общий объем диссертации составляет 179 страниц, включая 37 рисунков, 34 таблицы. Библиография содержит 177 наименований.

Во введении отмечена актуальность темы исследований, сформулирована цель работы и полученные результаты, описана новизна и практическая ценность.

В первой главе выполнен обзор известных подходов сокращения времени выполнения тензорных сверток. Описываются методы сокращения времени выполнения свертки тензоров, использующие ручную настройку, автонастройку и моделирование вычислений на целевой архитектуре процессора. Рассматриваются способы сведения сокращения времени выполнения свертки тензоров к сокращению времени выполнения матричного и матрично-векторных произведений.

Во второй главе представлено расширение модели целевой архитектуры процессора Лоу. Описаны новые алгоритмы для вычисления обобщений матричных и матрично-векторных произведений на замкнутые полукольца. Выведены формулы, позволяющие получить значения параметров представленных алгоритмов в зависимости от характеристик многоядерных процессоров общего назначения. Описан 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. Приводится описание архитектуры программной системы автоматической оптимизации тензорных операций и ее возможные реализации.

Четвертая глава посвящена оценке эффективности программой системы автоматической оптимизации тензорных операций при решении обратной задачи гравиметрии, общей задачи о путях, оптимизации тензорных операций.

В заключении сформулированы основные выводы и основные результаты диссертации, приведены рекомендации по использованию результатов диссертации и перспективы дальнейшей разработки темы.

Обоснованность и достоверность полученных результатов.

Обоснованность и достоверность полученных результатов подтверждается данными экспериментов, выполненных соискателем в соответствии с общепринятыми стандартами.

Соответствие паспорту специальности.

Содержание и результаты диссертации соответствуют п. 8 паспорта специальности 05.13.11 – Математическое и программное обеспечение вычислительных машин, комплексов и компьютерных сетей.

Теоретическая и практическая ценность основных положений диссертации.

Теоретическая значимость исследования Р.А. Гареева состоит в создании расширения модели целевой архитектуры процессора Лоу для сокращения времени выполнения матрично-векторных операций и их обобщений на замкнутые полукольца; в разработке оригинальных алгоритмов выполнения тензорных операций константной сложности относительно размерности тензоров для сокращения времени выполнения таких операций. Выведены новые формулы, позволяющие получить значения параметров алгоритмов выполнения тензорных операций в зависимости от характеристик многоядерных процессоров общего назначения. Практическая значимость диссертации Р.А. Гареева состоит в создании программной системы автоматической оптимизации тензорных операций, позволяющей автоматически оптимизировать время выполнения тензорных операций и осуществить их автоматическое распараллеливание в процессе компиляции программ для многоядерных процессоров общего назначения.

Рекомендации по использованию результатов диссертации.

Результаты, полученные в диссертационной работе, могут быть использованы для реализации численных решений задач механики, математической физики и математической геофизики на многоядерных процессорах общего назначения.

Публикации и апробация результатов.

Основные результаты по теме диссертации в полном объеме изложены в 7 публикациях, в том числе в 2 статьях в журналах, входящих в Перечень рецензируемых научных изданий ВАК, в 2 публикациях в изданиях, проиндексированных в Web of Science, в 2 публикациях в изданиях, проиндексированных в Scopus. По результатам исследования соискателем сделано 5 докладов на международных и всероссийских научных конференциях. Оптимизация матричного произведения, обобщенного на замкнутые полукольца, была внедрена в основной код Polly проекта Low Level Virtual Machine.

Замечания по диссертационной работе.

По диссертации имеются следующие замечания, которые не снижают ее общей значимости и высокого научного уровня.

1. На странице 14 есть такой текст:

«Методология и методы исследования. В диссертационной работе использован математический аппарат тензорного исчисления и теории лингвистических основ информатики.»

Но в работе нет описания того, как теория формальных грамматик используется в диссертации.

2. В описании модели процессора нужно было указать, что модель не описывает скалярные регистры и большую часть скалярных инструкций целевой архитектуры процессора, так как эти характеристики процессора не используются в моделировании вычислений, представленном в работе. Из описания, представленного в диссертации, может показаться, что в работе рассматриваются векторные процессоры, но термин «векторный процессор» не используется и остается неясность в этом вопросе.

