

Министерство науки и высшего образования Российской Федерации
НАЦИОНАЛЬНЫЙ ИССЛЕДОВАТЕЛЬСКИЙ
ТОМСКИЙ ГОСУДАРСТВЕННЫЙ УНИВЕРСИТЕТ (НИ ТГУ)

Институт прикладной математики и компьютерных наук

УТВЕРЖДАЮ:

Директор

 А. В. Замятин

«16» _____ 2023 г.

Рабочая программа дисциплины

Технологии высокопроизводительной обработки больших данных

по направлению подготовки

01.04.02 Прикладная математика и информатика

Направленность (профиль) подготовки:

Обработка данных, управление и исследование сложных систем

Форма обучения

Очная

Квалификация

Магистр

Год приема

2023

Код дисциплины в учебном плане: Б1.В.ДВ.01.01.01

СОГЛАСОВАНО:

Руководитель ОП

 Л.А. Нежелская

Председатель УМК

 С.П. Сущенко

Томск – 2023

1. Цель и планируемые результаты освоения дисциплины

Целью освоения дисциплины является формирование следующих компетенций:

- УК-1 – Способен осуществлять критический анализ проблемных ситуаций на основе системного подхода, выработать стратегию действий.
- ПК-5 – Способен управлять получением, хранением, передачей, обработкой больших данных.

Результатами освоения дисциплины являются следующие индикаторы достижения компетенций:

ИУК-1.1 Выявляет проблемную ситуацию, на основе системного подхода осуществляет её многофакторный анализ и диагностику.

ИУК-1.2 Осуществляет поиск, отбор и систематизацию информации для определения альтернативных вариантов стратегических решений в проблемной ситуации.

ИУК-1.3 Предлагает и обосновывает стратегию действий с учетом ограничений, рисков и возможных последствий.

ИПК-5.1 Осуществляет мониторинг и оценку производительности обработки больших данных.

ИПК-5.2 Использует методы и инструменты получения, хранения, передачи, обработки больших данных.

ИПК-5.3 Разрабатывает предложения по повышению производительности обработки больших данных.

2. Задачи освоения дисциплины

- Обучить студентов различным методам организации высокопроизводительных вычислений,
- Сформировать навыки использования различных программных инструментов для организации параллельных вычислений.

3. Место дисциплины в структуре образовательной программы

Дисциплина относится к части образовательной программы, формируемой участниками образовательных отношений, предлагается обучающимся на выбор. Дисциплина входит в модуль «Общепрофессиональные модули по выбору» – «Прикладной модуль».

4. Семестр(ы) освоения и форма(ы) промежуточной аттестации по дисциплине

Первый семестр, экзамен

5. Входные требования для освоения дисциплины

Для успешного освоения дисциплины требуются компетенции, сформированные в ходе освоения образовательных программ предшествующего уровня образования.

6. Язык реализации

Русский

7. Объем дисциплины

Общая трудоемкость дисциплины составляет 3 з.е., 108 часов, из которых:

- лекции: 16 ч.;
- семинарские занятия: 0 ч.
- практические занятия: 0 ч.;
- лабораторные работы: 32 ч.

в том числе практическая подготовка: 0 ч.

Объем самостоятельной работы студента определен учебным планом.

8. Содержание дисциплины, структурированное по темам

Тема 1. Актуальность, базовая терминология и тенденции развития. Параллелизм компьютерных вычислений.

Понятие высокопроизводительных вычислений. Способы повышения производительности вычислительной системы. Понятие параллельных вычислений.

Причины, порождающие вычислительный параллелизм. Мультипроцессирование. Классификация уровней параллелизма, предложенная П. Треливеном. Векторная обработка данных. Многофункциональная обработка данных. Конвейер команд.

Тема 2. Архитектура высокопроизводительных вычислительных систем.

Классификация вычислительных систем.

Классификация вычислительных систем Флинна. Системы с одним потоком команд и одним потоком данных. Системы с одним потоком команд и множеством потоков данных. Системы с множеством потоков команд и одним потоком данных. Системы с множеством потоков команд и множеством потоков данных.

