ПРИЛОЖЕНИЕ

МИНИСТЕРСТВО НАУКИ И ВЫСШЕГО ОБРАЗОВАНИЯ РОССИЙСКОЙ ФЕДЕРАЦИИ

ФЕДЕРАЛЬНОЕ ГОСУДАРСТВЕННОЕ БЮДЖЕТНОЕ ОБРАЗОВАТЕЛЬНОЕ УЧРЕЖДЕНИЕ ВЫСШЕГО ОБРАЗОВАНИЯ
«Рязанский государственный радиотехнический университет имени В.Ф. Уткина»

КАФЕДРА «ЭЛЕКТРОННЫЕ ВЫЧИСЛИТЕЛЬНЫЕ МАШИНЫ»

ОЦЕНОЧНЫЕ МАТЕРИАЛЫ

«Массово-параллельные вычисления»

Направление подготовки

09.03.01 Информатика и вычислительная техника

Профиль

Направленность (профиль) подготовки
«Программно-аппаратное обеспечение вычислительных комплексов и систем искусственного интеллекта»

Квалификация (степень) выпускника — бакалавр

Форма обучения — очная

Рязань

1 ОБЩИЕ ПОЛОЖЕНИЯ

Оценочные материалы – это совокупность учебно-методических материалов (практических заданий, описаний форм и процедур проверки), предназначенных для оценки качества освоения обучающимися данной дисциплины как части ОПОП.

Цель – оценить соответствие знаний, умений и владений, приобретенных обучающимся в процессе изучения дисциплины, целям и требованиям ОПОП в ходе проведения промежуточной аттестации.

Основная задача – обеспечить оценку уровня сформированности компетенций, закрепленных за дисциплиной.

Контроль знаний обучающихся проводится в форме промежуточной аттестации.

Промежуточная аттестация проводится в форме зачета. Форма проведения зачета – тестирование, письменный опрос по теоретическим вопросам.

1 ОПИСАНИЕ ПОКАЗАТЕЛЕЙ И КРИТЕРИЕВ ОЦЕНИВАНИЯ КОМПЕТЕНЦИЙ

Сформированность каждой компетенции (или ее части) в рамках освоения данной дисциплины оценивается по трехуровневой шкале:

1) пороговый уровень является обязательным для всех обучающихся по завершении освоения дисциплины;

2) продвинутый уровень характеризуется превышением минимальных характеристиксформированности компетенций по завершении освоения дисциплины;

3) эталонный уровень характеризуется максимально возможной выраженностью компетенций и является важным качественным ориентиром для самосовершенствования.

Уровень освоения компетенций, формируемых дисциплиной:

Описание критериев и шкалы оценивания тестирования:

Шкала оценивания	**Критерий**
3 балла (эталонный уровень)	уровень усвоения материала, предусмотренного программой: процент верных ответов на тестовые вопросы от 85 до 100%
2 балла (продвинутый уровень)	уровень усвоения материала, предусмотренного программой: процент верных ответов на тестовые вопросы от 70 до 84%
1 балл (пороговый уровень)	уровень усвоения материала, предусмотренного программой: процент верных ответов на тестовые вопросы от 50 до 69%
0 баллов	уровень усвоения материала, предусмотренного программой: процент верных ответов на тестовые вопросы от 0 до 49%

Описание критериев и шкалы оценивания теоретического вопроса:

Шкала оценивания	**Критерий**
3 балла (эталонный уровень)	выставляется студенту, который дал полный ответ на вопрос, показал глубокие систематизированные знания, смог привести примеры, ответил на дополнительные вопросы преподавателя
2 балла (продвинутый уровень)	выставляется студенту, который дал полный ответ на вопрос, но на некоторые дополнительные вопросы преподавателя ответил только с помощью наводящих вопросов
1 балл (пороговый уровень)	выставляется студенту, который дал неполный ответ на вопрос в билете и смог ответить на дополнительные вопросы только с помощью преподавателя
0 баллов	выставляется студенту, который не смог ответить на вопрос

Описание критериев и шкалы оценивания практического задания:

Шкала оценивания	**Критерий**

3 балла (эталонный уровень)	Задача решена верно
2 балла (продвинутый уровень)	Задача решена верно, но имеются неточности в логике решения
1 балл (пороговый уровень)	Задача решена верно, с дополнительными наводящими вопросами преподавателя
0 баллов	Задача не решена

На промежуточную аттестацию выносится тест, два теоретических вопроса и задача. Максимально студент может набрать 12 баллов. Итоговый суммарный балл студента, полученный при прохождении промежуточной аттестации, переводится в традиционную форму по системе «отлично», «хорошо», «удовлетворительно» и «неудовлетворительно».

Оценка «отлично» выставляется студенту, который набрал в сумме 12 баллов (выполнил все задания на эталонном уровне). Обязательным условием является выполнение всех предусмотренных в течение семестра практических заданий.

Оценка «хорошо» выставляется студенту, который набрал в сумме от 8 до 11 баллов при условии выполнения всех заданий на уровне не ниже продвинутого. Обязательным условием является выполнение всех предусмотренных в течение семестра практических заданий.

Оценка «удовлетворительно» выставляется студенту, который набрал в сумме от 4 до 7 баллов при условии выполнения всех заданий на уровне не ниже порогового. Обязательным условием является выполнение всех предусмотренных в течение семестра практических заданий.

Оценка «неудовлетворительно» выставляется студенту, который набрал в сумме менее 4 баллов или не выполнил всех предусмотренных в течение семестра практических заданий.

3 ПАСПОРТ ОЦЕНОЧНЫХ МАТЕРИАЛОВ ПО ДИСЦИПЛИНЕ

Контролируемые разделы (темы) дисциплины	***Код контролируемой компетенции (или её части)***	**Вид, метод, форма оценочного мероприятия**
Тема 1. Сферы применения специализированных вычислителей, их состав и характеристики	ПК-2.1	Экзамен
Тема 2. Типовые элементы аппаратной архитектуры вычислителя	ПК-2.1	Экзамен
Тема 3. Методы оценки и повышения производительности вычислителей	ПК-2.2	Экзамен
Тема 4. Низкоуровневые языки проектирования вычислителей	ПК-2.1, ПК-2.2, ПК-5.2, ПК-9.1	Экзамен
Тема 5. Подходы к проектированию высокопроизводительных вычислителей	ПК-2.1, ПК-2.2, ПК-5.1, ПК-5.2, ПК-17.1, ПК-9.1	Экзамен
Тема 6. Механизмы обмена данными	ПК-5.1, ПК-5.2, ПК-2.2, ПК-9.1	Экзамен
Тема 7. Вычисления на GPU	ПК-2.1, ПК-2.2, ПК-5.1, ПК-5.2, ПК-9.1, ПК-9.2, ПК-17.1	Экзамен
Тема 8. Встраиваемые решения (Embedded system)	ПК-2.1, ПК-5.1, ПК-5.2, ПК-9.1, ПК-9.2, ПК-17.1, ПК-17.2	Экзамен
Тема 9. Объединение идей GPU, NPU, в FPGA	ПК-2.1, ПК-5.1, ПК-5.2, ПК-9.1, ПК-9.2, ПК-17.1, ПК-17.2	Экзамен

4 ТИПОВЫЕ КОНТРОЛЬНЫЕ ЗАДАНИЯ ИЛИ ИНЫЕ МАТЕРИАЛЫ

4.1.Промежуточная аттестация в форме экзамена

Код компетенции	Результаты освоения ОПОП Содержание компетенций
ПК-2	Способен проектировать и разрабатывать программное обеспечение

ПК-2.1. Проектирует и разрабатывает программное обеспечение

Типовые тестовые вопросы

1. Закон Мура, долгое время являвшийся основным драйвером роста производительности процессоров, в современную эпоху столкнулся с физическими ограничениями. Какое из следующих утверждений наиболее точно описывает современную интерпретацию этого закоя и основную стратегию дальнейшего роста производительности?

- а) Закон Мура полностью перестал действовать, и рост производительности процессоров прекратился.
- б) Удвоение тактовой частоты процессоров продолжается каждые 2 года, что является основной стратегией.
- в) Удвоение количества транзисторов на кристалле продолжается, но связано с растущими затратами и сложностями, а основной стратегией стало распараллеливание вычислений (многоядерность).
- г) Рост производительности теперь достигается исключительно за счет перехода на новые материалы, такие как кремний-германий.

Правильный ответ: в)

2. При проектировании бортового вычислителя для космического аппарата критически важными являются определенные требования, которые отличают его от стандартного серверного процессора. Какое из перечисленных требований является для такого вычислителя НАИБОЛЕЕ специфичным и критичным?

- а) Высокая тактовая частота для обеспечения максимальной производительности в однопоточных задачах.
- б) Способность работать со стандартной версией операционной системы общего назначения, такой как Windows.
- в) Повышенная стойкость к радиации и экстремальным перепадам температур.
- г) Наличие мощной системы охлаждения с жидкостным циркулированием.

Правильный ответ: в)

3. Закон Амдала определяет теоретическое ускорение программы при увеличении количества вычислительных ресурсов. Какой вывод является ключевым следствием из этого закона для разработчика программного обеспечения?

- а) Любую программу можно ускорить практически линейно, просто увеличивая количество ядер процессора.
- б) Максимальное ускорение программы ограничено долей последовательного (не поддающегося распараллеливанию) кода.

- в) Производительность программы зависит исключительно от тактовой частоты самого быстрого ядра в системе.
- г) Эффективность распараллеливания не зависит от архитектуры алгоритма.

