

Министерство науки и высшего образования Российской Федерации
Федеральное государственное бюджетное образовательное учреждение
высшего образования
«Костромской государственный университет»

РАБОЧАЯ ПРОГРАММА ДИСЦИПЛИНЫ

ТЕХНОЛОГИЯ ОБРАБОТКИ БОЛЬШИХ ДАННЫХ

Направление подготовки *09.03.02 Информационные системы и технологии*

Направленность «Разработка и внедрение интеллектуальных компонентов
информационных систем»

Квалификация выпускника: бакалавр

**Кострома
2023**

Рабочая программа дисциплины «Технология обработки больших данных» разработана в соответствии с Федеральным государственным образовательным стандартом по направлению 09.03.02 Информационные системы и технологии (уровень бакалавриата), утвержден приказом Министерства образования и науки РФ № 926 от 19.09.17.

Разработал: Панин И.Г., д.т.н., доцент

Рецензент: Денисов А.Р., д.т.н., доцент

УТВЕРЖДЕНО:

На заседании кафедры информационных систем и технологий:
Протокол заседания кафедры №6 от 27.04.2023 г.
Заведующая кафедрой информационных систем и технологий:
Киприна Л.Ю., к.т.н., доцент

1. Цели и задачи освоения дисциплины

Цель дисциплины: получение слушателями представления о принципах обработки больших массивов данных, языках программирования, используемых для решения этих задач, прикладных пакетах и инструментах создания программ обработки, их тонкой настройки и оптимизации в целях создания высокоэффективного программного обеспечения и методик параллельных вычислений

Задачи дисциплины: знакомство с основными направлениями в области организации обработки больших массивов данных на многопроцессорных вычислительных системах; знакомство с технологиями параллельного программирования; приобретение навыков параллельного программирования при обработке больших массивов данных; знакомство с технологией параллельного программирования на системах с общей и разделенной оперативной памятью; знакомство с технологией параллельных вычислений на графических процессорах; приобретение навыков распараллеливания алгоритмов.

2. Перечень планируемых результатов обучения по дисциплине

В результате освоения дисциплины обучающийся должен:

знать: Теоретические и прикладные основы анализа больших данных; Современный опыт использования анализа больших данных; Содержание этапов жизненного цикла больших данных; Технологии хранения и обработки данных: базы данных, хранилища данных, распределенная и параллельная обработка данных, вычисления в оперативной памяти; Режимы получения и обработки данных, в том числе при потоковой обработке, поддержка режима реального времени; Современная технологическая инфраструктура высокопроизводительных и распределенных вычислений; современные методы и инструментальные средства анализа больших данных

уметь: Проводить интеграцию и преобразование данных, разрабатывать и оценивать модели данных; Использовать инструментальные средства для извлечения, преобразования, хранения и обработки данных из разнородных источников, в том числе в режиме реального времени; Использовать методологическую и технологическую инфраструктуры анализа больших данных для выполнения аналитических работ; Проводить сравнительный анализ методов и инструментальных средств анализа данных

Подготовка к трудовым действиям: Извлечение, проверка, очистка и агрегация данных, разработка представления данных; Разработка, проверка, оценка используемых моделей больших данных

Освоить компетенции: **ПК-4** (Способен осуществлять сбор, обработку и анализ больших данных с использованием существующей в организации методологической и технологической инфраструктуры)

Индикаторы освоенности компетенции:

ПК-4.1 Способен планировать и организовывать аналитические работы

ПК-4.2 Готов осуществлять подготовку данных для проведения аналитических работ

ПК-4.3 Способен проводить аналитические исследования в соответствии с согласованными требованиями

3. Место дисциплины в структуре ОП ВО

Дисциплина относится к части обязательных дисциплин вариативной части учебного плана. Изучается в 6 семестре обучения.

Изучение дисциплины основывается на ранее освоенных дисциплинах/практиках:

- основы программирования с использованием алгоритмических языков Си,
- вычислительные методы,
- основные понятия алгебры, комбинаторики, логики, информатики.

Изучение дисциплины является основой для освоения последующих дисциплин/практик:

- при проведении вычислительных экспериментов в случае выполнения итоговой квалификационной работы, связанной с реализацией высокоэффективных параллельных алгоритмов.

