МИНОБРНАУКИ РОССИИ
Федеральное государственное бюджетное образовательное учреждение
высшего образования
«САРАТОВСКИЙ НАЦИОНАЛЬНЫЙ ИССЛЕДОВАТЕЛЬСКИЙ ГОСУДАРСТВЕННЫЙ УНИВЕРСИТЕТ ИМЕНИ Н. Г. ЧЕРНЫШЕВСКОГО»
Кафедра дискретной математики и информационных технологий

РАЗРАБОТКА ПРОГРАММНОГО РЕШЕНИЯ ДЛЯ ОБРАБОТКИ РЕЗУЛЬТАТОВ ЧИСЛЕННОГО МОДЕЛИРОВАНИЯ

АВТОРЕФЕРАТ МАГИСТЕРСКОЙ РАБОТЫ

студента 2 курса 271 группы
направления 09.04.01 — Информатика и вычислительная техника
факультета КНиИТ
Родникова Кирилла Романовича

Научный руководитель
доцент, к. ф.-м. н. _____ А. Д. Панфёров

Заведующий кафедрой
доцент, к. ф.-м. н. _____ Л. Б. Тяпаев

Саратов 2024

ВВЕДЕНИЕ

Продолжающееся успешное развитие вычислительных возможностей современных компьютеров делает более доступными и точными средства численной имитации различных технических, химических и физических процессов. Всё большую роль играют математические модели, исследование свойств которых аналитическими методами не продуктивно и не даёт однозначных результатов.

При выборе темы в качестве цели представляемой выпускной квалификационной работы была определена разработка инструментов обработки, визуализации и анализа промежуточных результатов работы с моделями физических процессов, воспроизводящих поведение двумерного материала в условиях действия на него лазерных импульсов высокой интенсивности. Работы по данной тематике ведутся в СГУ и такие инструменты необходимы для анализа и обработки получаемых результатов. В этой связи тема работы важна и актуальна.

При разработке высокопроизводительных программных решений для работы с большими объёмами данных приходится явно учитывать то обстоятельство, что решение практических задач в этой области в настоящее время невозможно без использования распараллеливания с эффективным использованием аппаратных ресурсов многоядерных процессоров и мультикомпьютеров. Все современные инструменты для работы с данными по необходимости обеспечивают такие возможности. Для подробного знакомства с современным уровнем профессиональных решений в этой области в работу включён обзор современных решений.

В силу ряда обстоятельств выбор был сделан в пользу мощного универсального инструмента — Wolfram Mathematica. В значительной степени это было обусловлено её активным использованием в проекте, для которого было необходимо разработать новые инструменты.

Список задач, которые предстояло решить для достижения поставленной цели, следующий:

— изучить инструменты, используемые при обработке и анализе больших массивов данных;

— сделать выбор с учётом требований поставленной задачи, совместимости с ранее разработанными модулями и минимизации затрат времени

на освоение;

— разработать средства визуализации полученных в процессе моделирования промежуточных данных;
— разработать программу для вычисления по промежуточным данным временных рядов, описывающих поведение наблюдаемых характеристик модели с использованием как собственно результатов моделирования, так и исходных параметров физического процесса.

Магистерская работа состоит из ввдения, 3 глав: «Задача и выбор инструментов для её решения», «Инструменты Wolfram Mathematica для обработки данных», «Программная реализация поставленной задачи», заключения и списка использованных источников. Общий объём работы — 64 страницы, включая 32 рисунка, список использованных источников содержит 36 наименований.

КРАТКОЕ СОДЕРЖАНИЕ РАБОТЫ

В первой главе «Задача и выбор инструментов для её решения» рассмотрена поставленная задача и выполнен обзор стандартных средств для обработки данных, а также их выбор для дальнейшей программной реализации.

Jupyter Notebook — мощнейший инструмент для работы с данными, позволяющий обрабатывать, анализировать и визуализировать результаты [1]. Преимуществами данного программного средва являются: отличная поддержка математических библиотек Python, есть предопределённые модели визуализации, автоматическое создание контрольных точек. К недостаткам можно отнести сложность работы с несколькими ядрами [2].

Вычислительная платформа Apache Spark — фреймворк с открытым исходным кодом для реализации распределённой обработки неструктурированных и слабоструктурированных данных, входящий в экосистему проектов Hadoop [3]. Главным преимуществом Apache Spark является унификация использования на различных кластерных платформах и мощный функционал по работе с большими данными. Также можно выделить надёжность и отказоустойчивость, поддержку нескольких языков. Не смотря на преимущества, Apache Spark не является универсальным инструментом для работы с данными, потому что для каждой конкретной задачи необходимо его конфигурирование, а также обладает неудобной для поставленной задачи визуализацией данных [4].

