Программа учебной дисциплины «Высокопроизводительные вычисления»

Утверждена
Академическим советом ОП
Протокол №_____ от 08.09.2019

Разработчик	Посыпкин Михаил Анатольевич, Профессор, Базовая кафедра "Интеллектуальные технологии системного анализа и управления" ФИЦ РАН
Число кредитов	2
Контактная работа (час.)	60
Самостоятельная работа (час.)	0
Курс, Образовательная программа	3 (Б) курс, Прикладная математика и информатика
Формат изучения дисциплины	Без использования онлайн курса

1. Цель, результаты освоения дисциплины и пререквизиты

Цели:

1. Получение студентами знаний в области параллельных и распределенных вычислений, выработка у студентов навыков разработки, отладки и исследования производительности параллельных программ.

Планируемые результаты обучения (ПРО):

1. Знать основные типы высокопроизводительных архитектур
2. Знать определения и уметь вычислять базовые характеристики производительности параллельных алгоритмов
3. Владеть базовыми средствами многопоточного программирования
4. Знать архитектуру, принципы разработки программ и инструменты для программирования графических ускорителей
5. Владеть высокоуровневыми инструментами многопоточного программирования
6. Владеть средствами разработки программ для систем с распределенной памятью

Пререквизиты:

1. Базовые знания C/C++, навыки работы в Linux.

2. Содержание учебной дисциплины

Тема (раздел дисциплины)	**Объем в часах** лк см onl/cp	**Планируемые результаты обучения (ПРО), подлежащие контролю**	**Формы контроля**
Теоретические основы высокопроизводительных вычислений	4 0 0	• Знать основные типы высокопроизводительных архитектур • Знать определения и уметь вычислять базовые характеристики	ЭКЗ.

		производительности параллельных алгоритмов	
Введение в архитектуру микропроцессоров	4 0 0	• Знать основные типы высокопроизводительных архитектур	ЭКЗ.
Базовые средства многопоточного программирования	6 6 0	• Владеть базовыми средствами многопоточного программирования	ЛАБ1, ЭКЗ.
Программирование графических ускорителей	6 6 0	• Знать архитектуру, принципы разработки программ и инструменты для программирования графических ускорителей	ЛАБ2, ЭКЗ.
Высокоуровневые средства многопоточного программирования	8 8 0	• Владеть высокоуровневыми инструментами многопоточного программирования	ЛАБ3, ЭКЗ.
Средства разработки программ для систем с распределенной памятью	6 6 0	• Владеть средствами разработки программ для систем с распределенной памятью	ЭКЗ.
Часов по видам учебных занятий:	34 26 0		
Итого часов:	60		

Содержание разделов дисциплины:

1. **Теоретические основы высокопроизводительных вычислений**
   Рассматривается мотивация, история возникновения, такие базовые концепции высокопроизводительных вычислений, как ускорение, эффективность, масштабируемость. Существенное внимание будет уделено причинам снижения эффективности приложений: зависимостям по данным между инструкциями и накладным расходам на взаимодействие компонентов параллельной программы. Планируется рассмотреть модели высокопроизводительных вычислений, позволяющие теоретически оценивать время работы и ускорение приложений.
2. **Введение в архитектуру микропроцессоров**
   В теме рассматривается устройство современных многоядерных процессоров. Изучается общая схема микропроцессора и вычислительного ядра. На примере различных процессоров демонстрируются основные компоненты современных процессоров: функциональные устройства, регистры, буферы для работы с памятью, различные виды кэшей. Рассматриваются фазы обработки команд (ступени) в конвейерах, зависимости по данным и их устранение на аппаратном уровне. Внимание также уделяется таким механизмам, как предсказание ветвлений, переупорядочивание команд. Изучаются векторные устройства, позволяющие обрабатывать одновременно несколько однотипных данных одной командой. Детально рассматривается организация памяти, виды и принципы работы кэш-памяти, коггерентность кэшей в многоядерных системах. Рассматриваются основные компоненты современных графических ускорителей: потоковые процессоры, особенности режима SIMT, иерархия памяти (глобальная, константная, локальная, регистровая, кэш-память). Затрагиваются современные «тензорные» команды, нацеленные на задачи машинного обучения. Наиболее мощные современные суперкомпьютеры имеют кластерную архитектуру. Кластер представляет собой совокупность вычислительных узлов, соединенных высокопроизводительной сетью. В рамках данного раздела рассматривается устройство современных скоростных сетей. Изу-чаются принципы устройства сетей Ethernet и InfiniBand, механизмы классического взаимодействия путем

передачи сообщений и прямого удаленного доступа к памяти (RDMA), «выгрузка» логики взаимодействия (offloading), классические топологии высокопроизводительных сетей.