4. Объем дисциплины (модуля)

4.1. Объем дисциплины в зачетных единицах с указанием академических (астрономических) часов и виды учебной работы

Виды учебной работы,	Очная форма
Общая трудоемкость в зачетных единицах	5
Общая трудоемкость в часах	180
Аудиторные занятия в часах, в том числе:	48
Лекции	16
Практические занятия	-
Лабораторные занятия	32
Самостоятельная работа в часах	93,65
Форма промежуточной аттестации	экзамен

4.2. Объем контактной работы на 1 обучающегося

Виды учебных занятий	Очная форма
Лекции	16
Практические занятия	-
Лабораторные занятия	32
Консультации	-
Зачет/зачеты	-
Экзамен/экзамены	0.35
Курсовые работы	-
Курсовые проекты	-
Всего	48.35

5. Содержание дисциплины (модуля), структурированное по темам (разделам), с указанием количества часов и видов занятий

5.1 Тематический план учебной дисциплины

№	Название раздела, темы	Всего з.е. /час	Аудиторные занятия			Самостоятельная работа
			Лекции	Практические	Лаб.	
1	Введение. Цели, задачи обработки больших данных и параллельных вычислений.	7	1	-	-	6
2	Проблемы параллельных вычислений	8	2	-	-	6
3	Этапы подготовки и создания параллельных программ	8	2	-	-	6
4	Технология программирования OpenMP	56	5	-	16	35
5	Интерфейс передачи сообщений MPI	56	5	-	16	35
6	Технология CUDA	5,65	1	-	-	5,65
7	Подготовка к экзамену	36	-	-	-	36
	Итого:	177,65	16	-	32	93,65+36

5.2. Содержание:

1. Введение. Цели, задачи параллельных вычислений. Ограничения максимальной производительности однопроцессорных ЭВМ. Параллельные и распределенные вычисления и их техническая основа – вычислительные кластеры, ГРИД-системы и суперкомпьютеры. История развития параллельных вычислительных систем, виды параллелизма
2. Проблемы параллельных вычислений. Архитектура высокопроизводительных ЭВМ. Конвейерные и векторные вычисления. Многопроцессорные вычислительные системы с общей и распределенной памятью (мультипроцессоры и мультикомпьютеры). Классификация архитектур многопроцессорных вычислительных систем. Основные технологии распараллеливания программ. Закон Амдаля (о существовании последовательных алгоритмов). Закон Мура (о росте производительности последовательных компьютеров). Закон Гроша (о высокой стоимости параллельных систем). Гипотеза Минского (о влиянии потерь на взаимодействие на степень ускорения параллельных вычислений по сравнению с последовательными).
3. Этапы подготовки и создания параллельных программ. Возможные схемы взаимодействия ветвей параллельных алгоритмов и типовые топологии схем коммутации – кольцо, линейка, решетки, полный граф, гиперкуб, тор, дерево.
4. Технология программирования OpenMP. Последовательные и параллельные нити программы. Организация параллельных секций. Параллельные циклы. Директивы и функции. Синхронизация. Классы переменных. Спецификации OpenMP для языков Fortran или C и C++.
5. Интерфейс передачи сообщений MPI. Общие принципы построения и реализации MPI. Разработчики, история создания. Принципы разработки параллельных алгоритмов и программ, наличие общих функций MPI, коммуникаторы. Функции обмена сообщениями

типа «точка-точка»: блокирующий и неблокирующий обмен, синхронные и стандартные послышки сообщений. Предотвращение тупиков. Коллективные функции обмена данными: широковещательная рассылка, функции сбора и рассыпания данных. Функции редукции данных. Создание групп процессов, области связи, коммутаторы. Обмен данными внутри группы, межгрупповой обмен. Топология обменов. Декартовы топологии. Топологии произвольного графа. Примеры параллельных программ на основе обработки массивов.

6. Технология CUDA. GPU как массивно-параллельный процессор. Модель программирования CUDA. Иерархия памяти CUDA. Глобальная память. Параллельные решения задач умножения матриц и решения СЛАУ. Разделяемая память. Реализация примитивов параллельного суммирования (reduce) и префиксной суммы (scan) на CUDA.

5.3. Практическая подготовка

Код компетенции	Индикатор компетенции	Содержание задания на практическую подготовку по выбранному виду деятельности	Число часов практической подготовки			
			Всего	Лекции	Практ. занятия	Лаб. раб
ПК-4	ПК-4.1	Планирование и организация аналитических работ (лаб. работы 4,5)	4	-	-	4
ПК-4	ПК-4.2	Подготовка данных (лаб. работы 4,5)	6	-	-	6
ПК-4	ПК-4.3	Проведение аналитических исследований (лаб. работы 4,5)	8	-	-	8

6. Методические материалы для обучающихся по освоению дисциплины

6.1. Самостоятельная работа обучающихся по дисциплине (модулю)