Правильный ответ: б)

4. Специализированные вычислители (такие как GPU, FPGA, TPU) получают широкое распространение для решения определенных классов задач. Какой из перечисленных ниже классов задач НАИБОЛЕЕ эффективно решается с использованием графических процессоров (GPU)?

- а) Выполнение операционной системы и управление устройствами ввода-вывода.
- б) Обработка транзакций в базах данных с высокой частотой случайных запросов.
- в) Массово-параллельные вычисления над большими массивами данных с простыми, однотипными операциями (например, матричные вычисления, рендеринг графики, обучение нейронных сетей).
- г) Последовательные вычисления с сложной логикой ветвления и зависимостями по данным.

Правильный ответ: в)

5. Какое явление является основной физической причиной "частотной стены" (frequency wall), которая остановила экспоненциальный рост тактовых частот процессоров в середине 2000-х годов?

- а) Достижение предела миниатюризации транзисторов.
- б) Слишком высокое энергопотребление и тепловыделение, делающее дальнейшее увеличение частоты экономически и технически нецелесообразным.
- в) Отсутствие программного обеспечения, способного работать на высоких частотах.
- г) Ограничения, накладываемые законом Амдала на одноядерные процессоры.

Правильный ответ: б)

6. Какой тип параллелизма описывает ситуацию, когда одна и та же операция (команда) применяется одновременно к множеству различных элементов данных (например, к разным пикселям изображения)?

- а) Параллелизм на уровне инструкций (ILP).
- б) Параллелизм на уровне потоков (TLP).
- в) Параллелизм на уровне задач (Task-level parallelism).
- г) Параллелизм на уровне данных (DLP).

Правильный ответ: г)

7. Для какого КЛАССА ЗАДАЧ применение специализированных вычислителей, таких как FPGA (программируемые логические интегральные схемы), является НАИБОЛЕЕ предпочтительным по сравнению с универсальными CPU?

- а) Задачи, требующие высокой частоты обновления интерфейса пользователя.
- б) Задачи с жесткими требованиями к детерминизму времени отклика и обработки потоков данных в реальном времени (например, цифровая обработка сигналов).
- в) Задачи, связанные с выполнением сложных ветвящихся алгоритмов с непредсказуемым доступом к памяти.
- г) Задачи, требующие запуска стандартных операционных систем общего назначения.

Правильный ответ: б)

8. Какое из перечисленных требований к бортовым вычислителям летательных аппаратов НАПРЯМУЮ связано с минимизацией их массы и габаритов?

- а) Помехозащищенность.
- б) Ремонтопригодность.
- в) Высокая интеграция и компактность.
- г) Стойкость к вибрационным и ударным нагрузкам.

Правильный ответ: в)

9. Чем ОСНОВНОЕ архитектурное отличие графического процессора (GPU) от центрального (CPU) определяет его превосходство в задачах машинного обучения?

- а) GPU имеют значительно более высокую тактовую частоту, чем CPU.
- б) GPU построены на основе более современных технологических процессов.
- в) Архитектура GPU оптимизирована для большого количества простых арифметико-логических устройств (ALU), работающих параллельно над множеством данных, в то время как CPU имеет несколько сложных ядер, оптимизированных для последовательного выполнения кода.
- г) GPU используют специализированную систему команд, недоступную для CPU.

Правильный ответ: в)

10. Какой принцип лежит в основе работы аналоговых вычислителей и в каких задачах они сохраняют преимущество перед цифровыми?

- а) Принцип дискретных вычислений; преимущество в универсальности.
- б) Принцип моделирования физического процесса с помощью непрерывно изменяющихся величин (напряжения, тока); преимущество в скорости решения определенных дифференциальных уравнений.
- в) Принцип двоичной логики; преимущество в высокой точности вычислений.
- г) Принцип квантовой суперпозиции; преимущество в криптографии.

Правильный ответ: б)

1. **Типовые вопросы открытого типа:**

1. Объясните, в чем заключаются современные ограничения Закона Мура и какие альтернативные пути роста производительности вычислительных систем пришли на смену простому увеличению тактовой частоты.
2. Сравните архитектуры CPU и GPU. Объясните, какие классы задач наиболее эффективно решаются на каждом из них и почему.
3. Что такое "Частотная стена" (Frequency Wall) и каковы были ее основные причины?
4. Опишите, что такое конвейеризация (pipelining) в контексте процессорной архитектуры. Какие преимущества она дает и с какими проблемами может столкнуться разработчик?
5. Объясните, как Закон Амдала ограничивает ускорение параллельной программы. Приведите пример.
6. Назовите ключевые отличия специализированных вычислителей (таких как ASIC, FPGA) от универсальных (CPU). В каких сценариях их применение наиболее оправдано?
7. Что такое "узкое место" (bottleneck) в контексте производительности вычислительной системы? Приведите примеры различных типов "узких мест".

7. Опишите различия между арифметикой с фиксированной (Fixed-Point) и плавающей (Floating-Point) точкой. Каковы компромиссы при выборе того или иного формата для алгоритма ЦОС?
8. Что такое латентность (latency) и пропускная способность (throughput)? Приведите аналогию, объясняющую разницу между ними.
9. Что подразумевается под терминами "верификация" и "валидация" программного обеспечения? В чем между ними разница?
10. Объясните, что такое система контроля версий (например, Git) и какую роль она играет в процессе разработки ПО.
11. В чем заключаются основные преимущества и недостатки реализации алгоритмов на ПЛИС (FPGA) по сравнению с использованием процессоров общего назначения (CPU)?
12. Объясните, какую роль в современных вычислительных системах играют специализированные акселераторы, такие как NPU и TPU.

ПК-2.2 . Применяет современные инструментальные средства при разработке программного обеспечения

1. Чем ОСНОВНОЕ отличие в составе вычислителя автономной роботизированной системы (например, мобильного робота) по сравнению с составом стандартного промышленного компьютера?

- а) Наличие специализированных интерфейсов и модулей для подключения датчиков (лидаров, камер, энкодеров) и исполнительных механизмов (моторов, сервоприводов).
- б) Использование более мощного центрального процессора для сложных вычислений.
- в) Обязательное наличие дискретной видеокарты для обработки графики.
- г) Применение жидкостного охлаждения для стабильной работы.

Правильный ответ: а)

2. Какой компонент является КЛЮЧЕВЫМ и отличает графическую станцию от обычного мощного настольного компьютера, обеспечивая её высокую производительность в задачах 3D-моделирования и рендеринга?

- а) Высокоскоростной SSD-накопитель большого объема.
- б) Большой объем оперативной памяти с ECC (коррекцией ошибок).
- в) Профессиональная видеокарта (GPU), сертифицированная для работы со специализированным ПО (например, NVIDIA Quadro, AMD Radeon Pro).
- г) Многоканальный контроллер Ethernet.

Правильный ответ: в)

3. Для временного хранения данных в порядке "первый вошел – первый вышел" (FIFO) в цифровых схемах используется элемент, известный как:

- а) Стек (LIFO).
- б) Очередь (FIFO).
- в) Мультиплексор.
- г) Декодер.

Правильный ответ: б)

4. Какой тип памяти, часто используемый в составе бортовых вычислителей, характеризуется энергонезависимостью, относительно высоким быстродействием при чтении и используется для хранения программы и константных данных?

- а) DRAM (Dynamic RAM).

- б) SRAM (Static RAM).
- в) FLASH-память.
- г) Регистровая память процессора.

Правильный ответ: в)

5. Какая архитектура является ОСНОВОЙ для построения большинства современных СуперЭВМ?

- а) Однородные вычислительные системы с одним мощным мейнфреймом.
- б) Кластерная архитектура, объединяющая тысячи стандартных вычислительных узлов (часто на базе CPU и GPU) высокоскоростной сетью.
- в) Распределенные системы на основе медленных Ethernet-соединений.
- г) Массивно-параллельные системы с общей памятью для всех процессоров.

Правильный ответ: б)

6. Какое требование является КРИТИЧЕСКИ ВАЖНЫМ для вычислителей в составе оборудования связи (например, маршрутизаторов)?

- а) Возможность апгрейда видеокарты.
- б) Высокая пропускная способность сетевых интерфейсов и низкая латентность при обработке пакетов данных.
- в) Наличие сенсорного экрана для управления.
- г) Поддержка технологии трассировки лучей (Ray Tracing) в реальном времени.

Правильный ответ: б)

7. Какой базовый элемент цифровой схемы позволяет выбрать один из нескольких входных сигналов и направить его на единственный выход?

- а) Демультиплексор (Demux).
- б) Сумматор (Adder).
- в) Шифратор (Encoder).
- г) Мультиплексор (Mux).

Правильный ответ: г)

8. Для обработки видео потока в реальном времени (например, в системах видеонаблюдения) вычислитель должен обладать:

- а) Специализированными аппаратными блоками (например, кодеками) для сжатия и распаковки видео.
- б) Возможностью виртуализации для запуска нескольких операционных систем.
- в) Интерфейсами для подключения устройств виртуальной реальности.
- г) Системой водяного охлаждения для разгона процессора.

Правильный ответ: а)

9. Что такое "латентность" в контексте работы цифрового вычислителя?

- а) Общий объем данных, который можно обработать за единицу времени.
- б) Задержка по времени между началом выполнения операции и моментом получения её результата.
- в) Тактовая частота, на которой работает ядро процессора.

- г) Количество операций с плавающей точкой, выполняемых за секунду.