IBM SPSS Statistics — это универсальная среда количественного анализа данных, работающая в среде Microsoft Windows, а также в операционных системах Linux и macOS [5]. Преимуществами данной программы является: автоматизированная подготовка данных, точное моделирование линейных и нелинейных взаимосвязей, обнаружение аномалий и прогнозирование, поддержка алгоритмов и графиков R. Однако имеется существенный недостаток — большинство функций доступны только в платной версии.

RapidMiner является средой для проведения экспериментов, а также решения задач интеллектуального анализа данных и машинного обучения, в том числе загрузки и преобразования данных (ETL), визуализации и моделирования [6]. К преимуществам можно отнести богатый набор алгоритмов машинного обучения, понятный и удобный дизайн, полная автоматизация

необходимых функций и возможность подключения внешних полезных инструментов [7]. Недостатками является ограниченная поддержка языков программирования, в бесплатной версии ограничен функционал и возможности визуализации.

Apache Hadoop — распределяет большие наборы данных и аналитические задания по узлам вычислительного кластера, преобразуя их в более мелкие рабочие нагрузки, которые могут выполняться параллельно [8]. Преимуществами являются: высокая масштабируемость, отказоустойчивость, возможность использования в облачной среде или на обычном оборудовании, а также хранение данных в любом формате. Однако, в силу своей архитектуры, Hadoop может иметь высокую задержку при обработке данных. Кроме того, возможности Hadoop в обработке данных в режиме реального времени ограничены и для задач, требующих мгновенной реакции, может потребоваться интеграция с другими технологиями.

Wolfram Mathematica — это мощная система анализа данных, обеспечивает обширные возможности для анализа структурированных и неструктурированных данных, создания графиков и визуализаций, а также решения сложных математических задач. Mathematica имеет в наличии более 6000 встроенных функций, покрывающих все области технических расчётов, которые тщательно интегрированны для идеальной совместной работы [9]. Wolfram Mathematica построена с целью предоставления возможностей промышленной мощности, с крепкими эффективными алгоритмами во всех областях, способными решать крупномасштабные задачи с параллелизмом, вычислениями на графических процессорах и многим другим [10].

Таким образом, Wolfram Mathematica, которая обладает широким набором функций, являеется наиболее оптимальным и удобным инструментом для решения поставленной задачи.

Вторая глава «Инструменты Wolfram Mathematica для обработки данных» посвящена более детальному исследованию выбранной системы и языку программирования Wolfram Language, изучению инструментов визуализации Wolfram Mathematica, а также выбору функционала для программной реализации.

Mathematica представляет собой модульную систему программного обеспечения, которая состоит из следующих основных компонентов:

— MathKernel — ядро системы, обеспечивающее вычисления;
— FrontEnd — интерфейс, отвечающий за диалог с пользователем;
— процедур обмена данными MathLink, JLink, NETLink и т.д.
— пакетов расширений Addons packages, Dictionaries, Graphics [11].

Интерфейс пакета строится из нескольких базовых понятий: Блокнот (Notebooks), Ячейка (Cell) и Палитра (Palletes). Блокнотом называется файл, с которым работает пользователь. В нем создаются и вычисляются формулы, строятся графики и таблицы [12]. Блокнот состоит из ячеек, вся информация, которая есть в блокноте, храниться в его ячейках. Все ячейки можно разделить на три типа:

— Ячейки ввода — в них задаются команды (формулы), которые будут вычислены;
— Ячейки результата — в них Mathematica выводит результат вычислений;
— Другие ячейки — ячейки с текстом, заголовки, не требующие вычислений.

Wolfram Language — это символьный язык, специально разработанный с учётом широты и последовательности понятий, необходимых для быстрой разработки мощных программ [13].

В Wolfram Mathematica существует множество расширений, которые позволяют визуализировать различные типы данных и решать различные задачи. Различают следующие основные виды визуализации данных в Mathematica:

1. Plot — базовый метод Plot, который используется для построения графиков функций.
2. ListPlot — метод, который используется для построения графиков списков данных.
3. ScatterPlot — метод, который используется для построения графиков на основе точек.
4. Histogram — метод, который используется для построения гистограмм.
5. PieChart — метод, который используется для построения круговых диаграмм.
6. VectorPlot — метод, который используется для построения графиков векторных функций.
7. ParametricPlot — метод, который используется для построения графи-

ков параметрических функций [14].