Классификация вычислительных систем с множеством потоков команд и множеством потоков данных. Системы с разделяемой памятью. Системы с распределённой памятью. Вычислительные системы с централизованной общей памятью. Вычислительные системы с распределённой общей памятью. Симметричная многопроцессорная архитектура. Архитектура CC-NUMA. Архитектура COMA. Архитектура NCC-NUMA. Вычислители с массовым параллелизмом. Кластеры, их типы.

Тема 3. Облачные технологии, их свойства и типы.

Понятие облачных технологий и вычислений. Свойства облачных технологий. Классификация облачных сервисов по типу ресурса, к которому предоставляется доступ. Классификация облачных сервисов по модели развёртывания.

Тема 4. Технология вычислений MapReduce.

Технология обработки данных MapReduce, её сферы применения. Реализации MapReduce. Пример программы, реализованной с помощью технологии MapReduce.

Тема 5. Распределённые файловые системы.

Понятие распределённых файловых систем. Свойства распределённых файловых систем. Файловые системы Google File System, Hadoop distributed file system, их архитектура, сходства и различия.

Тема 6. Программирование для высокопроизводительных вычислений. Методология проектирования параллельных алгоритмов.

Проблемы параллельного программирования. Методология организации параллельных вычислений для SIMD архитектуры. Методология организации параллельных вычислений для MIMD архитектуры. Показатели качества параллельных методов. Библиотеки для обмена сообщениями. MPI (Message Passing Interface). OpenMP (Open Multi-Processing). Пример программы с использованием OpenMP.

9. Текущий контроль по дисциплине

Текущий контроль по дисциплине проводится в форме вопросов по лекционному материалу, лабораторных работ и фиксируется в форме контрольной точки не менее одного раза в семестр.

Вопросы для промежуточной аттестации:

Тема 1. Актуальность, базовая терминология и тенденции развития. Параллелизм компьютерных вычислений

1. Дать определение термину «высокопроизводительные вычисления».
2. Назвать основные способы повышения производительности вычислительной системы.
3. С чем связаны сложности дальнейшего повышения тактовой частоты процессора?
4. Дать определение термину «мультипроцессирование».
5. Рассказать о векторной обработке данных.
6. Применяется ли векторная обработка данных в современных персональных компьютерах? Назвать соответствующие технологии / расширения процессора.
7. Рассказать о многофункциональной обработке данных.
8. Рассказать о конвейере команд.
9. В чём разница пиковой и реальной производительности суперкомпьютера?
10. Какой суперкомпьютер на данный момент имеет максимальную пиковую производительность?

Тема 2. Архитектура высокопроизводительных вычислительных систем.

Классификация вычислительных систем.

1. По какому критерию классифицированы вычислительные системы в классификации Флинна?
2. Чем отличаются от прочих SISD системы?
3. Чем отличаются от прочих SIMD системы?
4. Чем отличаются от прочих MISD системы?
5. Чем отличаются от прочих MIMD системы?
6. Чем отличаются мультипроцессоры от мультикомпьютеров?
7. Чем отличаются UMA системы от NUMA систем?
8. Чем отличаются от прочих CC-NUMA системы?
9. Чем отличаются от прочих NCC-NUMA системы?
10. Чем отличаются от прочих SMP системы?
11. Чем отличаются от прочих мультипроцессорных вычислительных систем кластеры?

Тема 3. Облачные технологии, их свойства и типы.

1. Что такое «облачные технологии»?
2. Назовите основные свойства облачных технологий.
3. В чём причина широкого распространения облачных технологий?
4. Назовите основные типы облачных сервисов при проведении классификации по типу предоставляемого ресурса.
5. Приведите примеры облачных сервисов.
6. Почему IaaS (Infrastructure as a Service, Аппаратное обеспечение как сервис) не относится к сфере облачных технологий?
7. В чём отличие DaaS от SaaS?
8. Какие типы облачных сервисов предназначены в первую очередь для конечных пользователей, а какие для разработчиков ПО?
9. Какие модели развёртывания облачных систем вы знаете?
10. С какой целью создаются гибридные облака?

Тема 4. Облачные технологии, их свойства и типы.

