

**Федеральное государственное автономное образовательное  
учреждение высшего образования  
«Московский физико-технический институт  
(национальный исследовательский университет)»**

**УТВЕРЖДЕНО**

**Директор физтех-школы  
прикладной математики и  
информатики  
А.М. Райгородский**

	<b>Рабочая программа дисциплины (модуля)</b>
<b>по дисциплине:</b>	Алгоритмы во внешней памяти
<b>по направлению:</b>	Информатика и вычислительная техника
<b>профиль подготовки:</b>	Физтех-школа Прикладной Математики и Информатики кафедра анализа данных
<b>курс:</b>	4
<b>квалификация:</b>	бакалавр

Семестр, формы промежуточной аттестации: 7 (осенний) - Дифференцированный зачет

Аудиторных часов: 30 всего, в том числе:

лекции: 15 час.

семинары: 15 час.

лабораторные занятия: 0 час.

Самостоятельная работа: 15 час.

Всего часов: 45, всего зач. ед.: 1

Количество контрольных работ, заданий: 1

Программу составил: С.Н. Федотов, канд. физ.-мат. наук

Программа обсуждена на заседании кафедры анализа данных 06.03.2020

## Аннотация

Классические курсы по алгоритмам и структурам данных обычно используют стандартную модель для оценки сложности алгоритмов, то есть измеряют асимптотически количество элементарных операций процессора на худшем входе при фиксированном размере входных данных. Такой подход оказывается не всегда приемлем на практике по разным причинам. Например, если вы запустите классический алгоритм сортировки на наборе данных, которые живут на жестком диске и не помещаются в оперативную память, то время работы будет в сотни, а то и тысячи раз превосходить ожидаемые оценки, полученный на основе  $O$ -большое асимптотик. Подобная ситуация, может наблюдаться даже в случае, когда все данные помещаются в оперативную память, однако ваш алгоритм выполняет много случайных доступов к памяти. Масштаб трагедии в таком случае будет не таким существенным, однако, есть случаи, когда, применив специальные техники, можно ускорить алгоритм в разы отказавшись от случайных обращений.

Приведенные примеры говорят о том, что для построения алгоритмов в подобных случаях надо использовать другие модели для оценки эффективности алгоритмов, а также надо хорошо понимать, как устроена работа с жестким диском и памятью, чтобы получить хорошо работающее решение на практике.

В данном курсе мы рассмотрим следующие темы:

- Алгоритмы во внешней памяти;
- Cache-oblivious алгоритмы;
- Алгоритмы потоковой обработки данных.

Отметим, что в курсе не будет изучаться распределенные алгоритмы и модель MapReduce вычислений, которые обычно применяются для обработки больших объемов данных.

На семинарах мы рассмотрим вопросы устройства оперативной памяти и жесткого диска в современных компьютерах, а также научимся эффективно реализовывать на практике часть алгоритмов, рассказанных на лекциях.

## 1. Цели и задачи

### Цель дисциплины

ознакомление студентов с основными принципами построения алгоритмов для работы с данными, которые не помещаются в оперативную память компьютера.

### Задачи дисциплины

освоение студентами базовых знаний (понятий, концепций, методов и моделей) в области работы с большими данными; приобретение теоретических знаний и практических умений и навыков в области построения алгоритмов, работающих с данными, расположенными во внешней памяти, или в условиях недостаточности оперативной памяти.

## 2. Перечень формируемых компетенций

Освоение дисциплины направлено на формирование следующих компетенций:

Код и наименование компетенции	Индикаторы достижения компетенции
УК-1 Способен осуществлять поиск, критический анализ и синтез информации, применять системный подход для решения поставленных задач	УК-1.1 Анализирует задачу, выделяя этапы ее решения, действия по решению задачи
	УК-1.2 Находит, критически анализирует и выбирает информацию, необходимую для решения поставленной задачи
	УК-1.3 Рассматривает различные варианты решения задачи, оценивает их преимущества и недостатки
	УК-1.4 Грамотно, логично, аргументированно формирует собственные суждения и оценки
	УК-1.5 Определяет и оценивает практические последствия возможных вариантов решения задачи
УК-2 Способен определять круг задач в рамках поставленной цели и выбирать оптимальные способы их решения, исходя из	УК-2.1 Формулирует совокупность взаимосвязанных задач в рамках поставленной цели работы, обеспечивающих ее достижение. Определяет ожидаемые результаты решения поставленных задач