Правильный ответ: б)

10. Какой компонент типичного промышленного компьютера обеспечивает его устойчивость к вибрациям, повышенной запыленности и широкому температурному диапазону?

- а) Материнская плата форм-фактора ATX.
- б) Усиленный металлический корпус, пассивное охлаждение и компоненты, отобранные для промышленного применения.
- в) Наличие игровой видеокарты.
- г) Использование Wi-Fi модуля вместо проводного Ethernet.

Правильный ответ: б)

Типовые вопросы открытого типа:

1. Опишите типовой workflow разработки и отладки цифрового автомата на ПЛИС: от написания кода на Verilog до программирования кристалла и проверки на стенде. Какие инструменты (САПР) используются на каждом этапе?
2. Что такое статический временной анализ (Static Timing Analysis) и какую роль он играет в процессе проектирования цифровых вычислителей?
3. Объясните, как с помощью тестового окружения (testbench) проводится функциональная верификация кода на Verilog перед синтезом. Что такое временные диаграммы и как вы их анализируете?
4. Какие современные инструменты и методики вы применяли для отладки проектов на ПЛИС? (Например, использование встроенных логических анализаторов - ILA).
5. Опишите, какие инструменты вы использовали для оценки производительности системы на кристалле (SoC), включая взаимодействие процессорного ядра и программируемой логики.
6. На основе каких критериев и с использованием каких инструментов вы выбирали элементную базу (например, конкретную модель ПЛИС или процессора) для проекта бортового вычислителя?
7. Опишите типовой состав и особенности вычислительной системы автономного робота. Какие инструменты используются для разработки и отладки ПО для таких систем?
8. Какие специфические инструменты и средства разработки требуются для создания ПО для графических станций? В чем их отличие от инструментов для встраиваемых систем?
9. Какие современные инструментальные средства используются для разработки ПО для суперкомпьютерных систем? (Например, инструменты для отладки и профилирования параллельных приложений).
10. Опишите процесс разработки программного обеспечения для вычислителей в системах связи. Какие специализированные инструменты и библиотеки могут потребоваться?
11. С какими инструментами и SDK вы работали для разработки под специализированные вычислители обработки видеопотока? (Например, инструменты для работы с кодеками, DMA).
12. Как вы организуете процесс интеграции программного обеспечения с аппаратной частью на примере системы на кристалле (SoC)? Какие инструменты используете для отладки взаимодействия?
13. Как вы обеспечиваете и проверяете соответствие разрабатываемого ПО требованиям надежности и отказоустойчивости для бортовых вычислителей?
14. Какие инструменты статического анализа кода вы считаете наиболее полезными для проектов, связанных с аппаратно-ориентированным программированием, и почему?

Код компетенции	Результаты освоения ОПОП Содержание компетенций
ПК-5	Способен осуществлять программно-аппаратную реализацию алгоритмов цифровой обработки информации

ПК-5.1 . Проектирует и реализует программно-аппаратное описание алгоритмов цифровой обработки информации

1. При программно-аппаратном проектировании алгоритма ЦОС разработчик выбирает между форматами с фиксированной (fixed point) и плавающей (float point) точкой. Какое основное преимущество формата с фиксированной точкой?

- а) Более высокая и предсказуемая точность представления чисел в широком динамическом диапазоне.
- б) Меньшая аппаратная сложность и более высокое быстродействие при реализации в ПЛИС и ASIC.
- в) Автоматическая обработка исключительных ситуаций, таких как переполнение и потеря значимости.
- г) Прямая совместимость с математическим сопроцессором центрального процессора.

Правильный ответ: б)

2. Какая операция является основной (доминирующей) для оценки производительности вычислителей, ориентированных на выполнение задач сверточной нейронной сети?

- а) Операция ветвления (Conditional Branch).
- б) Операция "Умножение с Накоплением" (Multiply-Accumulate, MAC).
- в) Операция чтения/записи в память (Memory Access).
- г) Операция вычисления трансцендентных функций (sin, cos, exp).

Правильный ответ: б)

3. Что из перечисленного НАИБОЛЕЕ вероятно может стать "узким местом" (bottleneck), ограничивающим реальную производительность вычислительной системы при обработке больших данных, несмотря на высокую пиковую производительность её процессора?

- а) Высокая тактовая частота ядра CPU.
- б) Недостаточная пропускная способность (bandwidth) подсистемы памяти.
- в) Большое количество MAC-блоков в составе GPU.
- г) Использование формата данных с плавающей точкой одинарной точности (FP32).

Правильный ответ: б)

4. Пиковая производительность вычислителя измеряется в FLOPS. Что означает эта единица измерения и какой параметр системы она характеризует?

- а) (Frames Per Second) - характеризует скорость обработки кадров видео.
- б) (Floating Point Operations Per Second) - характеризует теоретически максимально возможное количество операций с плавающей точкой, которое процессор может выполнить за секунду.
- в) (Fixed Point Operations Per Second) - характеризует производительность при работе с целыми числами.
- г) (Frequency of Logical Operations Per Second) - характеризует тактовую частоту процессора.

Правильный ответ: б)

5. Какой основной КРИТЕРИЙ будет наиболее важен для сравнительной оценки производительности двух вычислительных систем, если ключевой задачей является inference (вывод) обученной нейронной сети на потоке видео данных в реальном времени?

- а) Пиковая производительность в TFLOPS для FP64.
- б) Средняя задержка (latency) обработки одного кадра от момента поступления до выдачи результата.
- в) Объем оперативной памяти, установленной в системе.
- г) Энергопотребление системы в режиме простоя.

Правильный ответ: б)

6. При реализации алгоритма на fix point арифметике разработчик должен в первую очередь озаботиться:

- а) Выбором оптимального формата числа (количество бит на целую и дробную часть) для предотвращения переполнения и обеспечения необходимой точности.
- б) Настройкой блока управления питанием для снижения энергопотребления.
- в) Оптимизацией кэш-памяти процессора.
- г) Выбором наиболее быстрого типа данных с плавающей точкой.

Правильный ответ: а)

7. Для оценки производительности выполнения сверточных нейронных сетей часто используют метрику:

- а) Килобайт в секунду (KB/s).
- б) Количество гигафлопс на ватт (GFLOPS/W).
- в) Количество кадров в секунду (Frames Per Second, FPS) при заданном разрешении входного изображения.
- г) Задержка доступа к оперативной памяти (RAM Latency) в наносекундах.

Правильный ответ: в)

8. Почему реальная производительность (например, на конкретном алгоритме) всегда ниже пиковой производительности вычислителя?

- а) Из-за необходимости охлаждения процессора.
- б) Из-за невозможности полностью загрузить все вычислительные блоки 100% времени и наличия "узких мест" (память, ввод-вывод).
- в) Из-за использования устаревших версий компиляторов.
- г) Из-за ошибок округления в арифметике с плавающей точкой.

Правильный ответ: б)

9. Какая из перечисленных единиц измерения НАИБОЛЕЕ уместна для оценки производительности специализированного акселератора, предназначенного исключительно для inference нейронных сетей с 8-битной целочисленной арифметикой (int8)?

- а) Гигафлопсы (GFLOPS) для FP32.
- б) Тактовая частота в гигагерцах (GHz).
- в) Тераоперации в секунду (TOPS) для INT8.
- г) Пропускная способность памяти в гигабайтах в секунду (GB/s).

Правильный ответ: в)

10. Методика оценки производительности выполнения сверточных нейронных сетей часто включает расчет:

- а) Общего количества MAC-операций, требуемых для обработки одного входного примера (например, изображения).
- б) Общего количества параметров (весов) в нейронной сети.
- в) Размера входного изображения в мегапикселях.
- г) Объема видеопамяти, требуемого для хранения весов.

Правильный ответ: а)

11. Какой из факторов НАИМЕНЕЕ существенно влияет на реальную производительность вычислителя при обработке потоковых данных в реальном времени?

- а) Задержки при доступе к памяти (Memory Latency).
- б) Пиковая производительность в FLOPS.
- в) Пропускная способность интерфейса ввода-вывода (I/O Bandwidth).
- г) Эффективность управления параллельными потоками исполнения.

Правильный ответ: б)

12. Основным преимуществом арифметики с плавающей точкой (float point) перед фиксированной (fix point) является:

- а) Аппаратная простота и низкое энергопотребление.
- б) Высокая скорость выполнения операций на одной и той же аппаратной платформе.
- в) Широкий динамический диапазон и удобство программирования, так как не требуется ручная масштабирование чисел.
- г) Абсолютная точность представления любых дробных чисел.

Правильный ответ: в)

Типовые вопросы открытого типа:

1. Обоснуйте выбор между арифметикой с фиксированной (fixed point) и плавающей (floating point) точкой для реализации алгоритма цифровой обработки сигналов на ПЛИС. Какие факторы будут ключевыми в этом выборе?
2. Как вы проводите анализ динамического диапазона сигналов в вашем алгоритме для определения оптимальной разрядности fixed point представления?
3. Какие "узкие места" (bottlenecks) наиболее характерны для систем цифровой обработки сигналов и как вы их выявляете на этапе проектирования?
4. Опишите методику оценки производительности выполнения сверточной нейронной сети. Какие параметры сети и вычислителя необходимо учитывать?
5. Объясните, почему операция MAC (Multiply-Accumulate) является ключевой метрикой для оценки производительности в задачах ЦОС и нейронных сетей.
6. Опишите, как вы проектируете конвейерную (pipeline) архитектуру для ресурсоемкого алгоритма ЦОС. Какие факторы учитываете при определении глубины конвейера?
7. Как вы распределяете вычислительную нагрузку между процессорным ядром и программируемой логикой (ПЛИС) при реализации сложного алгоритма обработки данных?
8. Какие методы вы применяете для оптимизации доступа к памяти в аппаратных ускорителях? Как учитываете иерархию памяти при проектировании?

