Приложение 4

МИНИСТЕРСТВО НАУКИ И ВЫСШЕГО ОБРАЗОВАНИЯ РОССИЙСКОЙ ФЕДЕРАЦИИ

федеральное государственное автономное
образовательное учреждение высшего образования
«Национальный исследовательский Нижегородский государственный университет
им. Н.И. Лобачевского»

Институт информационных технологий, математики и механики

УТВЕРЖДЕНО

решением Ученого совета ННГУ

протокол № 10 от 02.12.2024 г.

Рабочая программа дисциплины

Основы программирования AI акселераторов

Уровень высшего образования
Магистратура

Направление подготовки / специальность
02.04.02 - Фундаментальная информатика и информационные технологии

Направленность образовательной программы
Искусственный интеллект

Форма обучения
очная

г. Нижний Новгород

2025 год начала подготовки

1. Место дисциплины в структуре ОПОП

Дисциплина ФТД.04 Основы программирования AI акселераторов является факультативом в образовательной программе.

2. Планируемые результаты обучения по дисциплине, соотнесенные с планируемыми результатами освоения образовательной программы (компетенциями и индикаторами достижения компетенций)

Формируемые компетенции (код, содержание компетенции)	**Планируемые результаты обучения по дисциплине (модулю), в соответствии с индикатором достижения компетенции**		**Наименование оценочного средства**	
	Индикатор достижения компетенции (код, содержание индикатора)	**Результаты обучения по дисциплине**	**Для текущего контроля успеваемости**	**Для промежуточной аттестации**
ПК-8: Способен к разработке новых алгоритмических, методических и технологических решений в конкретной сфере профессиональной деятельности	*ПК-8.1: Знает методику разработки новых алгоритмических, методических и технологических решений* *ПК-8.2: Умеет применять полученные знания для разработки новых алгоритмических, методических и технологических решений* *ПК-8.3: Имеет практический опыт составления технического задания на разработку информационной системы*	*ПК-8.1:* *Знать основы ИТ в области ИИ и иметь навыки анализа современного состояния науки и ИТ в области ИИ; архитектуру и принципы работы графических и нейронных процессоров, современные подходы к разработке, анализу и отладке программных систем на GPU и NPU.* *ПК-8.2:* *Уметь проектировать, разрабатывать и развивать ИТ-решения на основе анализа современного состояния науки и ИТ в области ИИ; проектировать, разрабатывать и реализовывать программное обеспечение для нейронных и графических процессоров.* *ПК-8.3:* *Иметь навыки управления разработкой на основе анализа современного состояния науки и ИТ в области ИИ; навыками и методиками анализа и оптимизации производительности приложений на NPU и GPU.*	*Практическое задание*	*Зачёт: Контрольные вопросы*

3. Структура и содержание дисциплины

3.1 Трудоемкость дисциплины

	очная
Общая трудоемкость, з.е.	**1**
Часов по учебному плану	**36**
в том числе	
аудиторные занятия (контактная работа):	
- занятия лекционного типа	**16**
- занятия семинарского типа (практические занятия / лабораторные работы)	**16**
- КСР	**1**
самостоятельная работа	**3**
Промежуточная аттестация	**0** **Зачёт**

3.2. Содержание дисциплины

(структурированное по темам (разделам) с указанием отведенного на них количества академических часов и виды учебных занятий)

Наименование разделов и тем дисциплины	Всего (часы)	в том числе			
		Контактная работа (работа во взаимодействии с преподавателем), часы из них			Самостоятельная работа обучающегося, часы
		Занятия лекционного типа	Занятия семинарского типа (практические занятия/лабораторные работы), часы	Всего	
	ооо	ооо	ооо	ооо	ооо
Введение в программирование для ИИ-ускорителей	4	2	2	4	
Введение в архитектуру NPU/GPU	4	2	2	4	
Язык программирования CUDA C++	10	4	4	8	2
Использование тензорных ядер CUDA для оптимизации задач в области ИИ	4	2	2	4	
Язык программирования AscendC	9	4	4	8	1
Оптимизация приложений AscendC для ИИ- ускорителей семейства Ascend	4	2	2	4	
Аттестация	0				
КСР	1			1	
Итого	36	16	16	33	3