оптимальные способы их решения, исходя из действующих правовых норм, имеющихся ресурсов и ограничений	УК-2.2 Проектирует решение конкретной задачи проекта, выбирая оптимальный способ ее решения, исходя из действующих правовых норм и имеющихся ресурсов и ограничений
ОПК-4 Способен осуществлять сбор и обработку научно-технической и (или) технологической информации для решения фундаментальных и прикладных задач	ОПК-4.1 Владеет методами научного поиска и интеллектуального анализа информации при решении задач профессиональной деятельности
	ОПК-4.2 Знает основные источники научно-технической и (или) технологической информации в области профессиональной деятельности
	ОПК-4.3 Умеет составлять аннотации, рефераты, библиографические перечни и обзоры информации в области своей профессиональной деятельности
	ОПК-4.4 Владеет навыками работы с компьютером и компьютерными сетями с целью получения, хранения и обработки научной (технической, технологической) информации
ОПК-5 Способен участвовать в проведении фундаментальных и прикладных исследований и разработок, самостоятельно осваивать новые теоретические, в том числе, математические методы исследований и работать на современной экспериментальной научно-исследовательской, измерительно-аналитической и технологической аппаратуре)	ОПК-5.1 Способен решать поставленные задачи в области теоретических и экспериментальных исследований и разработок
	ОПК-5.2 Обладает способностью к освоению новых знаний на основе изучения литературы, научных статей и других источников
	ОПК-5.3 Способен к профессиональной эксплуатации современной экспериментальной научно-исследовательской (измерительно-аналитической и технологической) аппаратуры
ПК-1 Способен ставить, формализовывать и решать задачи, в том числе разрабатывать и исследовать математические модели изучаемых явлений и процессов, системно анализировать научные проблемы, получать новые научные результаты	ПК-1.1 Способен находить, анализировать и обобщать информацию об актуальных результатах исследований в рамках тематической области своей профессиональной деятельности
	ПК-1.3 Способен применять теоретические и (или) экспериментальные методы исследований к конкретной научной задаче и интерпретировать полученные результаты
	ПК-1.2 Способен выдвигать гипотезы, строить математические модели для описания изучаемых явлений и процессов, оценить качество разработанной модели
ПК-2 Способен самостоятельно или в качестве члена (руководителя) малого коллектива организовывать и проводить научные исследования и их апробацию	ПК-2.1 Знает принципы построения научной работы, методы сбора и анализа полученного материала, способы аргументации
	ПК-2.2 Способен планировать и проводить научные исследования самостоятельно или в качестве члена (руководителя) малого научного коллектива
	ПК-2.3 Способен проводить апробацию результатов научно-исследовательской работы посредством публикации научных статей и участия в конференциях

### 3. Перечень планируемых результатов обучения по дисциплине (модулю)

В результате освоения дисциплины обучающиеся должны

знать:

- основные алгоритмы для работы с данными, находящимися во внешней памяти, важнейшие cash-oblivious алгоритмы. Устройство SDD-дисков. Архитектуру компьютера, устройство жёсткого диска и процессорного кэша, особенности их устройства в системе Linux.

уметь:

- понять поставленную задачу; реализовать собственный алгоритм для работы с данными во внешней памяти;
- строить различные структуры данных во внешней памяти; работать с графами во внешней памяти;
- самостоятельно находить способы выполнения поставленных задач, в том числе и нестандартных, и проводить их анализ;
- самостоятельно видеть следствия полученных результатов. Делать оценки производительности алгоритмов.

владеть:

- навыками освоения большого объема информации и решения задач (в том числе, сложных); навыками самостоятельной работы и освоения новых дисциплин;
- культурой постановки, анализа и решения задач, возникающих при работе с большими данными.