9. Опишите процесс верификации точности вычислений при переходе от программной модели (C/C++) к аппаратной реализации (Verilog/VHDL).
10. Объясните, какие особенности архитектуры вычислителя критичны для достижения высокой производительности в задачах нейронных сетей.
11. Как вы оцениваете требуемую пропускную способность подсистемы памяти для реализации алгоритма с заданными характеристиками (частота, разрядность данных)?
12. Опишите методику анализа производительности системы с учетом ограничений по тепловыделению и энергопотреблению.
13. Какие инструменты и методики вы используете для профилирования производительности на системном уровне (процессор + ПЛИС + память + внешние интерфейсы)?
14. Как вы организуете процесс совместной оптимизации алгоритма и аппаратной архитектуры для достижения целевых показателей производительности?
15. Опишите, как вы проводите анализ компромиссов (trade-off) между точностью вычислений, производительностью и ресурсоемкостью реализации.
16. Как вы оцениваете производительность системы в условиях реальных рабочих нагрузок, отличающихся от идеализированных тестовых сценариев?

ПК-5.2 . Выполняет аргументированных выбор программно-аппаратных средств реализации алгоритмов цифровой обработки информации

1. Какой из перечисленных языков является ЯЗЫКОМ СИНТЕЗА СХЕМ (Hardware Description Language), а не языком программирования общего назначения, и используется для описания аппаратуры на уровне регистровых передач (RTL)?

- а) Python
- б) C++
- в) Verilog
- г) Java

Правильный ответ: в)

2. Какое утверждение наиболее точно описывает основную цель и преимущество использования High-Level Synthesis (HLS) по сравнению с проектированием на языках типа Verilog?

- а) HLS позволяет достичь более высокой тактовой частоты проектируемого устройства.
- б) HLS позволяет описывать поведение алгоритма на языках высокого уровня (например, C/C++) с последующей автоматической генерацией RTL-кода, что ускоряет процесс разработки и верификации.
- в) HLS полностью исключает необходимость понимания основ цифровой схемотехники.
- г) HLS генерирует более оптимальные по площади схемы, чем ручное проектирование.

Правильный ответ: б)

3. При реализации конвейерного (pipeline) алгоритма свертки на C++ для последующего синтеза в HLS, какой подход является КЛЮЧЕВЫМ для повышения производительности?

- а) Максимальное использование рекурсивных функций.
- б) Разбиение вычислений на последовательные этапы (стадии), разделенные регистрами, что позволяет обрабатывать несколько данных одновременно на разных стадиях.
- в) Применение динамического выделения памяти (оператор new).
- г) Использование сложных иерархий наследования классов.

Правильный ответ: б)

4. Какой тип моделирования в IDE для FPGA позволяет проверить функциональность проектируемого устройства ДО его синтеза в конкретную аппаратную конфигурацию, используя специальные тестовые воздействия (testbench)?

- а) Статический временной анализ (Static Timing Analysis).
- б) Функциональная симуляция (Functional Simulation).
- в) Внутрисхемное программирование (In-System Programming).
- г) Анализ использования ресурсов (Resource Utilization Analysis).

Правильный ответ: б)

5. Какой из перечисленных ресурсов НЕ является типичным для современной архитектуры FPGA?

- а) Программируемые логические блоки (Configurable Logic Blocks - CLB).
- б) Блоки памяти (Block RAM - BRAM).
- в) Блоки арифметики с плавающей точкой двойной точности (FPU Double Precision).
- г) Блоки цифровой обработки сигналов (DSP Slices).

Правильный ответ: в)

6. Какой подход к отладке проектов для FPGA позволяет наблюдать за внутренними сигналами устройства в реальном времени, после его программирования в ПЛИС?

- а) Использование отладочных утверждений (assertions) в testbench.
- б) Встроенный логический анализатор (Integrated Logic Analyzer, ILA).
- в) Функциональная симуляция.
- г) Статический временной анализ.

Правильный ответ: б)

7. При исследовании характеристик производительности конвейерной реализации свертки, увеличение тактовой частоты устройства выше определенного порога может привести к:

- а) Автоматическому увеличению пропускной способности конвейера.
- б) Нарушению временных ограничений (setup/hold time) и сбоям в работе устройства.
- в) Увеличению количества стадий конвейера.
- г) Уменьшению занимаемой площади на кристалле.

Правильный ответ: б)

8. Что из перечисленного является ПРЕИМУЩЕСТВОМ использования FPGA для реализации алгоритмов ЦОС по сравнению с выполнением на универсальном CPU?

- а) Простота программирования на языках высокого уровня.
- б) Возможность создания глубоко конвейеризированных и распараллеленных аппаратных структур, оптимальных для конкретного алгоритма.
- в) Более низкая стоимость единичного экземпляра устройства.
- г) Лучшая поддержка стандартных операционных систем и драйверов.

Правильный ответ: б)

9. Какой инструмент в составе САПР для FPGA используется для определения максимальной тактовой частоты, на которой может работать разработанная схема, и проверки соблюдения всех временных ограничений?

- а) Функциональный симулятор (Functional Simulator).
- б) Программа размещения и трассировки (Place & Route).
- в) Статический временной анализатор (Static Timing Analyzer).
- г) Логический синтезатор (Logic Synthesizer).

Правильный ответ: в)

10. При проектировании на Verilog, конструкция always @(posedge clk) описывает:

- а) Комбинационную логику.
- б) Логику, чувствительную к уровню сигнала.
- в) Последовательностную логику, синхронизируемую фронтом тактового импульса.
- г) Асинхронный сброс.

Правильный ответ: в)

11. Для аргументированного выбора между реализацией алгоритма на CPU, GPU или FPGA, разработчик должен в первую очередь проанализировать:

- а) Стоимость Software Developer Kit (SDK).
- б) Структуру алгоритма (наличие параллелизма, требования к задержке, регулярность вычислительных операций) и целевые показатели (производительность, энергоэффективность, стоимость).
- в) Популярность каждого из подходов в интернет-форумах.
- г) Наличие готовых драйверов для операционной системы Windows.

Правильный ответ: б)

Типовые вопросы открытого типа:

1. На основе каких критериев вы выбираете между реализацией алгоритма на CPU, GPU, FPGA или ASIC для конкретной задачи ЦОС? Аргументируйте на примере.
2. Опишите архитектуру современной FPGA. Какие ключевые компоненты (CLB, BRAM, DSP, SerDes) и как влияют на выбор конкретной микросхемы для проекта?
3. В каких случаях использование High-Level Synthesis (HLS) предпочтительнее проектирования на Verilog/VHDL? Какие ограничения HLS вы учитываете?
4. Опишите полный цикл проектирования устройства на FPGA: от создания RTL-кода до программирования кристалла. Какие инструменты и на каких этапах вы используете?
5. Объясните, как вы проводите функциональную верификацию проекта на этапе RTL-моделирования. Что такое testbench и как вы его создаете?
6. Какие методы и инструменты вы применяете для отладки временных характеристик (timing closure) в проектах на FPGA?
7. Опишите принцип конвейерной обработки (pipelining) при реализации алгоритмов на FPGA. Как вы определяете оптимальную глубину конвейера?
8. Объясните, как вы реализуете доступ к внешней памяти в FPGA-проектах для обеспечения максимальной пропускной способности.
9. Как вы организуете взаимодействие между программной частью (процессор) и аппаратным ускорителем (FPGA) в системах на кристалле (SoC)?

10. Опишите методику исследования производительности конвейерной реализации свертки на C++. Какие метрики вы измеряете и как интерпретируете результаты?
11. Какие техники оптимизации вы применяете для увеличения тактовой частоты проекта на FPGA без нарушения временных ограничений?
12. Как вы оцениваете и оптимизируете потребление ресурсов FPGA (логика, память, DSP блоки) при реализации сложных алгоритмов?
13. Опишите ваш подход к отладке FPGA-проектов с использованием встроенных логических анализаторов (ILA/SignalTap).
14. Какие методы вы используете для совместной отладки программного обеспечения и аппаратной части в SoC-системах?
15. Как вы оцениваете производительность системы при реализации свертки на FPGA? Какие параметры критичны для достижения целевых характеристик?
16. Опишите, как вы обеспечиваете переносимость кода между разными семействами FPGA и как учитываете их архитектурные особенности.
17. Какие стратегии тестирования вы применяете для проверки корректности работы FPGA-проекта в реальных условиях?
18. Как вы организуете процесс верификации сложного алгоритма, реализованного средствами HLS?

Код компетенции	Результаты освоения ОПОП Содержание компетенций
ПК-9	**Способен применять языки программирования C/C++ для решения задач в области ИИ**

ПК-9.1 . Разрабатывает и отлаживает эффективные многопоточные решения на C++, тестирует, испытывает и оценивает качество таких решений

1. При реализации Pipeline слоев (Max, Avg Pooling) на C++ для высокопроизводительной системы ИИ, какой подход будет НАИБОЛЕЕ эффективен для увеличения скорости выполнения на современном CPU?