№ п/п	Раздел (тема) дисциплины	Задание	Часы	Методические рекомендации по выполнению задания	Форма контроля
1.	Введение. Цели, задачи параллельных вычислений.	Изучить материалы лекции	6	Использовать материалы лекций и рекомендованную литературу	Прохождение теста
2.	Проблемы параллельных вычислений	Изучить материалы лекции	6	Использовать материалы лекций и рекомендованную литературу	Прохождение теста
3	Этапы подготовки и создания параллельных программ	Изучить материалы лекции	6	Использовать материалы лекций и рекомендованную литературу	Прохождение теста
4	Технология программирования OpenMP	Изучить материалы лекции.	35	Использовать материалы лекций и рекомендованную литературу.	Прохождение теста

		Выполнение заданий по вариантам		Выполнить задания, используя материалы лекций. Подготовить отчет по л/р	Проверка выполнения заданий
5	Интерфейс передачи сообщений MPI	Изучить материалы лекции. Выполнение заданий по вариантам	35	Использовать материалы лекций и рекомендованную литературу. Выполнить задания, используя материалы лекций. Подготовить отчет по л/р	Прохождение теста Проверка выполнения заданий
6	Технология CUDA	Изучить материалы лекции.	5,65	Использовать материалы лекций и рекомендованную литературу.	Прохождение теста
7	Подготовка к экзамену	Изучение всех лекций и выполнение всех лабораторных работ	36	Использование материалов лекций, лабораторных работ и рекомендованной литературы	Экзамен

6.2. Тематика и задания для практических занятий (при наличии)

6.3. Тематика и задания для лабораторных занятий

Изучение параллельных языков программирования.

OpenMP: Первая параллельная программа. Компиляция и запуск программ. Директивы OpenMP, Переменные окружения. Библиотечные функции. Средства синхронизации. Расширенные возможности. Отладка, трассировка, профилирование параллельных программ, оптимизация.

MPI: Первая параллельная программа. Компиляция и запуск программ. Тестирование коммуникационной среды кластера. Тестирование производительности кластера. Передача данных с помощью блокирующих коммуникационных функций типа “Точка-Точка”. Другие виды передачи данных с помощью коммуникационных функций типа “Точка-Точка”. Одновременная передача данных. Коллективные операции. Глобальные вычислительные операции. Создание производных типов данных. Передача упакованных данных. Работа с группами и коммутаторами, создание виртуальных топологий. Отладка, трассировка, профилирование параллельных программ, оптимизация. Сравнение теоретических оценок эффективности распараллеливания с полученными численными оценками.

Реализация простейших параллельных алгоритмов матричной алгебры. MPI, OpenMP,

6.4. Методические рекомендации для обучающихся по освоению дисциплины (модуля)

Рекомендуется обязательное посещение лекций и лабораторных работ студентами

ввиду ограниченного количества литературы и постоянного обновления теоретического и практического материала.

Самостоятельная работа студентов заключается в изучении материала лекций и рекомендованной литературы, самостоятельном изучении указанных разделов и тем дисциплины, подготовке к лабораторным работам, подготовке отчетов по лабораторным работам, выполнении индивидуальных заданий, подготовке к защите лабораторных работ, подготовке реферата. Отчет по лабораторной работе может представляться в электронной форме в виде листинга программного кода или файла в формате *.doc или *.pdf с включением изображений (скриншотов) в соответствии с заданием на лабораторную работу. Контроль самостоятельной работы студентов осуществляется в форме теоретического и практического опроса согласно перечню тем, предусмотренных в рабочей программе дисциплины.

Лекционное обучение осуществляется в аудиториях, оснащенных специализированным оборудованием, таким как: ПК, видеопроектор, оптический проектор, аудио и видеосистемы.

Лабораторные задания выполняются в соответствии с тематикой лабораторных работ, приведенной в рабочей программе дисциплины, в компьютерных классах, оснащенных 7-9 ПК, объединенными в локальную сеть.

7. Перечень основной и дополнительной литературы, необходимой для освоения дисциплины (модуля)

а) основная:

1. Воеводин В.В. Параллельные вычисления / Воеводин В.В., Воеводин Вл.В. СПб: БХВ-Петербург, 2002. 608с.
2. Немнюгин С.А. Параллельное программирование для многопроцессорных вычислительных систем / С.А. Немнюгин, О.Л. Стесик – СПб.: БХВ-Петербург. 2002. 400с.
3. Гергель В.П. Теория и практика параллельных вычислений. М.: БИНОМ.ЛЗ, 2007.
4. Гергель В.П., Стронгин Р.Г. Основы параллельных вычислений для многопроцессорных вычислительных систем. – Н.Новгород, ННГУ, 2003.
5. Корнеев В.Д. Параллельное программирование в MPI. – Ижевск, 2004.