#### 4. Содержание дисциплины (модуля), структурированное по темам (разделам) с указанием отведенного на них количества академических часов и видов учебных занятий

##### 4.1. Разделы дисциплины (модуля) и трудоемкости по видам учебных занятий

№	Тема (раздел) дисциплины	Трудоемкость по видам учебных занятий, включая самостоятельную работу, час.			
		Лекции	Семинары	Лаборат. работы	Самост. работа
1	Введение. Сортировка во внешней памяти	1	1		1
2	Задача о ранжировании списка (list ranking) и ее приложения	1	1		1
3	Онлайн-деревья поиска	1	1		1
4	Оффлайн-деревья поиска	1	1		1
5	Кучи во внешней памяти	1	1		1
6	Графы, простейшие алгоритмы	2	2		1
7	Обходы графов	1	1		1
8	Связные компоненты и оптимальные остовные деревья	1	1		1
9	Кеширование (caching)	1	1		1
10	Нечувствительные к кешированию (cache-oblivious) алгоритмы и структуры данных	1	1		2
11	Потоковые алгоритмы	1	1		2
12	Хеширование (hashing) и создание эскизов (sketching)	3	3		2
Итого часов		15	15		15
Подготовка к экзамену		0 час.			
Общая трудоёмкость		45 час., 1 зач.ед.			

##### 4.2. Содержание дисциплины (модуля), структурированное по темам (разделам)

Семестр: 7 (Осенний)

###### 1. Введение. Сортировка во внешней памяти

-Модель вычислений во внешней памяти, измерение сложности алгоритмов

- Оценки сложности сортировки во внешней памяти
- Нижние оценки для I/O-сложности сортировки
- MergeSort
- DistributionSort

## 2. Задача о ранжировании списка (list ranking) и ее приложения

- Понятие о ранжировании списка
- Эйлеровы обходы деревьев

## 3. Онлайн-деревья поиска

- B-деревья (B-trees) поиска и их разновидности (B+, B\*)
- Использование B-деревьев в DBMS-системах

## 4. Оффлайн-деревья поиска

- Buffered-деревья
- Buffered repository-деревья

## 5. Кучи во внешней памяти

- Реализация куч на основе buffered-деревьев
- Time-forward processing и его приложения
- Tournament-деревья

## 6. Графы, простейшие алгоритмы

- Представление графов во внешней памяти
- Подграфы максимальной плотности и их приближенное вычисление
- Эйлеровы обходы графов

## 7. Обходы графов

- Обход в ширину: алгоритмы Munagala-Ranade и Mehlhorn-Meyer
- Обход в глубину

## 8. Связные компоненты и оптимальные остовные деревья

- Техника стягиваний, алгоритм Борувки
- Техника спарсификации

## 9. Кеширование (caching)

- Кеши, их типы и организация
- Стратегии замещения
- Competitive-анализ
- Оптимальная оффлайн-стратегия

## 10. Нечувствительные к кешированию (cache-oblivious) алгоритмы и структуры данных

- Задача об умножении матриц
- Деревья van Emde Boas
- Оптимальный статический бинарный поиск
- Оптимальная сортировка (funnel sort)

-COLA

#### 11. Поточковые алгоритмы

- Поиск порядковых статистик: точных (Munro-Paterson) и приближенных (Manku-Rajagopalan-Lindsay)
- Приближенный подсчет числа различных элементов
- Поиск частотных элементов

#### 12. Хеширование (hashing) и создание эскизов (sketching)

- Гипотеза равномерного хеширования, универсальные семейства хеш-функций
- Хеш-функции, учитывающие близость (locality-sensitive hashing), эскизы (sketches)
- Мера сходства Жаккара, система LSH на основе min-wise перестановок
- Косинусная близость, система LSH на основе случайных проекций
- CountMin-эскизы

### **5. Описание материально-технической базы, необходимой для осуществления образовательного процесса по дисциплине (модулю)**

учебная аудитория, оснащенная медиапроектором и экраном.

### **6.Перечень рекомендуемой литературы**

#### Основная литература

1. Алгоритмы. Руководство по разработке [Текст] = The Algorithm Design Manual : [учеб. пособие для вузов] / С. Скиена ; [пер. с англ. С. Таранушенко] .— 2-е изд. — СПб. : БХВ-Петербург, 2011, 2014 .— 720 с.

#### Дополнительная литература

1. Алгоритмы [Текст] /Т. Кормен, Ч. Лейзерсон, Р. Ривест, построение и анализ. -М., МЦНМО, 2001

### **7. Перечень ресурсов информационно-телекоммуникационной сети "Интернет", необходимых для освоения дисциплины (модуля)**

Не используются

### **8. Перечень информационных технологий, используемых при осуществлении образовательного процесса по дисциплине (модулю), включая перечень необходимого программного обеспечения и информационных справочных систем (при необходимости)**

пакет программ Microsoft office.