- а) Использование рекурсивных функций для обработки каждого окна
- б) Векторизация вычислений с помощью SIMD инструкций (например, SSE, AVX)
- в) Динамическое выделение памяти для каждого временного буфера
- г) Реализация с помощью виртуальных функций и полиморфизма

Правильный ответ: б)

2. Для обеспечения надежной передачи данных между модулями FPGA, работающими на разных тактовых частотах, следует использовать:

- а) Прямое соединение сигналов без дополнительной синхронизации
- б) Механизм квитирования (handshake) с синхронизаторами
- в) Единую глобальную тактовую частоту для всех модулей
- г) Асинхронные сбросы для всех регистров

Правильный ответ: б)

3. Какой интерфейс передачи данных в FPGA обеспечивает наиболее эффективный механизм для высокоскоростного обмена большими массивами данных между программируемой логикой и процессорной системой?

- а) GPIO (General Purpose Input/Output)
- б) UART (Universal Asynchronous Receiver-Transmitter)
- в) AXI DMA с поддержкой Scatter/Gather
- г) SPI (Serial Peripheral Interface)

Правильный ответ: в)

4. При исследовании производительности реализации Pipeline свертки на Verilog, ключевым показателем эффективности является:

- а) Количество строк кода в проекте
- б) Количество используемых логических элементов и достижимая тактовая частота
- в) Размер откомпилированного файла прошивки
- г) Время компиляции проекта

Правильный ответ: б)

5. Для обработки высокоскоростных последовательных данных в системах компьютерного зрения наиболее подходящим интерфейсом физического уровня является:

- а) LVDS (Low-Voltage Differential Signaling)
- б) CMOS single-ended
- в) TTL (Transistor-Transistor Logic)
- г) RS-232

Правильный ответ: а)

6. При реализации LSTM ячейки на Verilog, основной проблемой проектирования является:

- а) Отсутствие поддержки операций с плавающей точкой в FPGA
- б) Высокая вычислительная сложность и необходимость конвейеризации матричных операций
- в) Невозможность реализации нелинейных функций активации
- г) Ограничения по количеству входных/выходных сигналов

Правильный ответ: б)

7. Какой механизм позволяет эффективно организовать передачу данных между разнородными вычислительными модулями в системе на кристалле (SoC)?

- а) Разделяемая память с программной синхронизацией
- б) AXI шины с различными вариантами (Lite, Stream, Full)
- в) Глобальные переменные в C/C++
- г) Почтовые ящики на основе UART

Правильный ответ: б)

8. При реализации upscaling FIR фильтра на Verilog, основной выигрыш от конвейеризации достигается за счет:

- а) Уменьшения количества используемых DSP блоков
- б) Увеличения максимальной тактовой частоты и пропускной способности
- в) Упрощения отладки проекта
- г) Снижения энергопотребления

Правильный ответ: б)

9. Для каких задач в системах ИИ наиболее целесообразно использовать реализацию на Verilog вместо C++?

- а) Прототипирование алгоритмов машинного обучения
- б) Высокопроизводительная обработка потоковых данных с детерминированной задержкой
- в) Реализация пользовательского интерфейса системы
- г) Работа с файловой системой и базами данных

Правильный ответ: б)

10. Какой подход к синхронизации данных следует использовать при передаче между асинхронными тактовыми доменами?

- а) Использование комбинационной логики
- б) Применение двухступенчатых синхронизаторов
- в) Непосредственное соединение регистров
- г) Использование глобальной асинхронной шины

Правильный ответ: б)

11. Преимущество использования DMA Scatter/Gather в системах ИИ заключается в:

- а) Упрощении программного кода за счет использования библиотек STL
- б) Возможности работы с несмежными областями памяти без вмешательства CPU
- в) Автоматической векторизации вычислений
- г) Динамическом изменении тактовой частоты процессора

Правильный ответ: б)

12. При выборе между C++ и Verilog для реализации компонента системы ИИ, решающим фактором в пользу Verilog является:

- а) Требование к минимальной задержке обработки и предсказуемому времени отклика
- б) Необходимость частых изменений алгоритма в процессе разработки
- в) Простота отладки и тестирования кода
- г) Наличие готовых библиотек машинного обучения

Правильный ответ: а)

Типовые вопросы открытого типа:

1. На основе каких критериев вы выбираете между программной реализацией на C++ и аппаратной на Verilog для различных слоев нейронной сети в системе ИИ?
2. Объясните, как вы реализуете конвейерную обработку (pipeline) для слоев Pooling на C++ для максимизации производительности. Какие техники оптимизации применяете?
3. Опишите методику исследования характеристик производительности pipeline реализации свертки. Какие метрики вы измеряете и как интерпретируете результаты?
4. Как вы обеспечиваете балансировку конвейера при реализации сложных вычислений, таких как LSTM ячейка?
5. Опишите особенности pipeline реализации свертки на Verilog. Как вы организуете параллельные вычисления и управление памятью?

6. Какие архитектурные решения вы применяете при реализации LSTM ячейки на Verilog для обеспечения требуемой производительности?
7. Объясните принцип конвейерной реализации upscaling свертки (FIR фильтра) на Verilog. В чем особенности обработки данных?
8. Объясните проблему Cross Clock Domain и методы ее решения. В каких случаях вы применяете различные механизмы синхронизации?
9. Когда вы выбираете механизм рукопожатия (handshake) для передачи данных между модулями и почему?
10. Сравните различные типы шин AXI (Lite, Stream, Full) от Xilinx. В каких сценариях системы ИИ вы применяете каждый тип?
11. Опишите преимущества использования DMA с Scatter/Gather для задач нейронных сетей. Как это влияет на общую производительность системы?
12. В каких случаях вы применяете LVDS для передачи данных в системах ИИ и какие особенности учитываете при работе с этим интерфейсом?
13. Объясните, как вы организуете сериализацию-десериализацию данных в распределенной системе ИИ. Какие факторы учитываете при выборе формата?
14. Опишите, как вы управляете потоком данных между процессорной системой и программируемой логикой в FPGA-based системах ИИ.
15. Какие методы вы используете для совместной отладки C++ кода и Verilog модулей в системе ИИ?
16. Опишите процесс верификации корректности работы pipeline реализаций как на C++ моделях, так и на Verilog описании.
17. Как вы проводите тестирование производительности всей системы ИИ с учетом взаимодействия всех компонентов?
18. Опишите, как вы обеспечиваете детерминизм временных характеристик в real-time системах ИИ с гетерогенной архитектурой.

ПК-9.2 . Разрабатывает и отлаживает системы ИИ на C++ под конкретные аппаратные платформы с ограничениями по вычислительной мощности, в том числе для встроенных систем

1. Какое утверждение наиболее точно описывает ключевое архитектурное отличие NPU (например, Rockchip RKNN) от GPU в контексте выполнения нейронных сетей?

- а) NPU имеют более высокую тактовую частоту, чем GPU.
- б) NPU специализированы на матрично-векторных операциях и эффективном выполнении типичных для ИИ workload'ов (свертки, активации), в то время как GPU более универсальны.
- в) NPU используют технологию трассировки лучей (ray tracing) для ускорения графики.
- г) Архитектура NPU оптимизирована для выполнения произвольного C++ кода с большим количеством ветвлений.

Правильный ответ: б)

2. Какой инструмент от Xilinx используется для развертывания и выполнения обученных нейронных сетей (например, YOLO) на программируемой логике FPGA платформы Zynq?

- а) Vivado HLS
- б) Vitis AI
- в) Xilinx SDK
- г) Petalinux

Правильный ответ: б)

3. Что такое DPU (Deep Learning Processing Unit) в контексте платформы Xilinx?

- а) Многоядерный центральный процессор общего назначения.
- б) Программируемый логический блок для реализации произвольных цифровых схем.
- в) Конфигурируемое вычислительное ядро, оптимизированное для выполнения сверточных нейронных сетей, которое реализуется в ресурсах FPGA.
- г) Графический процессор для рендеринга 3D-сцен.

Правильный ответ: в)

4. Основное преимущество использования связки "FPGA + DPU" (например, на Zynq-7020) для задач компьютерного зрения по сравнению с выполнением на CPU заключается в:

- а) Более простой отладке программного кода.
- б) Значительно более высокой энергоэффективности и предсказуемой низкой задержке обработки кадра (low latency).
- в) Возможности использования более современных языков программирования.
- г) Автоматическом динамическом изменении тактовой частоты.

Правильный ответ: б)

5. Какой этап работы с нейронной сетью на NPU Rockchip (RKNN) предполагает преобразование модели, обученной в фреймворке (например, TensorFlow, PyTorch), в собственный формат NPU?

- а) Компиляция модели.
- б) Выполнение inference.
- в) Кросс-компиляция C++ кода.
- г) Синтез RTL кода.

Правильный ответ: а)

6. При разработке прикладного C++ приложения, которое использует NPU для выполнения нейронной сети, разработчик взаимодействует с ускорителем через:

- а) Прямое программирование регистров NPU на языке Verilog.
- б) Вызовы функций специализированной программной библиотеки (SDK), предоставляемой производителем NPU (например, RKNN API).
- в) Стандартные инструкции SSE/AVX процессора.
- г) Интерфейс системных вызовов операционной системы.

Правильный ответ: б)

7. Платформа Zynq-7020, часто используемая для прототипирования систем ИИ, объединяет в одной микросхеме:

- а) Два независимых FPGA кристалла.
- б) Процессорные ядра ARM и программируемую логику (FPGA).
- в) NPU и GPU.
- г) Только программируемую логику.