1. Дайте определение «функции высшего порядка» в программировании.
2. Какие функции высшего порядка используются в технологии облачных вычислений MapReduce?
3. Назовите шаги обобщённого алгоритма MapReduce.

4. Какие типы узлов создаются при использовании технологии MapReduce и каково их назначение?
5. Какими характеристиками должна обладать вычислительная задача, чтобы она могла быть эффективно решена с помощью технологии MapReduce?
6. Приведите пример задачи, которая может быть эффективно решена с помощью технологии MapReduce?
7. Запишите сигнатуру функции map() на псевдоязыке программирования для приведённого примера.
8. Запишите сигнатуру функции reduce() на псевдоязыке программирования для приведённого примера.
9. Назовите наиболее известные реализации MapReduce.
10. Для каких типов вычислительных систем в классификации Флинна имеет смысл использовать технологию MapReduce?

Тема 5. Распределённые файловые системы.

1. Дайте определение «распределённой файловой системы».
2. Дайте определение «репликации данных».
3. С какой целью выполняется репликация данных?
4. Каковы сферы применения распределённых файловых систем?
5. В чём состоит отличие распределённой файловой системы от распределённого хранилища данных?
6. Какие распределённые файловые системы вы знаете?
7. Опишите принцип работы Google File System.
8. Что происходит в Google File System, если узел данных, хранящий основную копию фрагмента файла, выходит из строя?
9. Опишите принцип работы Hadoop distributed file system.
10. Что происходит в Hadoop distributed file system, если узел данных, хранящий основную копию фрагмента файла, выходит из строя?
11. Назовите сходства и различия Google File System и Hadoop distributed file system.

Тема 6. Программирование для высокопроизводительных вычислений. Методология проектирования параллельных алгоритмов.

1. Назовите основные проблемы параллельного программирования.
2. Расскажите о методологии организации параллельных вычислений для SIMD архитектуры.
3. Расскажите о методологии организации параллельных вычислений для MIMD архитектуры.
4. Назовите основные показатели качества параллельных методов.
5. Как связаны между собой такие показатели как ускорение и эффективность.
6. Может ли значение ускорения превышать количество процессоров, используемых для запуска параллельной программы?
7. Чем отличается сильная масштабируемость от слабой?
8. Для чего используется библиотека MPI, какие основные возможности предоставляет?
9. Что такое OpenMP, какие возможности предоставляет?
10. Имеет ли смысл использовать MPI на вычислителях с общей памятью?

10. Порядок проведения и критерии оценивания промежуточной аттестации

Экзамен по окончании изучения курса проводится в письменно-устной форме. Допуском до экзамена является выполнение двух лабораторных работ. При выполнении менее трёх лабораторных работ максимальная оценка на экзамене – 4.

Примерный перечень теоретических вопросов:

1. Виды параллелизма компьютерных вычислений.
2. Классификацию вычислительных систем Флинна.
3. Классификация MIMD систем.
4. Свойства и типы облачных технологий.
5. Технология распределённых вычислений MapReduce.
6. Реализации MapReduce: сходства и различия.
7. Распределённые файловые системы.
8. Распределённая файловая система Google File System.
9. Распределённая файловая система Hadoop distributed file system.
10. Методология организации параллельных вычислений для SIMD, MIMD архитектур.
11. Показатели качества параллельных методов.
12. OpenMP и MPI: назначение каждой из технологий и их сравнительный анализ.

Оценка «отлично» ставится, если студент полноценно ответил на вопрос билета, а также на дополнительный вопрос, требующий аналитического сопоставления знаний, полученных при изучении различных тем данной дисциплины. Оценка «хорошо» ставится, если студент полноценно ответил на вопрос билета, но не ответил на дополнительный вопрос. Оценка «удовлетворительно» ставится, если студент частично ответил на вопрос билета.

11. Учебно-методическое обеспечение

- а) Электронный учебный курс по дисциплине в электронном университете «Moodle».
- б) Оценочные материалы текущего контроля и промежуточной аттестации по дисциплине.
- в) Учебное пособие
- г) Видеолекции
- д) Методические указания по проведению лабораторных работ.