Содержание разделов и тем дисциплины

1. Введение в программирование ускорителей для задач искусственного интеллекта. Программные модели. Архитектуры NPU и GPU ускорителей.
2. Язык программирования CUDA C++. CUDA Host API. Потоковое (stream) исполнение и синхронизация с помощью событий. Общая (unified) память.
3. Оптимизация приложений на языке CUDA C++. Использование тензорных ядер для ускорения ИИ-вычислений. Библиотека CUTLASS.
4. Язык программирования AscendC. Обзор архитектуры Huawei Ascend и программных моделей.
5. Оптимизация приложений на языке AscendC. Векторные, тензорные (cube) и смешанные вычисления.

4. Учебно-методическое обеспечение самостоятельной работы обучающихся

Самостоятельная работа обучающихся включает в себя подготовку к контрольным вопросам и заданиям для текущего контроля и промежуточной аттестации по итогам освоения дисциплины приведенным в п. 5.

Самостоятельная работа обучающихся включает в себя подготовку к контрольным вопросам и заданиям для текущего контроля и промежуточной аттестации по итогам освоения дисциплины.

1. Основы работы с технологией CUDA / Боресков А.В., Харламов А.А. - Москва : ДМК-пресс, 2010., https://e-lib.unn.ru/MegaPro/UserEntry?Action=FindDocs&ids=636496&idb=0
2. Руководство по программированию на AscendC - https://www.hiascend.com/document/detail/en/canncommercial/700/operatordev/Ascendcopdevg/atlas_ascendc_10_0001.html

5. Фонд оценочных средств для текущего контроля успеваемости и промежуточной аттестации по дисциплине (модулю)

5.1 Типовые задания, необходимые для оценки результатов обучения при проведении текущего контроля успеваемости с указанием критериев их оценивания:

5.1.1 Типовые задания (оценочное средство - Практическое задание) для оценки сформированности компетенции ПК-8:

1. *Лабораторная работа (проект) «Сверточный уровень нейросети на языке CUDA C++».*

Требуется разработать программу, реализующую типовой двумерный сверточный слой нейросети на языке CUDA C++. Исходными данными являются монохромное изображение и двумерный фильтр. Результатом работы будет матрица, получившаяся после применения сверточного преобразования над входным изображением и фильтром. Оценивание результатов выполняется в смысле а) работоспособности программы б) понимания применения заданного алгоритма и особенностей его эффективной реализации с использованием различных возможностей языка CUDA C++ и имеющегося аппаратного обеспечения (GPU).
Подробное описание задачи и ее обсуждение происходит в начале семестра. Допускается корректировка формулировки с учетом научных, курсовых и личных предпочтений студентов.

2. Лабораторная работа (проект) «Сверточный уровень нейросети на языке AscendC».

Требуется разработать программу, реализующую типовой двумерный сверточный слой нейросети на языке AscendC. Исходными данными являются монохромное изображение и двумерный фильтр. Результатом работы будет матрица, получившаяся после применения сверточного преобразования над

входным изображением и фильтром. Оценивание результатов выполняется в смысле а) работоспособности программы б) понимания применения заданного алгоритма и особенностей его эффективной реализации с использованием различных возможностей языка AscendC и имеющегося аппаратного обеспечения (NPU).
Подробное описание задачи и ее обсуждение происходит в начале семестра. Допускается корректировка формулировки с учетом научных, курсовых и личных предпочтений студентов.

Критерии оценивания (оценочное средство - Практическое задание)

Оценка	Критерии оценивания
зачтено	Выполнены все или большая часть этапов решения задачи или задача решена со незначительными недочетами. Код и результаты работы представлены преподавателю в срок.
не зачтено	Выполнены не все лабораторные работы или выполнены не в полном объеме (представлено не полное описание этапов выполнения заданий, код работает некорректно, результаты работы не представлены преподавателю).