Правильный ответ: б)

8. Основная идея объединения подходов GPU и NPU внутри FPGA заключается в:

- а) Создании универсального процессора для настольных компьютеров.
- б) Возможности создания на одной микросхеме гибкой и переконфигурируемой вычислительной системы, где программируемая логика может реализовать специализированный акселератор (подобный NPU) для конкретной нейронной сети.
- в) Упрощении процесса написания C++ кода.
- г) Эмуляции архитектуры видеокарт NVIDIA.

Правильный ответ: б)

9. Типичный пайплайн работы прикладного C++ решения с YOLO на платформе Zynq с использованием Vitis AI включает:

- а) Захват изображения с камеры на ARM-процессоре -> Передача в DDR-память -> Обработка DPU в FPGA -> Получение результатов bounding boxes на ARM.
- б) Компиляцию C++ кода непосредственно в битстрим для FPGA.
- в) Непосредственный доступ камеры к выводам программируемой логики без участия процессора.
- г) Выполнение всей нейронной сети на процессорных ядрах ARM.

Правильный ответ: а)

10. При отладке C++ приложения, использующего NPU, если inference выполняется некорректно, в первую очередь следует проверить:

- а) Соответствие формата и размера входных данных ожиданиям модели.
- б) Напряжение питания ядра процессора.
- в) Настройки BIOS материнской платы.
- г) Версию используемого веб-браузера.

Правильный ответ: а)

11. Какой фактор является ОСНОВНЫМ ограничением при выборе сложности нейронной сети (например, версии YOLO) для запуска на DPU Zynq-7020?

- а) Объем оперативной памяти DDR, доступный для DPU.
- б) Разрешение монитора, подключенного к системе.
- в) Версия языка C++, поддерживаемая компилятором.
- г) Количество USB-портов на отладочной плате.

Правильный ответ: а)

12. По сравнению с выполнением на CPU, выполнение нейронной сети на NPU или DPU обычно дает наибольший выигрыш в производительности (FPS) для моделей:

- а) С большим количеством последовательных условных переходов (if/else).
- б) С преобладанием матричных операций (сверток), которые могут быть распараллелены.
- в) Которые реализованы с использованием сложных деревьев решений.
- г) Которые работают с данными в формате двойной точности (double precision).

Правильный ответ: б)

Типовые вопросы открытого типа:

1. Опишите архитектурные особенности NPU Rockchip RKNN. В чем его преимущества и ограничения по сравнению с GPU для задач компьютерного зрения?
2. На основе каких критериев вы выбираете между использованием NPU, GPU и FPGA для развертывания моделей ИИ в embedded-системах?
3. Объясните принцип выполнения нейронных моделей на NPU. Как происходит распределение вычислений между ядрами NPU?
4. Опишите процесс подготовки и конвертации модели YOLO для запуска на NPU Rockchip с использованием RKNN Toolkit. С какими типичными проблемами сталкиваетесь?
5. Как вы организуете пред- и постобработку данных в C++ приложении при работе с NPU? Какие особенности передачи данных учитываете?
6. Какие методы оптимизации модели вы применяете для достижения максимальной производительности на NPU Rockchip?
7. Опишите процесс сборки DPU ядра для Zynq-7020. Какие конфигурационные параметры DPU наиболее критичны для производительности?
8. Как вы интегрируете DPU в общую систему на кристалле Zynq? Опишите взаимодействие PS и PL частей.
9. Объясните процесс компиляции модели YOLO с использованием Vitis AI. Какие этапы включает pipeline развертывания?
10. Какие инструменты и методики вы используете для отладки C++ приложений, работающих с NPU/DPU?
11. Как вы проводите профилирование производительности системы ИИ на Zynq? Какие метрики отслеживаете?
12. Опишите методы оптимизации энергопотребления системы при работе с NPU/DPU в embedded-решениях.
13. Как вы организуете конвейер обработки данных в системе на Zynq с использованием DPU? Как распределяете нагрузку между ARM ядрами и программируемой логикой?
14. Какие стратегии тестирования вы применяете для верификации корректности работы модели на NPU/DPU?
15. Как вы решаете проблему несовместимости операторов модели при переносе на целевую платформу?
16. Опишите вашу методику оценки производительности модели YOLO на различных платформах (NPU vs DPU vs CPU).
17. Как вы обеспечиваете детерминизм времени выполнения при работе с NPU в real-time приложениях?
18. Какие методы используете для калибровки квантования модели при подготовке к запуску на NPU/DPU?
19. Опишите подходы к балансировке нагрузки между различными вычислительными элементами в системе ИИ.

ПК-9.2 . Тестирует, испытывает и оценивает качество решений с элементами ИИ, реализованных с использованием языка программирования C/C++

1. Какое утверждение наиболее точно описывает архитектурное отличие GPU от CPU, определяющее их преимущество в задачах машинного обучения?

- а) GPU имеют более высокую тактовую частоту каждого вычислительного ядра.
- б) Архитектура GPU оптимизирована для большого количества простых арифметико-логических устройств (ALU), работающих параллельно над множеством данных (SIMD/SIMT модель).
- в) GPU используют более быструю оперативную память, чем CPU.
- г) GPU лучше справляются с задачами, содержащими много условных переходов и нерегулярный доступ к памяти.

Правильный ответ: б)

2. В модели выполнения OpenCL, что такое "host"?

- а) Устройство, которое выполняет вычислительное ядро (kernel).
- б) Основная программа, работающая на центральном процессоре, которая управляет выполнением кода на устройствах (GPU, CPU и др.).
- в) Специальный тип памяти с произвольным доступом.
- г) Многомерное пространство для организации параллельных вычислений.

Правильный ответ: б)

3. При тестировании производительности кода обработки изображений на OpenCL, обнаружено, что производительность на GPU ниже ожидаемой. Что из перечисленного является НАИБОЛЕЕ вероятной причиной?

- а) Частые передачи небольших объемов данных между хостом и устройством, создающие высокие накладные расходы.
- б) Использование современных оптимизирующих компиляторов C++.
- в) Отсутствие кэширования на CPU.
- г) Использование 64-битной точности (double) для всех вычислений.

Правильный ответ: а)

4. Для чего в OpenCL используются барьеры (barriers) внутри ядра (kernel)?

- а) Для синхронизации выполнения различных ядер.
- б) Для обеспечения того, что все work-items в work-group достигнут определенной точки выполнения, прежде чем любой из них продолжит работу.
- в) Для синхронизации между хост-программой и устройством.
- г) Для автоматической оптимизации использования кэша CPU.

Правильный ответ: б)

5. Какой тип памяти в модели памяти OpenCL является самым быстрым, но разделяется только между work-items внутри одной work-group?

- а) Global memory
- б) Constant memory
- в) Local memory
- г) Private memory

Правильный ответ: в)

6. При реализации сверточного слоя модели YOLO на OpenCL, какой подход к организации данных, скорее всего, обеспечит наилучшую производительность?

- а) Использование атомарных операций для каждого выходного пикселя.
- б) Хранение весов свертки в private memory каждого work-item.
- в) Использование локальной памяти (local memory) для кэширования tile'а входного изображения и весов, к которым обращается work-group.
- г) Последовательная обработка каждого канала изображения в отдельном work-item.

Правильный ответ: в)

7. Что из перечисленного является типичным МЕТРИКОЙ для оценки качества реализации нейронной сети на OpenCL?

- а) Количество строк кода в ядре (kernel).
- б) Время выполнения одного forward pass (инференса) на заданном наборе данных.
- в) Размер исполняемого файла host-программы.
- г) Количество предупреждений (warnings) компилятора.

Правильный ответ: б)

8. Что из перечисленного является типичным МЕТРИКОЙ для оценки качества реализации нейронной сети на OpenCL?

- а) Количество строк кода в ядре (kernel).
- б) Время выполнения одного forward pass (инференса) на заданном наборе данных.
- в) Размер исполняемого файла host-программы.
- г) Количество предупреждений (warnings) компилятора.

Правильный ответ: б)

9. При портировании модели NetworkPrediction landmark с CPU на OpenCL, какой этап является наиболее критичным для обеспечения корректности результатов?

- а) Обеспечение идентичности представления данных с плавающей точкой на CPU и GPU.
- б) Использование самых современных графических драйеров.
- в) Увеличение тактовой частоты GPU.
- г) Запуск host-программы с правами администратора.

Правильный ответ: а)

10. Какой инструмент НАИБОЛЕЕ полезен для профилирования OpenCL кода с целью поиска "узких мест" (bottlenecks)?

- а) Статический анализатор кода C++ (например, cppcheck).
- б) Профайлер, входящий в состав SDK для GPU (например, Nvidia Nsight, AMD CodeXL).
- в) Менеджер задач операционной системы.
- г) Бенчмарк для измерения скорости работы жесткого диска.

Правильный ответ: б)

11. Какое из перечисленных ограничений является типичным при программировании на OpenCL?

- а) Невозможность использования условных операторов (if/else) внутри kernel.
- б) Ограниченный размер локальной памяти (local memory) на work-group.
- в) Обязательность использования только скалярных операций.
- г) Запрет на использование циклов.

Правильный ответ: б)

12. При тестировании реализации на OpenCL обнаружено, что производительность на GPU лишь незначительно превышает производительность на CPU. Какая из перечисленных причин НАИМЕНЕЕ вероятна?

- а) Вычислительное ядро (kernel) является memory-bound, а не compute-bound.
- б) Плохая оптимизация доступа к глобальной памяти (non-coalesced access).
- в) Высокие накладные расходы на передачу данных между хостом и устройством.
- г) Использование устаревшей версии стандарта OpenCL.