### **9. Методические указания для обучающихся по освоению дисциплины (модуля)**

для успешной сдачи дифференцированного зачета рекомендуется внимательно выполнять практические задания, а также тщательно прорабатывать рекомендуемую литературу.

**ОЦЕНОЧНЫЕ МАТЕРИАЛЫ ПО ДИСЦИПЛИНЕ (МОДУЛЮ)**

**по направлению:** Информатика и вычислительная техника

**профиль подготовки:** Физтех-школа Прикладной Математики и Информатики  
кафедра анализа данных

**курс:** 4

**квалификация:** бакалавр

Семестр, формы промежуточной аттестации: 7 (осенний) - Дифференцированный зачет

**Разработчик:** С.Н. Федотов, канд. физ.-мат. наук

## 1. Компетенции, формируемые в процессе изучения дисциплины

Код и наименование компетенции	Индикаторы достижения компетенции
УК-1 Способен осуществлять поиск, критический анализ и синтез информации, применять системный подход для решения поставленных задач	УК-1.1 Анализирует задачу, выделяя этапы ее решения, действия по решению задачи
	УК-1.2 Находит, критически анализирует и выбирает информацию, необходимую для решения поставленной задачи
	УК-1.3 Рассматривает различные варианты решения задачи, оценивает их преимущества и недостатки
	УК-1.4 Грамотно, логично, аргументированно формирует собственные суждения и оценки
	УК-1.5 Определяет и оценивает практические последствия возможных вариантов решения задачи
УК-2 Способен определять круг задач в рамках поставленной цели и выбирать оптимальные способы их решения, исходя из действующих правовых норм, имеющихся ресурсов и ограничений	УК-2.1 Формулирует совокупность взаимосвязанных задач в рамках поставленной цели работы, обеспечивающих ее достижение. Определяет ожидаемые результаты решения поставленных задач
	УК-2.2 Проектирует решение конкретной задачи проекта, выбирая оптимальный способ ее решения, исходя из действующих правовых норм и имеющихся ресурсов и ограничений
ОПК-4 Способен осуществлять сбор и обработку научно-технической и (или) технологической информации для решения фундаментальных и прикладных задач	ОПК-4.1 Владеет методами научного поиска и интеллектуального анализа информации при решении задач профессиональной деятельности
	ОПК-4.2 Знает основные источники научно-технической и (или) технологической информации в области профессиональной деятельности
	ОПК-4.3 Умеет составлять аннотации, рефераты, библиографические перечни и обзоры информации в области своей профессиональной деятельности
	ОПК-4.4 Владеет навыками работы с компьютером и компьютерными сетями с целью получения, хранения и обработки научной (технической, технологической) информации
ОПК-5 Способен участвовать в проведении фундаментальных и прикладных исследований и разработок, самостоятельно осваивать новые теоретические, в том числе, математические методы исследований и работать на современной экспериментальной научно-исследовательской, измерительно-аналитической и технологической аппаратуре)	ОПК-5.1 Способен решать поставленные задачи в области теоретических и экспериментальных исследований и разработок
	ОПК-5.2 Обладает способностью к освоению новых знаний на основе изучения литературы, научных статей и других источников
	ОПК-5.3 Способен к профессиональной эксплуатации современной экспериментальной научно-исследовательской (измерительно-аналитической и технологической) аппаратуры
ПК-1 Способен ставить, формализовывать и решать задачи, в том числе разрабатывать и исследовать математические модели изучаемых явлений и процессов, системно анализировать научные проблемы, получать новые научные результаты	ПК-1.1 Способен находить, анализировать и обобщать информацию об актуальных результатах исследований в рамках тематической области своей профессиональной деятельности
	ПК-1.3 Способен применять теоретические и (или) экспериментальные методы исследований к конкретной научной задаче и интерпретировать полученные результаты
	ПК-1.2 Способен выдвигать гипотезы, строить математические модели для описания изучаемых явлений и процессов, оценить качество разработанной модели



ПК-2 Способен самостоятельно или в качестве члена (руководителя) малого коллектива организовывать и проводить научные исследования и их апробацию	ПК-2.1 Знает принципы построения научной работы, методы сбора и анализа полученного материала, способы аргументации
	ПК-2.2 Способен планировать и проводить научные исследования самостоятельно или в качестве члена (руководителя) малого научного коллектива
	ПК-2.3 Способен проводить апробацию результатов научно-исследовательской работы посредством публикации научных статей и участия в конференциях

## 2. Показатели оценивания компетенций

В результате изучения дисциплины «Алгоритмы во внешней памяти» обучающийся должен:

### знать:

- основные алгоритмы для работы с данными, находящимися во внешней памяти, важнейшие cash-oblivious алгоритмы. Устройство SSD-дисков. Архитектуру компьютера, устройство жёсткого диска и процессорного кэша, особенности их устройства в системе Linux.