б) дополнительная:

1. Д.С. Глызин, Д.С. Кащенко. Параллельное и функциональное программирование. Методические указания. Ярослав. гос. ун-т им. П.Г. Демидова. Ярославль: ЯрГУ, 2009.
2. А.С. Антонов. Параллельное программирование с использованием технологии OpenMP: учеб. пособие. М.: Изд-во МГУ, 2009.
3. Д. Кнут, Искусство программирования, том 1. Основные алгоритмы. М.: Издательский дом «Вильямс», 2008.
4. Д. Уоткинс, Основы матричных вычислений. М.: БИНОМ. Лаборатория знаний, 2006.
5. Барский А.Б. Параллельные информационные технологии. - М.: Интернет-университет информационных технологий, 2007.
6. Левин М.П. Параллельное программирование с использованием OpenMP. - М.: БИНОМ.ЛЗ, 2008.
7. Эндрюс Г.Р. Основы многопоточного, параллельного программирования. М.: Издательский дом "Вильямс". 2003. 512с.
8. Афанасьев К.Е. Многопроцессорные системы: построение, развитие, обучение /

8. Перечень ресурсов информационно-телекоммуникационной сети «Интернет», необходимых для освоения дисциплины

1. Богданов А., Мареев В., Станнова Е., Корхов В. Архитектуры и топологии многопроцессорных вычислительных систем // электронный учебник <http://www.informika.ru/text/teach/topolog/index.htm>
2. Букатов А.А., Дацюк В.Н., Жегуло А.И. Программирование многопроцессорных вычислительных систем. Ростов-на-Дону. Издательство ООО «ЦВВР», 2003, 208 с. (<http://rsusu1.rnd.runnet.ru/tutor/method/index.html>)
3. Быкова В. В. Теоретические основы анализа параметризованных алгоритмов [Электронный ресурс] : Монография / В. В. Быкова. - Красноярск: Сиб. федер. ун-т, 2011. 180 с. URL: <http://www.znaniium.com/bookread.php?book=441165>
4. Сырецкий Г. А. Информатика. Фундаментальный курс. Том II. Информационные технологии и системы /Г. А. Сырецкий. СПб.: БХВ-Петербург, 2007. 846 с. <http://znaniium.com/bookread.php?book=350042>
5. Богачёв К.Ю. Основы параллельного программирования: учебное пособие. - 2-е (эл.). - М.: Бином. Лаборатория знаний, 2013. - 342 с. http://e.lanbook.com/books/element.php?pl1_id=42626
6. Информационно-аналитические материалы по параллельным вычислениям Научно-исследовательский вычислительный центр МГУ (<http://www.parallel.ru>)
7. Гергель В.П. Теория и практика параллельных вычислений. (<http://www.intuit.ru/departament/calculate/paraltp/index.html>)
8. Немнюгин С.А. Средства программирования для многопроцессорных вычислительных систем. – СПб: СПбГУ, 2007 (http://www.phys.spb.ru/content/File/Library/studentlectures/Nemnygin/Metod_Nemnygin_Intel.pdf)
9. Foster I. Designed and Building Parallel Programs. – Addison Wesley, 1994. (<http://www.mcs.anl.gov/dvpp.html>)
10. Introduction to parallel Computing (Teaching Course). (<http://www.ece.northwestern.edu/course/358.html>)

9. Описание материально-технической базы, необходимой для осуществления образовательного процесса по дисциплине

Наименование специальных помещений и помещений для самостоятельной работы	Оснащенность специальных помещений и помещений для самостоятельной работы	Перечень лицензионного программного обеспечения. Реквизиты подтверждающего документа
ауд. Е-326 (занятия лекционного типа, групповые консультации, промежуточная аттестация)	Лекционная аудитория. Число посадочных мест – 80. Имеется: мультимедиа – проектор с компьютером, выход в интернет; усилитель; колонки.	Лицензионное программное обеспечение не используется
ауд. Е-327 (лабораторные занятия, индивидуальные	Компьютерный класс. Число посадочных мест – 16. Число	Лицензионное программное обеспечение не используется

консультации, промежуточная аттестация, самостоятельная работа обучающихся)	мест, оборудованных компьютерами – 8 с выходом в интернет. Имеется: мультимедиа – проектор с компьютером; интерактивная доска.	
ауд. Е-324 (лабораторные занятия, индивидуальные консультации, промежуточная аттестация, самостоятельная работа обучающихся)	Компьютерный класс. Число посадочных мест – 16. Число мест, оборудованных компьютерами – 8 с выходом в интернет. Имеется: мультимедиа – проектор с компьютером; колонки.	Лицензионное программное обеспечение не используется

Проведение занятий лекционного типа, лабораторных работ, индивидуальных и групповых консультаций, промежуточной аттестации возможно в других аудиториях КГУ, имеющих аналогичное техническое и программное оснащение.