5.2. Описание шкал оценивания результатов обучения по дисциплине при промежуточной аттестации

Шкала оценивания сформированности компетенций

Уровень сформированности компетенций (индикатора достижения компетенций)	**плохо**	**неудовлетворительно**	**удовлетворительно**	**хорошо**	**очень хорошо**	**отлично**	**превосходно**
	не зачтено		**зачтено**				
Знания	Отсутствие знаний теоретического материала. Невозможность оценить полноту знаний вследствие отказа обучающегося от ответа	Уровень знаний ниже минимальных требований. Имели место грубые ошибки	Минимально допустимый уровень знаний. Допущено много негрубых ошибок	Уровень знаний в объеме, соответствующем программе подготовки. Допущено несколько негрубых ошибок	Уровень знаний в объеме, соответствующем программе подготовки. Допущено несколько несущественных ошибок	Уровень знаний в объеме, соответствующем программе подготовки. Ошибок нет.	Уровень знаний в объеме, превышающем программу подготовки.
Умения	Отсутствие минимальных умений. Невозможность оценить наличие умений вследствие отказа обучающегося от ответа	При решении стандартных задач не продемонстрированы основные умения. Имели место грубые ошибки	Продемонстрированы основные умения. Решены типовые задачи с негрубыми ошибками. Выполнены все задания, но не в	Продемонстрированы все основные умения. Решены все основные задачи с негрубыми ошибками. Выполнены все задания в полном	Продемонстрированы все основные умения. Решены все основные задачи. Выполнены все задания в полном объеме, но некоторые	Продемонстрированы все основные умения. Решены все основные задачи с отдельными несущественными	Продемонстрированы все основные умения. Решены все основные задачи. Выполнены все задания, в полном объеме без недочетов

			полном объеме	объеме, но некоторые с недочетами	с недочетами.	недочетами, выполнены все задания в полном объеме	
Навыки	Отсутствие базовых навыков. Невозможность оценить наличие навыков вследствие отказа обучающегося от ответа	При решении стандартных задач не продемонстрированы базовые навыки. Имели место грубые ошибки	Имеется минимальный набор навыков для решения стандартных задач с некоторыми недочетами	Продемонстрированы базовые навыки при решении стандартных задач с некоторыми недочетами	Продемонстрированы базовые навыки при решении стандартных задач без ошибок и недочетов	Продемонстрированы навыки при решении нестандартных задач без ошибок и недочетов	Продемонстрирован творческий подход к решению нестандартных задач

Шкала оценивания при промежуточной аттестации

Оценка		Уровень подготовки
зачтено	**превосходно**	Все компетенции (части компетенций), на формирование которых направлена дисциплина, сформированы на уровне не ниже «превосходно», продемонстрированы знания, умения, владения по соответствующим компетенциям на уровне выше предусмотренного программой
	отлично	Все компетенции (части компетенций), на формирование которых направлена дисциплина, сформированы на уровне не ниже «отлично».
	очень хорошо	Все компетенции (части компетенций), на формирование которых направлена дисциплина, сформированы на уровне не ниже «очень хорошо»
	хорошо	Все компетенции (части компетенций), на формирование которых направлена дисциплина, сформированы на уровне не ниже «хорошо».
	удовлетворительно	Все компетенции (части компетенций), на формирование которых направлена дисциплина, сформированы на уровне не ниже «удовлетворительно», при этом хотя бы одна компетенция сформирована на уровне «удовлетворительно»
не зачтено	**неудовлетворительно**	Хотя бы одна компетенция сформирована на уровне «неудовлетворительно».
	плохо	Хотя бы одна компетенция сформирована на уровне «плохо»

5.3 Типовые контрольные задания или иные материалы, необходимые для оценки результатов обучения на промежуточной аттестации с указанием критериев их оценивания:

5.3.1 Типовые задания (оценочное средство - Контрольные вопросы) для оценки сформированности компетенции ПК-8

1. Вычисления в области искусственного интеллекта (ИИ). Типы ускорителей для ИИ- вычислений.
2. Методы и технологии для программирования на нейронных и графических процессорах.
3. Архитектура графических процессоров. Общее описание.

4. Иерархия памяти графических процессоров.
5. Архитектура нейронных процессоров. Общее описание.
6. Сравнение архитектур NPU и GPU. Основные различия.
7. Технология NVIDIA CUDA. Основные принципы и инструменты.
8. Язык CUDA C++. Модель исполнения.
9. Язык CUDA C++. Модель памяти. Типы памяти.
10. CUDA Host API. Потоковое исполнение (streams) и события.
11. Механизмы общей (unified) памяти в CUDA.
12. Тензорные ядра. Особенности программирования. CUTLASS.
13. Подходы к оптимизации приложений на языке CUDA C++.
14. Язык AscendC. Модель исполнения.
15. Язык AscendC. Модель памяти.
16. Язык AscendC. Векторные вычисления.
17. Язык AscendC. Тензорные (cube) вычисления.
18. Язык AscendC. Смешанные вычисления.
19. Подходы к оптимизации приложений на языке AscendC.