Правильный ответ: г)

Типовые вопросы открытого типа:

1. Как архитектурные особенности GPU (SIMD/SIMT, иерархия памяти) влияют на стратегию тестирования и оценки производительности?
4. Опишите модель выполнения OpenCL. Как взаимодействуют хост, платформа, устройство и ядро в процессе вычислений?
5. Объясните иерархию модели памяти в OpenCL. Как особенности каждого типа памяти влияют на производительность и как вы это учитываете при тестировании?
6. Какие методы вы используете для тестирования эффективности передачи данных между хостом и устройством в OpenCL?
7. Опишите методику сравнительного анализа производительности реализации обработки изображений на OpenCL и CPU. Как вы оцениваете влияние кеша CPU?
8. Как вы проводите исследование влияния атомарных операций и барьеров синхронизации на производительность OpenCL кода?
9. Какие методы вы применяете для тестирования эффективности различных стратегий загрузки данных с host на device?
10. Опишите процесс тестирования и отладки реализации модели NetworkPrediction landmark на OpenCL. Какие специфические проблемы характерны для таких задач?
11. Как вы подходите к тестированию отдельных слоев модели YOLO, реализованных на OpenCL? Какие критерии корректности работы вы проверяете?
12. Объясните, как вы тестируете и оптимизируете использование векторных операций в OpenCL для задач обработки данных ИИ.
13. Какие методы тестирования вы применяете для верификации корректности синхронизации work-items в OpenCL?
14. Опишите подход к тестированию производительности при использовании различных паттернов доступа к памяти в OpenCL kernel.
15. Как вы проверяете корректность работы механизмов OpenCL-OpenGL interoperation в графических приложениях с элементами ИИ?
16. Какие методы валидации результатов вычислений на GPU вы используете для обеспечения соответствия эталонной CPU реализации?
17. Опишите процесс тестирования на устойчивость к ошибкам округления и проблемам численной стабильности в GPU-реализациях ИИ.
21. Какие техники тестирования вы применяете для оценки эффективности использования различных типов памяти (global, local, private) в OpenCL kernel?
22. Как вы тестируете масштабируемость OpenCL реализации при работе на различных GPU архитектурах?

Код компетенции	Результаты освоения ОПОП Содержание компетенций
ПК-17	**Способен проводить фронтирные исследования в области управления, решения, агентных и мультиагентных систем**

ПК-17.1. Исследует и создает агентные системы

1. Какой из перечисленных методов сжатия нейронных сетей предполагает замену весов с плавающей точкой (FP32) на целочисленные представления с меньшей разрядностью (например, INT8) с целью уменьшения размера модели и увеличения скорости inference?

- а) Обрезка (Pruning)

- б) Квантизация (Quantization)
- в) Дистилляция знаний (Teacher-Student)
- г) Низкоуровневая оптимизация

Правильный ответ: б)

2. Основная идея конвейерного (pipeline) подхода при построении вычислительной архитектуры для нейросетей заключается в:

- а) Увеличении тактовой частоты каждого вычислительного блока.
- б) Разбиении вычислений на последовательные этапы (стадии), что позволяет обрабатывать несколько данных одновременно, повышая общую пропускную способность (throughput).
- в) Уменьшении разрядности обрабатываемых данных.
- г) Динамическом изменении структуры сети в процессе выполнения.

Правильный ответ: б)

3. Какой из перечисленных методов сжатия предполагает обучение компактной модели (студент) так, чтобы она имитировала поведение большой и точной модели (учитель), используя как правильные ответы, так и "мягкие" метки (soft labels) от учителя?

- а) Квантизация
- б) Обрезка весов
- в) Дистилляция знаний (Knowledge Distillation)
- г) Векторизация инструкций

Правильный ответ: в)

4. При низкоуровневой конвейерной реализации нелинейной функции активации (например, гиперболического тангенса) на аппаратном уровне (ПЛИС), основной проблемой является:

- а) Необходимость аппроксимации сложной математической функции с помощью конвейера простых операций (сдвиги, сложения, умножения).
- б) Отсутствие аппаратной поддержки операций умножения.
- в) Невозможность использования конвейерной архитектуры для таких задач.
- г) Слишком высокая тактовая частота, требуемая для вычислений.

Правильный ответ: а)

5. Метод обрезки (pruning) нейронной сети направлен на:

- а) Увеличение количества слоев в сети.
- б) Удаление наименее значимых весов или нейронов для создания разреженной модели.
- в) Замену всех весов на случайные значения.
- г) Увеличение разрядности весов для повышения точности.

Правильный ответ: б)

6. Какой аппаратный механизм синхронизации наиболее эффективен для управления доступом к разделяемому ресурсу (например, блоку памяти) со стороны нескольких параллельных конвейерных вычислительных ядер?

- а) Программные циклы ожидания (spinlock).
- б) Аппаратные семафоры, реализованные на триггерах и логических элементах.
- в) Условные операторы в коде C++.

- г) Увеличение тактовой частоты.

Правильный ответ: б)

7. При потоковой обработке данных в конвейерной архитектуре, основным требованием для достижения максимальной производительности является:

- а) Непрерывная подача данных на вход конвейера и отсутствие "пузырей" (stall) в его работе.
- б) Использование самой медленной тактовой частоты.
- в) Максимальное усложнение каждой стадии конвейера.
- г) Случайный порядок обработки данных.

Правильный ответ: а)

8. Какая проблема многопоточности возникает, когда несколько конвейерных ядер пытаются одновременно записать данные в одну и ту же ячейку памяти?

- а) "Состояние гонки" (Race Condition).
- б) Инверсия приоритетов.
- в) Фрагментация памяти.
- г) Утечка памяти.

Правильный ответ: а)

9. После применения агрессивной обрезки (pruning) нейронной сети, стандартной практикой является:

- а) Немедленное развертывание модели в продакшен.
- б) Дообучение (fine-tuning) обрезанной модели для восстановления точности.
- в) Удаление оставшихся весов.
- г) Увеличение learning rate до максимального значения.

Правильный ответ: б)

10. Для чего в конвейерной архитектуре используются аппаратные барьеры (barriers)?

- а) Для синхронизации работы различных стадий конвейера, обеспечивая корректную передачу данных между ними.
- б) Для физического разделения кристалла процессора.
- в) Для уменьшения энергопотребления системы.
- г) Для увеличения разрядности обрабатываемых данных.

Правильный ответ: а)

11. Низкоуровневая конвейерная реализация ячейки LSTM на аппаратном уровне является сложной задачей прежде всего из-за:

- а) Наличия нескольких независимых гейтов (input, forget, output), требующих параллельных матричных умножений и нелинейных преобразований.
- б) Отсутствия операций сложения в аппаратных акселераторах.
- в) Необходимости использования только 64-битной арифметики.
- г) Простоты математического аппарата LSTM.

Правильный ответ: а)

12. Какой из методов сжатия моделей может быть применен на этапе inference без изменения процесса обучения исходной модели?

- а) Дистилляция знаний (требует переобучения студенческой модели)
- б) Пост-тренировочная квантизация (Post-training quantization)
- в) Обучение с подкреплением
- г) Накопительное обучение

Правильный ответ: б)

Типовые вопросы открытого типа:

1. Опишите процесс обрезки (pruning) нейронной сети от выбора метрики значимости весов до финального fine-tuning. Какие критерии вы используете для определения "важных" весов?
2. Объясните метод дистилляции знаний (Knowledge Distillation). Как вы выбираете архитектуру студенческой модели и температуру для функции потерь?
3. Какие метрики вы используете для оценки эффективности сжатия модели? Как вы балансируете между степенью сжатия и падением точности?
4. Опишите принцип конвейерной обработки (pipelining) применительно к выполнению нейронных сетей. Как конвейеризация помогает скрыть латентность операций?
5. Как особенности потоковой обработки данных влияют на архитектурные решения при реализации ускорителей для нейросетей?
6. Объясните, как вы проектируете конвейерную реализацию LSTM ячейки. Какие стадии конвейера выделяете и как обеспечиваете их балансировку?
7. Опишите низкоуровневую конвейерную реализацию сверточного слоя. Как вы организуете параллельную обработку карт признаков и ядер свертки?
8. Какие подходы вы используете для аппаратной реализации нелинейных функций активации (сигмоид, гиперболический тангенс) с сохранением точности?
9. Объясните, как вы реализуете конвейерный FIR фильтр для обработки временных рядов в нейросетевых приложениях.
10. Опишите проблемы многопоточности, характерные для параллельного выполнения нейросетевых вычислений. Как вы их диагностируете и решаете?
11. Объясните разницу между барьерами (barriers) и семафорами (semaphores) в контексте синхронизации потоков в нейросетевых ускорителях.
12. Как вы реализуете аппаратные механизмы синхронизации для управления доступом к разделяемым ресурсам в многопоточных вычислителях?
13. Как методы сжатия нейросетей влияют на архитектурные решения при проектировании специализированных вычислителей?
14. Опишите, как совместное применение квантизации и обрезки позволяет достичь синергетического эффекта в сжатии моделей.
15. Какие особенности аппаратной реализации вы учитываете при применении различных методов сжатия к нейросетевым моделям?

ПК-17.2. Исследует и создает мультиагентные системы

1. Какое из перечисленных требований является КРИТИЧЕСКИ ВАЖНЫМ для большинства встраиваемых систем с нейросетевыми алгоритмами, но обычно не является столь жестким для серверных решений?