3. **Базовые средства многопоточного программирования**
   Наиболее распространенной и широко доступной платформой для высокопроизводительных вычислений являются многоядерные системы с общей памятью. К этому классу относятся практи-чески любой персональный компьютер. Поэтому особое внимание в курсе уделяется основной па-радигме разработки программ для подобных систем: многопоточному программированию. Рас-сматриваются понятия потока и процесса с точки зрения операционной системы. Для кратко рас-сматриваются классическая библиотека для разработки многопоточных программ: POSIX Threads, на примере которой изучаются способы создания потоков и их синхронизация: семафоры, мьютек-сы, блокировки на чтение-запись, условные переменные. POSIX Threads является простым, эффективным, но достаточно неудобным в использовании средством разработки многопоточных приложений, на смену которому пришли многопоточные расширения современных диалектов языка C++. В рамках этой темы рассматриваются классы и ме-тоды для создания и синхронизации потоков. Изучаются атомарные данные, операции над ними, и модели констистентности памяти. Изучается концепция безблокировочного взаимодействия пото-ков (lock-free programming).
4. **Программирование графических ускорителей**
   Рассматривается CUDA – базовый пакет для программирования графических ускорителей. Изучается процесс сборки, запуска и отладки программ для графических ускорителей. Рассматриваются концепции вычислительного ядра, потоков, блоков. Определяется модель исполнения CUDA-программы – модель SIMT, понятие расхождения потоков и его влияния на производительность. Рассматриваются виды памяти, используемые в графических ускорителях: глобальная, локальная, константная, кэш-память. Также уделяется внимание передаче данных с «хоста» на «устройство» посредством копирования или с использованием отображаемой памяти устройства. Изучаются способы повышения эффективности использования памяти с помощью агрегации обра-щений к памяти и других приемов. Также дается обзор поддержки машинного обучения в совре-менных графических процессорах, рассматриваются «тензорные» операции.
5. **Высокоуровневые средства многопоточного программирования**
   Многопоточное программирование на основе OpenMP. OpenMP позволяет распараллеливать программы, прилагая минимальные усилия. Код про-граммы изменяется с помощью специальных директив, указывающих компилятору, каким образом провести распараллеливание последующего цикла или блока. Рассматриваются директивы органи-зации параллельного выполнения, синхронизации, распределения работы, векторизации. Недавно появившийся инструментарий OpenACC построен на той же парадигме директив транслятору, что и OpenMP, но в отличие от последнего, позволяет разрабатывать программы, ко-торые могут выполняться как на многоядерном центральном процессоре, так и на графическом ускорителе. Рассматриваются основные инструменты для работы с OpenACC, базовые директивы для инициации параллельного выполнения, распределения вычислительной нагрузки и обмена данными между потоками. Программная среда OpnCL позволяет разрабатывать программы для широкого класса много-ядерных платформ, включая центральные процессоры и графические ускорители, а также FPGA. В курсе рассматривается архитектура среды OpenCL, основные концепции, понятия устройства и яд-ра. Изучаются методы разработки параллельных программ, способных задействовать различные устройства в разнородной высокопроизводительной среде.
6. **Средства разработки программ для систем с распределенной памятью**
   Основным средством разработки программ для систем с распределенной памятью, т.е. вычислительных кластеров, является библиотека MPI (Message Passing Interface). В курсе рассматривается структура библиотеки, способ сборки и запуска MPI-программы. Изучаются понятия коммуникатора, средств управления коммуникаторами в MPI-приложении. Система типов в MPI-программах, базовые и производные типы. Взаимодействие процессов в MPI. Рассматриваются парные и коллективные взаимодействия MPI-процессов, синхронные и асинхронные взаимодействия. Рассматриваются односторонние обмены, использующие возможно-сти прямого удаленного доступа к памяти. Рассматривается практика разработки и отладки MPI-программ, источники ошибок в MPI-программах, взаимные блокировки, недетерминизм. Разделённое глобальное адресное пространство (partitioned global address space, сокр. PGAS) – одна из перспективных моделей разработки программ для систем с распределенной памятью, при которой вся память параллельного вычислительного комплекса

является адресуемой и разделена на логические разделы, каждый из которых локален для какого-то процесса или потока. В курсе рассматривается программный интерфейс OpenSHMEM, разработанный международным консорциумом. Изучаются библиотечные вызовы для записи и чтения данных, синхронизации и коллективных обменов. Концепции PGAS иллюстрируются на различных примерах.