Критерии оценивания (оценочное средство - Контрольные вопросы)

Оценка	Критерии оценивания
зачтено	Студент ответил на большую часть вопросов возможно с незначительными недочетами.
не зачтено	При ответе студент допускает грубые ошибки в основном материале и решении стандартных задач.

6. Учебно-методическое и информационное обеспечение дисциплины (модуля)

Основная литература:

1. Основы работы с технологией CUDA / Боресков А.В., Харламов А.А. - Москва : ДМК-пресс, 2010., https://e-lib.unn.ru/MegaPro/UserEntry?Action=FindDocs&ids=636496&idb=0.
2. Сандерс Дж. Технология CUDA в примерах: введение в программирование графических процессоров : монография / Сандерс Дж.; Кэндрот Э. - Москва : ДМК-пресс, 2013. - 232 с. - ISBN 978-5-94074-889-2., https://e-lib.unn.ru/MegaPro/UserEntry?Action=FindDocs&ids=772935&idb=0.

Дополнительная литература:

1. Малявко А. А. Параллельное программирование на основе технологий openmp, cuda, opencl, mpi : учебное пособие / А. А. Малявко. - 3-е изд. ; испр. и доп. - Москва : Юрайт, 2023. - 135 с. - (Высшее образование). - ISBN 978-5-534-14116-0. - Текст : электронный // ЭБС "Юрайт"., https://e-lib.unn.ru/MegaPro/UserEntry?Action=FindDocs&ids=847643&idb=0.
2. Многоядерные процессоры / Калачев А.В. - Москва : ИНТУИТ, 2016., https://e-lib.unn.ru/MegaPro/UserEntry?Action=FindDocs&ids=662986&idb=0.
3. Теория и практика параллельных вычислений / Гергель В.П. - Москва : ИНТУИТ, 2016., https://e-lib.unn.ru/MegaPro/UserEntry?Action=FindDocs&ids=663423&idb=0.

4. Рыбальченко Михаил Викторович. Архитектура информационных систем : Учебное пособие для вузов / Рыбальченко М. В. - Москва : Юрайт, 2016. - 91 с. - (Высшее образование). - ISBN 978-5-9916-9326-4 : 179.00. - Текст : электронный // ЭБС "Юрайт"., https://e-lib.unn.ru/MegaPro/UserEntry?Action=FindDocs&ids=566682&idb=0.

Программное обеспечение и Интернет-ресурсы (в соответствии с содержанием дисциплины):

1. Программное обеспечение CUDA - https://developer.nvidia.com/cuda-downloads
2. Руководство по программированию на CUDA C++ - https://docs.nvidia.com/cuda/cuda-c-programming-guide/index.html
3. Программное обеспечение Ascend CANN - https://www.hiascend.com/en/software/cann
4. Руководство по программированию на AscendC - https://www.hiascend.com/document/detail/en/canncommercial/700/operatordev/Ascendcopdevg/atlas_ascendc_10_0001.html

7. Материально-техническое обеспечение дисциплины (модуля)

Учебные аудитории для проведения учебных занятий, предусмотренных образовательной программой, оснащены мультимедийным оборудованием (проектор, экран), техническими средствами обучения, компьютерами.

Помещения для самостоятельной работы обучающихся оснащены компьютерной техникой с возможностью подключения к сети "Интернет" и обеспечены доступом в электронную информационно-образовательную среду.

Программа составлена в соответствии с требованиями ОС ННГУ по направлению подготовки/специальности 02.04.02 - Фундаментальная информатика и информационные технологии.

Автор(ы): Горшков Антон Валерьевич, кандидат технических наук.

Заведующий кафедрой: Мееров Иосиф Борисович, кандидат технических наук.

Программа одобрена на заседании методической комиссии от 02.12.2024, протокол № 5.