8. ContourPlot — метод, который используется для построения контурных графиков.
9. 3DPlot — метод, который используется для построения трёхмерных графиков и многие другие.

Помимо этого Wolfram Mathematica обладает богатым функционалом настроек внешнего вида функций визуализаций. Например, можно настроить размер графика, стиль и свойства линий, отображение осей и легенд.

В Wolfram Mathematica имеется встроенная поддержка параллельных вычислений, которая позволяет эффективно использовать вычислительные ресурсы многоядерных процессоров для ускорения выполнения вычислительных задач [15]. Это позволяет ускорить обработку больших объёмов данных, выполнение сложных вычислений и алгоритмов, которые могут быть разделены на независимые части для параллельного выполнения. Для работы с параллельными вычислениями в Mathematica, используются специализированные функции и конструкции языка, которые позволяют распределить вычислительную нагрузку на несколько ядер процессора.

В третьей главе «Программная реализация поставленной задачи» были визуализированы промежуточные данные систем моделирования, описана структура входных файлов данных, приведён алгоритм вычисления временных рядов наблюдаемых параметров, а также его программная реализация средствами Wolfram Mathematica. Для уменьшения времени выполнения задачи, было принято решение реализовать численное интегрирование с использованием методов параараллельных вычислений.

Под наблюдаемыми параметрами понимаются:

1. поверхностная плотность заряженных носителей (скаляр);
2. поверхностная плотность тока проводимости (двухкомпонентный вектор);
3. поверхностная плотность поляризационного тока (двухкомпонентный вектор).

Эти параметры получаются путём интегрирования по импульсному пространству (узлам сетки) для каждого момента времени t_i.

Описать алгоритм вычисления временных рядов наблюдаемых параметров можно следующим образом:

1. Считывание входных параметров импульса внешнего поля из файла task.txt.
2. Принятие результатов численного воспроизведения эволюции квантовых состояний двухуровневой системы, которые зафиксированы в файле results.txt.
3. Сортировка данных из файла results.txt по четвёртой позиции в строке (момент времени) по возрастанию.
4. Определение количества наборов координат p_1, p_2 для первого момента времени.
5. Определение нормировочного коэффициента для тока проводимости как коэффициента перехода из графеновой системы единиц (элементарный заряд в единицу времени через отрезок единичной длины, что есть 26470 А/см), умноженный на $2\gamma^2$, что равно 8/3 и делённый на $(2\pi)^2$. Итоговое значение 1788. Для поляризационного тока коэффициент в два раза меньше 894. Выполнение численного интегрирования по первым двум координатам для каждого уникального значения третьей координаты.
6. Запись результатов в массив, после чего данный массив сортируется по первой позиции (метка времени) по возрастанию.
7. Сохранение полученного массива с временными рядами в текстовый файл observedvalues.txt.
8. Визуализация временных рядов значений, наблюдаемых параметров в виде графиков.

Организация данных в файле observedvalues.txt построчная, каждая строка включает шесть значений:

1. Метка времени.
2. Значение плотности.
3. Первая компонента тока проводимости.
4. Вторая компонента тока проводимости.
5. Первая компонента поляризационного тока.
6. Вторая компонента поляризационного тока.

Полученные результаты вычисления наблюдаемых параметров визуализированы в виде графиков временных рядов для плотности, для каждой компоненты тока проводимости, для каждой компоненты поляризационного

тока. По оси абсцисс задан временной ряд, а по оси ординат — наблюдаемые параметры.

Структура данных, используемая при разработке программного решения, предполагает работу с текстовыми файлами, содержащими достаточно большое количество строк. Например, текстовый файл results.txt имеет размер 3,93 Гб и состоит из 38125716 записей. Чтобы уменьшить время выполнения задач, численное интегрирование было реализовано с помощью параллельных вычислений. В Wolfram Mathematica множество различных инструментов для реализации параллельных вычислений, однако, выбор был остановлен на ParallelDo и Parallelize.

Далее программное решение, реализованное с использованием функций ParallelDo и Parallelize, было протестировано различными наборами входных параметров с вычислениями на 4, 2, и 1 ядрах процессора. Результат представлен на рисунке 1.