### уметь:

- понять поставленную задачу; реализовать собственный алгоритм для работы с данными во внешней памяти;
- строить различные структуры данных во внешней памяти; работать с графами во внешней памяти;
- самостоятельно находить способы выполнения поставленных задач, в том числе и нестандартных, и проводить их анализ;
- самостоятельно видеть следствия полученных результатов. Делать оценки производительности алгоритмов.

### владеть:

- навыками освоения большого объема информации и решения задач (в том числе, сложных); навыками самостоятельной работы и освоения новых дисциплин;
- культурой постановки, анализа и решения задач, возникающих при работе с большими данными.

## 3. Перечень типовых (примерных) вопросов, заданий, тем для подготовки к текущему контролю

Перечень вопросов для промежуточного контроля:

- Эйлеровы обходы деревьев;
- Buffered-деревья;
- Buffered repository-деревья;
- Tournament-деревья;
- Деревья van Emde Boas.

## 4. Перечень типовых (примерных) вопросов и тем для проведения промежуточной аттестации обучающихся

Примерный список вопросов к дифференцированному зачету:

1. Сортировка во внешней памяти
2. Онлайн-деревья поиска
3. Оффлайн-деревья поиска
4. Связные компоненты и оптимальные остовные деревья

Критерии оценивания

- оценка «отлично (10)» выставляется студенту, показавшему всесторонние, систематизированные, глубокие знания учебной программы дисциплины и умение уверенно применять их на практике при решении конкретных задач, свободное и правильное обоснование принятых решений
- оценка «отлично (9)» выставляется студенту, показавшему всесторонние, систематизированные, глубокие знания учебной программы дисциплины и умение применять их на практике при решении конкретных задач, свободное и правильное обоснование принятых решений
- оценка «отлично (8)» выставляется студенту, показавшему всесторонние систематизированные, глубокие знания учебной программы дисциплины и умение применять их на практике при решении конкретных задач, и правильное обоснование принятых решений
- оценка «хорошо (7)» выставляется студенту, если он твердо знает материал, грамотно и по существу излагает его, умеет применять полученные знания на практике, но допускает в ответе или в решении задач некоторые неточности;
- оценка «хорошо (6)» выставляется студенту, если он знает материал, грамотно и по существу излагает его, умеет применять полученные знания на практике, но допускает в ответе или в решении задач некоторые неточности;
- оценка «хорошо (5)» выставляется студенту, если он знает материал, и по существу излагает его, умеет применять полученные знания на практике, но допускает в ответе или в решении задач некоторые неточности;
- оценка «удовлетворительно (4)» выставляется студенту, показавшему фрагментарный, разрозненный характер знаний, недостаточно правильные формулировки базовых понятий, нарушения логической последовательности в изложении программного материала, но при этом он владеет основными разделами учебной программы, необходимыми для дальнейшего обучения и может применять полученные знания по образцу в стандартной ситуации;
- оценка «удовлетворительно (3)» выставляется студенту, показавшему фрагментарный, разрозненный характер знаний, недостаточно правильные формулировки базовых понятий, нарушения логической последовательности в изложении программного материала, но при этом он владеет фрагментарно основными разделами учебной программы, необходимыми для дальнейшего обучения и может применять полученные знания по образцу в стандартной ситуации;
- оценка «неудовлетворительно (2)» выставляется студенту, который не знает большей части основного содержания учебной программы дисциплины, допускает грубые ошибки в формулировках основных понятий дисциплины и не умеет использовать полученные знания при решении типовых практических задач
- оценка «неудовлетворительно (1)» выставляется студенту, который не знает формулировок основных понятий дисциплины.

## **5. Методические материалы, определяющие процедуры оценивания знаний, умений, навыков и (или) опыта деятельности**

Во время дифференцированного зачета студенты могут пользоваться материалом лекций.