12. Перечень учебной литературы и ресурсов сети Интернет

- а) основная литература:
 - S. Srinivasan. Cloud Computing Basics electronic resource / New York, NY : Springer New York : Imprint: Springer, 2014. – 400 с.
 - A Ohri. R for Cloud Computing electronic resource : An Approach for Data Scientists. – New York, NY : Springer New York : Imprint: Springer, 2014. – 336 с.
 - Zaigham Mahmood. Cloud Computing: Methods and Practical Approaches, – London : Springer London : Imprint: Springer, 2013. – 259 с.
 - К. Ю. Богачев. Основы параллельного программирования / Москва БИНОМ. Лаборатория знаний, 2013. – 389 с.
 - В. П. Гергель. Теория и практика параллельных вычислений : учебное пособие / Москва : Интернет-Университет Информационных Технологий, 2016. – 426 с.
- б) дополнительная литература:
 - Christoph Fehling, Frank Leymann, Ralph Retter, Walter Schupeck, Peter Arbitter. Cloud Computing Patterns electronic resource : Fundamentals to Design, Build, and Manage Cloud Applications - Vienna : Springer Vienna : Imprint: Springer, 2014. - 280 с.
 - Xiaolin Li, Judy Qiu. Cloud Computing for Data-Intensive Applications electronic resource – New York, NY : Springer New York : Imprint: Springer, 2014. – 226 с.

- Zaigham Mahmood. Cloud Computing electronic resource : Challenges, Limitations and R&D Solutions – New York, NY : Springer New York : Imprint: Springer, 2014. – 321 с.
- А. В. Линев, Д. К. Боголепов, С. И. Бастраков. Технологии параллельного программирования для процессоров новых архитектур – Нижегородский гос. ун-т им. Н. И. Лобачевского – Москва : Изд-во Московского университет, 2010. – 572 с.

в) ресурсы сети Интернет:

- Облачный сервис: <http://www.ncloudtech.ru>
- Статья «Бизнес в облаках. Чем полезны облачные технологии для предпринимателя» <https://kontur.ru/articles/225>
- Батура Т.В., Мурзин Ф.А., Семич Д.Ф. Облачные технологии: основные понятия, задачи и тенденции развития. <http://swsys-web.ru/cloud-computing-basic-concepts-problems.html>
- MapReduce Tutorial: https://hadoop.apache.org/docs/r1.2.1/mapred_tutorial.html
- HDFS Architecture Guide: https://hadoop.apache.org/docs/r1.2.1/hdfs_design.html

13. Перечень информационных технологий

а) лицензионное и свободно распространяемое программное обеспечение:

- Microsoft Visual Studio Community Edition 2013 (C++), библиотека классов, реализующая технологию MapReduce (например, <https://github.com/cdmh/mapreduce>). Все используемые программные продукты являются свободно распространяемыми.

б) информационные справочные системы:

- Электронный каталог Научной библиотеки ТГУ – <http://chamo.lib.tsu.ru/search/query?locale=ru&theme=system>
- Электронная библиотека (репозиторий) ТГУ – <http://vital.lib.tsu.ru/vital/access/manager/Index>
- ЭБС Лань – <http://e.lanbook.com/>
- ЭБС Консультант студента – <http://www.studentlibrary.ru/>
- Образовательная платформа Юрайт – <https://urait.ru/>
- ЭБС ZNANIUM.com – <https://znanium.com/>
- ЭБС IPRbooks – <http://www.iprbookshop.ru/>

14. Материально-техническое обеспечение

Аудитории для проведения занятий лекционного типа.

Аудитории для проведения занятий семинарского типа, индивидуальных и групповых консультаций, текущего контроля и промежуточной аттестации.

Помещения для самостоятельной работы, оснащенные компьютерной техникой и доступом к сети Интернет, в электронную информационно-образовательную среду и к информационным справочным системам.

Для проведения лабораторных работ требуются компьютеры, оснащённые дискретными видеокартами фирмы Nvidia с поддержкой технологии CUDA.

15. Информация о разработчиках

Дружинин Денис Вячеславович, канд. техн.наук, доцент кафедры теоретических основ информатики.