- а) Максимальная производительность в TFLOPS
- б) Минимизация энергопотребления
- в) Поддержка виртуализации
- г) Возможность горячей замены компонентов

Правильный ответ: б)

2. Какой инструмент используется для преобразования моделей нейронных сетей, обученных в фреймворках типа TensorFlow/PyTorch, в формат, исполняемый на NPU Rockchip (RKNN)?

- а) rknn-toolkit2
- б) OpenVINO Toolkit
- в) TensorRT
- г) ONNX Runtime

Правильный ответ: а)

3. Какой из перечисленных подходов к оптимизации нейронной сети для встраиваемых систем позволяет значительно уменьшить объем памяти, требуемый для хранения весов модели, и увеличить скорость inference?

- а) Квантование весов (например, переход от FP32 к INT8)
- б) Увеличение количества слоев в сети
- в) Использование более сложных функций активации
- г) Добавление dropout-слоев

Правильный ответ: а)

4. Что означает концепция "разделение на soft и hard обработку слоев сети" при оптимизации нейросети для встраиваемых систем?

- а) Выполнение простых, регулярных слоев (свертки) на специализированном аппаратном ускорителе (NPU - hard), а сложных, нерегулярных - на CPU (soft)
- б) Использование разных learning rate для разных слоев
- в) Применение мягких и жестких функций активации
- г) Разделение данных обучения на простые и сложные примеры

Правильный ответ: а)

5. Основное преимущество использования SoM (System-on-Module) подхода при разработке встраиваемых систем с нейросетевыми возможностями заключается в:

- а) Упрощении процесса разработки за счет использования предварительно отлаженного вычислительного модуля
- б) Возможности самостоятельного проектирования процессорного ядра
- в) Отсутствии любых ограничений по энергопотреблению
- г) Невозможности кастомизации периферийных интерфейсов

Правильный ответ: а)

6. При подготовке модели YOLO для запуска на NPU RKNN, какой этап является необходимым для достижения максимальной производительности?

- а) Компиляция модели с указанием целевой платформы и оптимизаций
- б) Ручное переписывание модели на языке C++
- в) Увеличение размера входного изображения
- г) Добавление не поддерживаемых NPU операций

Правильный ответ: а)

7. Какая архитектура процессоров доминирует на рынке встраиваемых систем с требованиями к энергоэффективности?

- а) ARM
- б) x86
- в) PowerPC
- г) MIPS

Правильный ответ: а)

8. Что из перечисленного является характерной ОСОБЕННОСТЬЮ встраиваемых систем с нейросетевыми алгоритмами?

- а) Жесткие требования к детерминизму времени отклика (работа в реальном времени)
- б) Неограниченные ресурсы памяти и вычислений
- в) Возможность постоянного подключения к облачным сервисам
- г) Отсутствие ограничений по массогабаритным показателям

Правильный ответ: а)

9. Какой из этапов работы rknn-toolkit2 позволяет оценить производительность и точность модели до ее развертывания на целевом устройстве?

- а) Запуск инференса на симуляторе (simulator)
- б) Просмотр структуры модели
- в) Экспорт модели в формат ONNX
- г) Создание документации по модели

Правильный ответ: а)

10. При оптимизации нейросети для встраиваемой системы, концепция "миниатюризация систем" подразумевает:

- а) Сокращение физических размеров вычислительного устройства при сохранении функциональности
- б) Увеличение количества параметров модели
- в) Использование только облачных вычислений
- г) Добавление дополнительных периферийных устройств

Правильный ответ: а)

11. Какое преимущество дает использование RISC-V архитектуры в перспективных встраиваемых системах с ИИ?

- а) Открытая архитектура, позволяющая кастомизацию под конкретные задачи
- б) 100% совместимость с программным обеспечением x86
- в) Наибольшая производительность в серверных приложениях
- г) Автоматическая оптимизация нейросетевых моделей

Правильный ответ: а)

12. При переносе модели YOLO на NPU RKNN, с какой типичной проблемой может столкнуться разработчик?

- а) Не все операции исходной модели могут поддерживаться аппаратным ускорителем

- б) NPU требует наличия дискретной видеокарты
- в) Модель автоматически становится более точной
- г) Скорость инференса всегда уменьшается

Правильный ответ: а)

Типовые вопросы открытого типа:

1. Какие специфические требования встраиваемых систем оказывают наибольшее влияние на оптимизацию нейросетевых вычислений?
2. Опишите стратегии минимизации энергопотребления нейросетевых алгоритмов при работе в реальном времени.
3. Как требования детерминированного времени отклика влияют на выбор архитектуры нейросети и методов оптимизации?
4. Сравните подходы к оптимизации нейросетей для ARM и RISC-V архитектур. В чем ключевые различия?
5. Какие преимущества и ограничения SoM-решений для развертывания нейросетевых моделей в embedded-системах?
6. Как вы распределяете вычислительную нагрузку между различными ядрами в гетерогенной SoC-архитектуре?
7. Опишите процесс интеграции NPU/DPU ядер в общую вычислительную систему. Какие аспекты наиболее критичны для производительности?
8. Как вы проводите анализ целесообразности использования специализированных ускорителей для конкретной нейросетевой задачи?
9. Какие методы оптимизации вы применяете для эффективного использования памяти в системах с NPU?
10. Опишите полный пайплайн подготовки модели YOLO для запуска на NPU Rockchip с использованием rknn-toolkit2.
11. С какими типичными проблемами совместимости операторов сталкиваетесь при конвертации моделей в RKNN-формат?
12. Как вы оптимизируете модель для достижения максимального FPS на конкретном NPU Rockchip?
13. Какие инструменты и методики вы используете для отладки производительности нейросетевых вычислений на embedded-системах?
14. Как вы обеспечиваете воспроизводимость результатов оптимизации на различных экземплярах устройств?
15. Какие новые подходы к оптимизации нейросетевых вычислений для embedded-систем вы считаете наиболее перспективными?
16. Как развитие аппаратных ускорителей влияет на методы оптимизации программного обеспечения?

Типовые теоретические вопросы для экзамена по дисциплине

1. Сравните архитектуры CPU, GPU и NPU. Объясните, какие классы задач нейронных сетей наиболее эффективно решаются на каждом из них, и приведите примеры конкретных операций.
2. Опишите полный цикл проектирования конвейерного вычислителя для операции свертки на ПЛИС - от высокоуровневого описания до низкоуровневой реализации и верификации.
3. Что такое "Закон Амдала" и "Частотная стена"? Как эти концепции повлияли на развитие современных вычислительных систем?

4. Объясните проблему "узких мест" (bottleneck) в контексте гетерогенной вычислительной системы. Опишите методику выявления такого "узкого места".
5. Каков типовой состав и ключевые требования к бортовому вычислителю автономного робота? Как это влияет на выбор элементной базы?
6. Обоснуйте выбор между арифметикой с фиксированной и плавающей точкой для реализации CNN на ПЛИС. Опишите методику перевода модели.
7. Опишите принцип организации вычислений в OpenCL. Объясните взаимодействие: хост, устройство, ядро, work-group, work-item.
8. Сравните подходы к реализации на ПЛИС: Verilog/VHDL vs HLS. В каких сценариях каждый предпочтительнее?
9. Опишите процесс подготовки и запуска модели YOLO на NPU Rockchip с использованием rknn-toolkit2.
10. Что такое "разделение на soft и hard обработку" нейронной сети? Приведите пример для архитектуры YOLO.
11. Сравните три метода сжатия нейросетей: квантование, обрезка и дистилляция знаний. Как они влияют на производительность?
12. Объясните, как сжатие нейросетей влияет на архитектурные решения при проектировании специализированных вычислителей.
13. Опишите процесс обрезки (pruning) нейронной сети. Почему этап дообучения критически важен?
14. Что такое "квантование-осознанное обучение" (QAT)? Чем оно отличается от посттренировочного квантования?
15. Какова роль DMA в гетерогенных системах? Объясните на примере Zynq.
16. Опишите типовой workflow отладки проекта на ПЛИС - от симуляции до ILA.
17. Какие инструменты вы используете для профилирования C++ кода в embedded-системах?
18. Опишите методику сравнительного анализа производительности модели на CPU, GPU и NPU.
19. Что такое "конвейеризация" и как она применяется для ускорения вычислений на разных уровнях?
20. Опишите принципы синхронизации в гетерогенных системах. Разница между барьерами, семафорами и мьютексами.
21. Опишите процесс сборки и интеграции DPU ядра Xilinx в проект для Zynq-7020.
22. Объясните, как реализовать конвейерную реализацию функции активации на Verilog.
23. Каковы ключевые требования к вычислителям для обработки видеопотока в реальном времени?
24. Опишите организацию обмена данными между асинхронными тактовыми доменами внутри ПЛИС.
25. В чем преимущества и недостатки RISC-V против ARM для встраиваемых систем с ИИ?
26. Разработайте систему компьютерного зрения для беспилотного аппарата. Опишите выбор платформы, оптимизацию модели и тестирование.
27. Проанализируйте производительность системы с NPU. Почему реальная производительность ниже пиковой?
28. Спроектируйте конвейерный accelerator для LSTM на ПЛИС. Какие проблемы ожидаете?
29. Сравните инструментальные цепочки для разработки под NPU Rockchip и DPU Xilinx.
30. Оптимизируйте ResNet-50 для embedded-устройства. Какие методы сжатия примените и в какой последовательности?