программа	количество ядер	Тест 1	Тест 2	Тест 3	Тест 4	Тест 5
		время выполнения				
ParallelDo	4	11 минут	12 минут	15 минут	28 минут	53 минуты
	2	14 минут	16 минут	20 минут	38 минут	1 час 11 минут
	1	22 минуты	24 минуты	32 минуты	58 минут	1 час 46 минут
Parallelize	4	10 минут	11 минут	15 минут	29 минут	53 минуты
	2	14 минут	15 минут	21 минута	38 минут	1 час 9 минут
	1	22 минуты	24 минуты	32 минуты	1 час 17 минут	1 час 47 минут

Рисунок 1 – Зависимость времени выполнения программного решения от количества ядер

ЗАКЛЮЧЕНИЕ

В ходе выполнения выпускной квалификационной работы поставленные задачи были решены, а цель достигнута.

Основным достижением является разработка программы для вычисления по промежуточным данным временных рядов, описывающих поведение наблюдаемых характеристик модели с использованием как собственно результатов моделирования, так и исходных параметров физического процесса. Программы реализована как в последовательной однопоточной версии, так и с использованием распараллеливания на этапе выполнения выборок данных. Это обеспечило высокий уровень масштабируемости, что подтверждено представленными результатами численных экспериментов.

Работа с параллельной версией программы дала возможность познакомиться с различными инструментами распараллеливания, используемыми в Wolfram Mathematica, получить практические навыки их использования, выявления и решения типичных проблем и узких мест параллельного кода на примерах с реальными данными.

Основные источники информации:

1 Сравнительный анализ систем обработки данных Jupyter Notebook и Anaplan / А.Н. Абдрашитова, Е. Муханов // Научно-образовательный журнал для студентов и преподавателей «StudNet». — 2022. — Т. 5, № 6. — С. 7016–7026.

2 A Tool for Creating and Grading Assignments in the Jupyter Notebook / D. Bourgin, M. Bussonnier, J. Frederic, B. Ragan-Kelley // The Journal of Open Source Education. — 2019. — Vol. 2. — P. 32.

3 Мониторинг кластера анализа больших данных Apache Spark на основе Kubernetes / Д.К. Загребаев // Достижения науки и образования. — 2019. — Т. 46, № 5. — С. 34–42.

4 DIS-groupe [Электронный ресурс] Apache Spark: 6 фактов, которые нужно знать каждому URL: https://dis-group.ru/company-news/articles/6-faktov-ob-apache-spark-kotorye-nuzhno-znat-kazhdomu/ (дата обращения — 31.10.2022) Загл. с экрана. Яз. рус.

5 J. Wiktorowicz, M. Grzelak, K. Grzeszkiewicz-Radulska Analiza statystyczna z IBM SPSS Statistics — Lodz: University of Lodz, 2020, 204 s.

6 Auto Modelling for Machine Learning: A Comparison Implementation between

RapidMiner and Python / N. Baharun, N. Faezah, S. Masrom, N. A. Mohamad Yusri, A.S. Abd Rahman // International Journal of Emerging Technology and Advanced Engineering. — 2022. — Vol. 12. — Pp. 15–27.

7 Soware [Электронный ресурс] Описание системы RapidMiner URL: https://soware.ru/products/rapidminer/ (дата обращения — 07.11.2022) Загл. с экрана. Яз. рус.

8 Apache Hadoop 3 Quick Start Guide / H.V. Karambelkar. — Packt Publishing, 2018.

9 Wolfram Mathematica [Электронный ресурс] Wolfram Mathematica: Современные технические вычисления URL: https://www.wolfram.com/mathematica/ (дата обращения — 21.11.2022) Загл. с экрана. Яз. рус.

10 Основные функции систем компьютерной алгебры / В.Б. Таранчук — Минск: БГУ, 2013. — 59 с.

11 УрФУ [Электронный ресурс] Wolfram Mathematica URL: https://dit.urfu.ru/soft/mathematica/ (дата обращения — 06.12.2022) Загл. с экрана. Яз. рус.

12 Exponenta [Электронный ресурс] Введение в Wolfram Mathematica URL: http://old.exponenta.ru/educat/news/vygovskiy/vygovskiy.asp/ (дата обращения — 29.12.2022) Загл. с экрана. Яз. рус.

13 Оптимизация в системе Mathematica / В.Р. Кристалинский — Санкт-Петербург: Лань, 2023. — 76 с.

14 О решении задач классификации графических образов в системе Wolfram Mathematica / В.Р. Кристалинский // Современные информационные технологии и ИТ-образование. — 2021. — Т. 17, № 2. — С. 464–472.

15 Wolfram [Электронный ресурс] Wolfram Language & System documentation center URL: https://reference.wolfram.com/language/ (дата обращения — 17.02.2023) Загл. с экрана. Яз. англ.