3. Оценивание

- **ЛАБ1**, Не блокирующее, Лабораторная работа
  Средства многопоточного программирования
- **ЛАБ2**, Не блокирующее, Лабораторная работа
  Средства разработки программ для графических процессоров
- **ЛАБ3**, Не блокирующее, Лабораторная работа
  Высокоуровневые средства многопоточного программирования
- **ЛАБ4**, Не блокирующее, Лабораторная работа
  Средства разработки высокопроизводительных приложений для систем с распределенной памятью
- **ЭКЗ**, Не блокирующее, Экзамен (устный)
  Итоговый экзамен

Формула округления: Стандартное арифметическое округление
Шкала оценки: Десятибалльная
Вид формулы оценивания: Линейная
Формула оценивания:

Окончательная оценка = Округление(1/8 * ЛАБ1 + 1/8 *ЛАБ2 + 1/8 * ЛАБ3 + 1/8 *ЛАБ4 + 1/2*ЭКЗ) - Окончательная оценка за дисциплин

4. Примеры оценочных средств

Лабораторная работа: разработка многопоточного варианта LU-разложения

Примеры вопросов на итоговой аттестации (экзамен):

Многопоточное программирование: обзор технологий POSIX Threads, функции для создания и завершения потоков.
2. Проблема недетерминизма в многопоточных программах.
3. Поддержка синхронизации потоков в POSIX Threads. Семафоры.
4. Критические секции в POSIX Threads.
5. Условные переменные в POSIX Threads.
6. Блокирование на чтение-запись в POSIX Threads.
7. Общая характеристика пакета OpenMP. Последовательные и параллельные участки. Ди-рективы распараллеливания.
8. Общие правила для переменных в OpenMP программе, опции private, shared, firstprivate, директива threadprivate.
9. Директивы распределения работы в OpenMP.

5. Ресурсы

5.1. Рекомендуемая основная литература

п/п	Наименование
1	*Эндрюс Г. Р.* Основы многопоточного, параллельного и распределенного программирования
2	*Petersen W. P.* Introduction to parallel computing, New York; Oxford: Oxford University Press, 2004
3	*А. В. Боресков, А. А. Харламов, Н. Д. Марковский, и др.* Параллельные вычисления на GPU, Изд-

	во Моск. ун-та
4	*Сандерс Дж.* Технология CUDA в примерах: введение в программирование графических процессоров

5.2. Рекомендуемая дополнительная литература
Не требуется

5.3. Программное обеспечение

п/п	Наименование	Условия доступа/скачивания
1	Microsoft Windows 7 Professional RUS Microsoft Windows 8.1 Professional RUS Microsoft Windows 10	*Из внутренней сети университета (договор)*
2	Microsoft Office Professional Plus 2010	*Из внутренней сети университета (договор)*

5.4. Профессиональные базы данных, информационные справочные системы, интернет-ресурсы (электронные образовательные ресурсы)

п/п	Наименование	Условия доступа/скачивания
	Профессиональные базы данных, информационно-справочные системы	
1	Электронно-библиотечная система Юрайт	URL: https://biblio-online.ru/
	Интернет-ресурсы (электронные образовательные ресурсы)	
1	Открытое образование	URL: https://openedu.ru/

5.5. Материально-техническое обеспечение дисциплины
Учебные аудитории для лекционных по дисциплине обеспечивают использование и демонстрацию тематических иллюстраций, соответствующих программе дисциплины в составе:
- ПЭВМ с доступом в Интернет (операционная система, офисные программы, антивирусные программы);
- мультимедийный проектор с дистанционным управлением.
Учебные аудитории для семинарских и самостоятельных занятий по дисциплине оснащены ПЭВМ, с возможностью подключения к сети Интернет и доступом к электронной информационно-образовательной среде НИУ ВШЭ.

Компьютерные классы оборудованы ПЭВМ с доступом в Интернет, операционными системами и программным обеспечением, необходимыми для освоения дисциплины. При необходимости допускается замена оборудования его виртуальными аналогами.

6. Особенности организации обучения для лиц с ограниченными возможностями здоровья и инвалидов

В случае необходимости, обучающимся из числа лиц с ограниченными возможностями здоровья (по заявлению обучающегося) а для инвалидов также в соответствии с индивидуальной программой реабилитации инвалида, могут предлагаться следующие варианты восприятия учебной информации с учетом их индивидуальных психофизических особенностей, в том числе с применением электронного обучения и дистанционных технологий:
6.1.1. *для лиц с нарушениями зрения:* в печатной форме увеличенным шрифтом; в форме электронного документа; в форме аудиофайла (перевод учебных материалов в аудиоформат); в печатной форме на языке Брайля; индивидуальные консультации с привлечением тифлосурдопереводчика; индивидуальные задания и консультации.
6.1.2. *для лиц с нарушениями слуха:* в печатной форме; в форме электронного документа; видеоматериалы с субтитрами; индивидуальные консультации с привлечением сурдопереводчика; индивидуальные задания и консультации.
6.1.3. *для лиц с нарушениями опорно-двигательного аппарата:* в печатной форме; в форме электронного документа; в форме аудиофайла; индивидуальные задания и консультации.