3. В работе моделируются процессоры с суперскалярной архитектурой и векторным расширением, к которым относятся и стековые процессоры. Можно ли в них использовать результаты диссертации?

4. В диссертации соискателем разработаны новые алгоритмы выполнения тензорных операций, экспериментально подтверждена

применимость созданных алгоритмов. Дополнительно следовало бы выполнить оценку точности предложенных алгоритмов. Кроме этого, отсутствует оценка вычислительной сложности описанных алгоритмов вычисления обобщенного матричного и матрично-векторного произведений.

5. Соискателем разработано новое расширение модели целевой архитектуры процессора Лоу для сокращения времени выполнения матрично-векторных операций и их обобщений на замкнутые полукольца, являющиеся частными случаями алгебр. Указанное обобщение используется для сокращения времени выполнения решений общей задачи о путях. Несмотря на это, не рассмотрен случай сокращения времени выполнения вычислений в конечных полях, представляющих собой другие важные частные случаи алгебр.

6. Для выполнения экспериментов использовались процессоры фирмы Intel, IBM и TSMC. Дополнительно, следовало бы, использовать процессоры отечественного производства.

7. Текст диссертации не свободен от опечаток, пунктуационных ошибок и стилистических погрешностей. Например, на странице 9 слово «криволинейных» использовано с опечаткой. На странице 41 в слове «кэш-память» отсутствует дефис. На странице 94 использовано неправильное словосочетание «оптимизировать времени». На странице 107 после причастного оборота «выполняемой библиотекой LLVM Core в ПС АОТО» отсутствует запятая. На странице 113 слово «повлияло» должно быть использовано во множественном числе.

Заключение о работе.

Диссертационная работа Гареева Романа Альбертовича «Методы оптимизации выполнения тензорных операций на многоядерных процессорах» представляет собой законченную научно-квалификационную работу, в которой разработаны новые модели, методы и алгоритмы для решения задач сокращения времени выполнения тензорных операций и их автоматического распараллеливания на многоядерных процессорах. Решение данных задач имеет важное значение в области создания программных систем для параллельной обработки данных. Работа в полной мере соответствует требованиям, предъявляемым к диссертациям на соискание ученой степени

кандидата физико-математических наук согласно п. 9 «Положения о присуждении ученых степеней» (утверждено постановлением Правительства Российской Федерации от 24 сентября 2013 г. № 842 «О порядке присуждения ученых степеней» с изменениями постановления Правительства Российской Федерации от 21 апреля 2016 г. № 335 «О внесении изменений в Положение о присуждении ученых степеней»), а соискатель заслуживает присуждения ученой степени кандидата физико-математических наук по специальности 05.13.11 «Математическое и программное обеспечение вычислительных машин, комплексов и компьютерных сетей».

Официальный оппонент:

Я, Соколов Андрей Владимирович, даю согласие на включение моих персональных данных в документы, связанные с работой Диссертационного совета ЮУрГУ 212.298.18 и их дальнейшую обработку.

Соколов Андрей Владимирович

«17» февраля 2021 г.

Доктор физико-математических наук, профессор,
ведущий научный сотрудник Федерального государственного бюджетного учреждения науки Федеральный исследовательский центр «Карельский научный центр Российской академии наук», г. Петрозаводск.
Адрес организации: 185910, Россия, Республика Карелия, г. Петрозаводск, ул. Пушкинская, д. 11
Телефон: +7 (814) 276-63-13
Email: avs@krc.karelia.ru

Подпись А.В. Соколова удостоверяю:

