Ответы на контрольные вопросы

1. Различие между последовательной и параллельной сортировкой слиянием

Последовательная сортировка выполняется одним потоком и обрабатывает массив шаг за шагом. Параллельная версия делит массив на части и сортирует их одновременно в нескольких потоках. Это позволяет сократить общее время выполнения при больших объемах данных.

2. Влияние потоков и блоков на производительность

Количество потоков и блоков определяет сколько элементов обрабатываются одновременно. Если потоков слишком мало GPU используется не полностью. Если слишком много возникает лишняя нагрузка и потери на синхронизацию.

3. Сложности быстрой сортировки на GPU

Основная сложность связана с рекурсией и зависимостями между элементами. GPU плохо подходит для глубокой рекурсии. Поэтому алгоритм приходится упрощать или ограничивать.

4. Когда GPU может быть медленнее CPU

Для маленьких массивов накладные расходы на копирование данных и запуск kernel превышают выигрыш от параллелизма. В таких случаях CPU работает быстрее.

5. Почему важен выбор размера блоков

Размер блока влияет на загрузку вычислительных ресурсов GPU. Неправильный выбор приводит к простаиванию потоков или перерасходу памяти.

6. Влияние разделяемой памяти

Разделяемая память быстрее глобальной. Ее использование уменьшает задержки при доступе к данным и ускоряет сортировку внутри блока.

7. Принцип разделяй и властвуй

Алгоритм разбивает задачу на меньшие части. Каждая часть решается отдельно. После этого результаты объединяются в итоговое решение.