```
!nvcc task1_reduction.cu -o task1
!./task1
```

```
TASK 1 — Reduction
CPU sum = 511.593
GPU sum = 0
Diff    = 511.593
```

```
[3]
3 сек.
!nvcc -O3 -std=c++17 task2_scan.cu -o task2
!./task2
```

```
TASK 2 — Prefix sum
Last element = 0 (expected 1024)
```

```
[4]
2 сек.
!nvcc -O3 -std=c++17 task3_benchmark.cu -o task3
!./task3
```

```
TASK 3 — Benchmark
GPU time: 7.2664 ms
CPU time: 0.00301 ms
```

Контрольные вопросы

_____

1. В чём разница между редукцией и сканированием?

Редукция — это операция сведения массива к одному значению

(сумма, минимум, максимум).

Сканирование (prefix sum) — это операция, при которой вычисляется

накопленный результат для каждого элемента.

Пример:

[1, 2, 3, 4]

Редукция → 10

Сканирование → [1, 3, 6, 10]

_____

2. Какие типы памяти CUDA используются для оптимизации?

- Глобальная память — хранение входных и выходных данных
- Разделяемая память (shared) — ускорение доступа внутри блока
- Регистры — локальные переменные потоков

В данной работе ключевую роль играет shared memory.

---

3. Как можно оптимизировать префиксную сумму на GPU?

- Использовать Blelloch Scan (upsweep + downsweep)
- Минимизировать __syncthreads()
- Выполнять scan иерархически (block → grid)
- Использовать warp-level primitives (__shfl_*)

---

4. Приведите пример задачи, где применяется сканирование

- Построение гистограмм
- Stream compaction
- Сортировка (Radix Sort)
- Генерация индексов
- Алгоритмы графов (BFS)