Ученый секретарь КарНЦ РАН,
Кандидат биологических наук, / Фокина Н.Н. /

СВЕДЕНИЯ ОБ ОФИЦИАЛЬНОМ ОППОНЕНТЕ

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

Фамилия, имя, отчество	Соколов Андрей Владимирович
Ученая степень (с указанием номера и шифра специальности)	Доктор физико-математических наук. 05.13.18 - Математическое моделирование, численные методы и комплексы программ. 05.13.17 - Теоретические основы информатики
Ученое звание	Профессор
Организация основного места работы	Федеральное государственное бюджетное учреждение науки Федеральный исследовательский центр «Карельский научный центр Российской академии наук»
Ведомственная принадлежность	
Занимаемая должность	Ведущий научный сотрудник
Почтовый адрес	185910, Россия, Республика Карелия, г. Петрозаводск, ул. Пушкинская, д. 11
Телефон	+7 (814) 276-63-13
Адрес электронной почты	avs@krc.karelia.ru

Список основных публикаций по теме диссертации в рецензируемых научных изданиях

1. Aksenova E.A., Barkovsky E.A., Sokolov A.V., The Models and Methods of Optimal Control of Three Work-Stealing Deques Located in a Shared Memory // Lobachevskii Journal of Mathematics. 2019. Vol. 40, no. 11. P. 1763–1770. DOI: 10.1134/S1995080219110052.

2. Kuchumov R., Sokolov A., Korkhov V. Staccato: shared-memory work-stealing task scheduler with cache-aware memory management // International Journal of Web and Grid Services. 2019. Vol. 15, no. 4. P. 394–407. DOI: 10.1504/IJWGS.2019.103233.

3. Barkovsky E.A., Lazutina A.A., Sokolov A.V. The Optimal Control of Two Work-Stealing Deques, Moving One After Another in a Shared Memory //

Program Systems: Theory and Applications. 2019. Vol. 10, no. 1(40). P. 19–32. DOI: 10.25209/2079-3316-2019-10-1-19-32.
4. Kuchumov R., Sokolov A., Korkhov V. Staccato: Cache-Aware Work-Stealing Task Scheduler for Shared-Memory Systems // Lecture Notes in Computer Science. 2018. Vol. 10963. P. 91–102. DOI: 10.1007/978-3-319-95171-3_8.
5. Aksenova E.A., Sokolov A.V. Modeling of the Memory Management Process for Dynamic Work-Stealing Schedulers // 2017 Ivannikov ISPRAS Open Conference (Moscow, Russia, November 30 – December 1, 2017). Massachusetts, IEEE Xplore Digital Library, 2018. P. 12–15. DOI: 10.1109/ISPRAS.2017.00009.
6. Барковский Е.А., Лазутина А.А., Соколов А.В. Построение и анализ модели процесса работы с двумя деками, двигающимися друг за другом в общей памяти // Программные системы: теория и приложения. 2019. Т. 10, № 1 (40). С. 3-17. DOI: 10.25209/2079-3316-2019-10-1-3-17.
7. Барковский Е.А., Кучумов Р.И., Соколов А.В.Оптимальное управление двумя work-stealing деками в общей памяти при различных стратегиях перехвата работы // Программные системы: теория и приложения. 2017. Т. 10, № 1 32. С. 83-103. DOI: 10.25209/2079-3316-2017-8-1-83-103.
8. Сазонов А.М., Соколов А.В. Математическая модель оптимального управления настраиваемой очередью из двух последовательных циклических FIFO-очередей в общей памяти // Информационно-управляющие системы. 2017. Т. 4. С. 44–50. DOI: 0.15217/issnl684-8853.2017.4.44.
9. Барковский Е.А., Соколов А.В., Модель управления двумя параллельными FIFO-очередями, двигающимися друг за другом в общей памяти // Информационно-управляющие системы. 2016. Т 1. С. 65–73. DOI: 10.15217/issn1684-8853.2016.1.65.
10. Соколов А.В., Сазонов А.М., Морозов Е.В., Некрасова Р.С., Разумчик Р.В. Математические модели и алгоритмы оптимального управления FIFO-очередями в общей памяти // Труды Карельского научного центра РАН. 2016. Т. 8. С. 98–107. DOI: 10.17076/mat396.

_____/ Соколов А.В. /

Подпись А.В. Соколова удостоверяю:

Ученый секретарь КарНЦ РАН,
Кандидат биологических наук, _____/ Фокина Н.Н. /