CUDA: Reduction

План семинара

1. Задача Reduction. Оптимизация
2. Задача Scan. Оптимизация

Код этого семинара

Reduction (вычисление суммы всех элементов)

Scan (вычисление префиксной суммы)

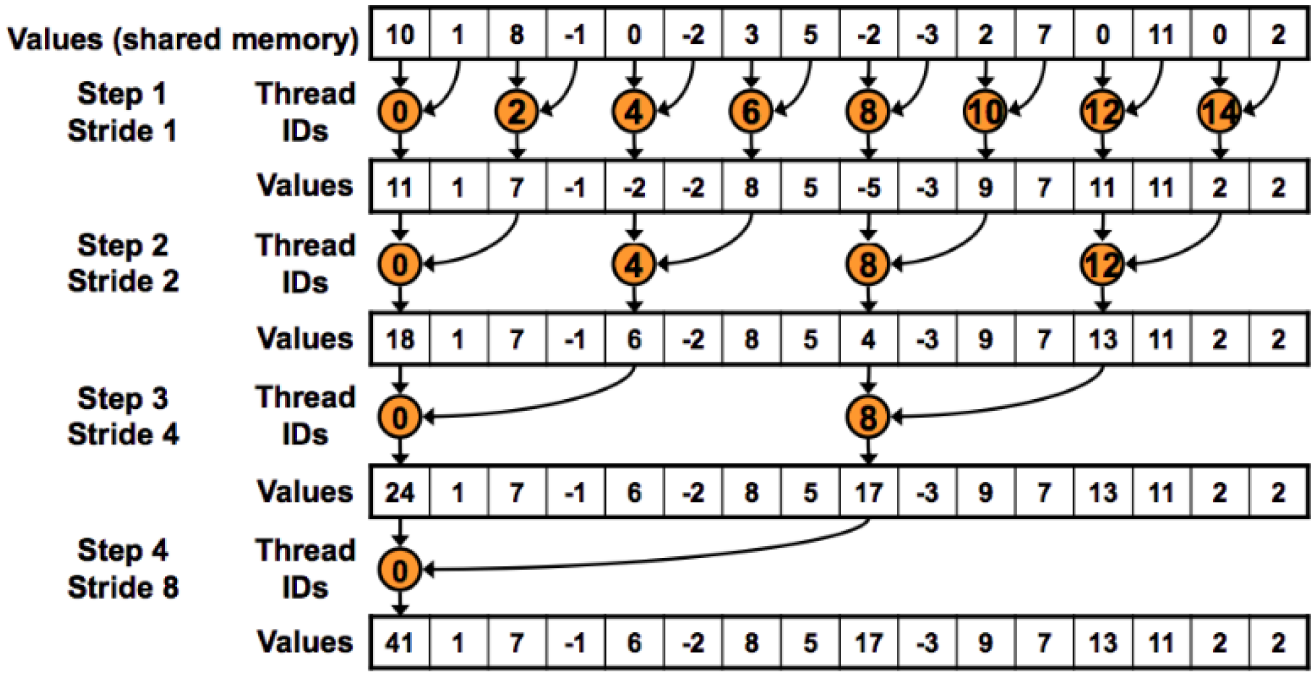
Reduction

Задача: Найти сумму всех элементов массива.

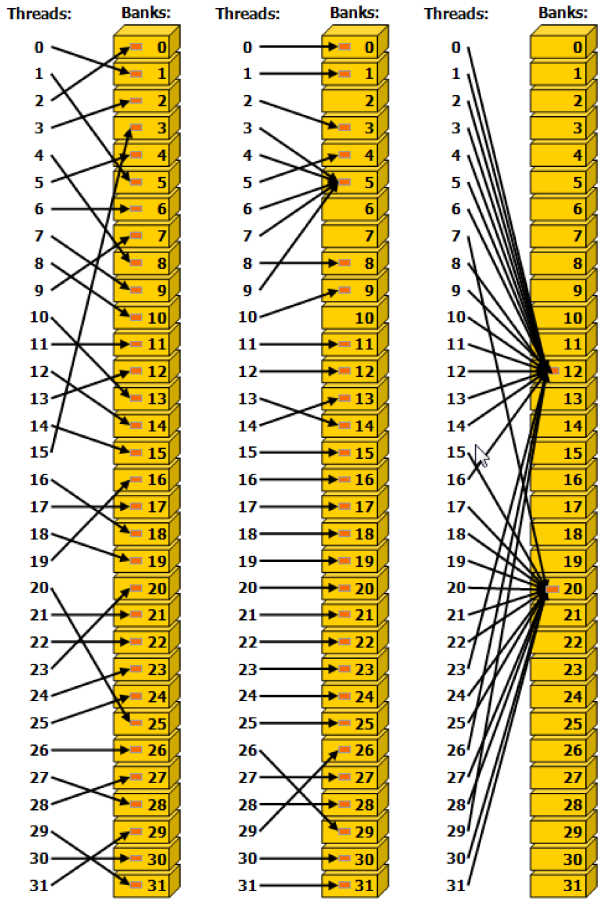
Reduction

Задача: Найти сумму всех элементов массива.

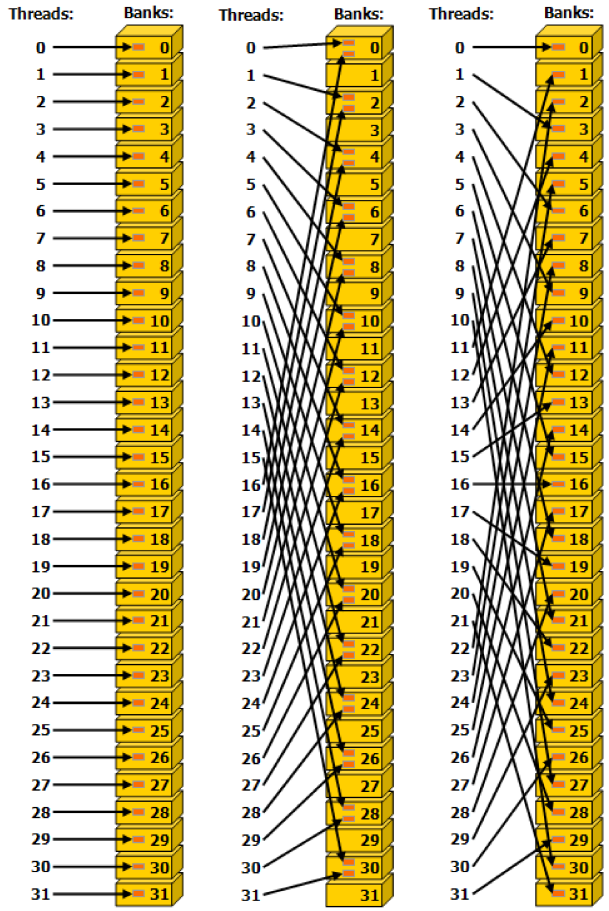
Решение:



Reduction: shared память



Плохо



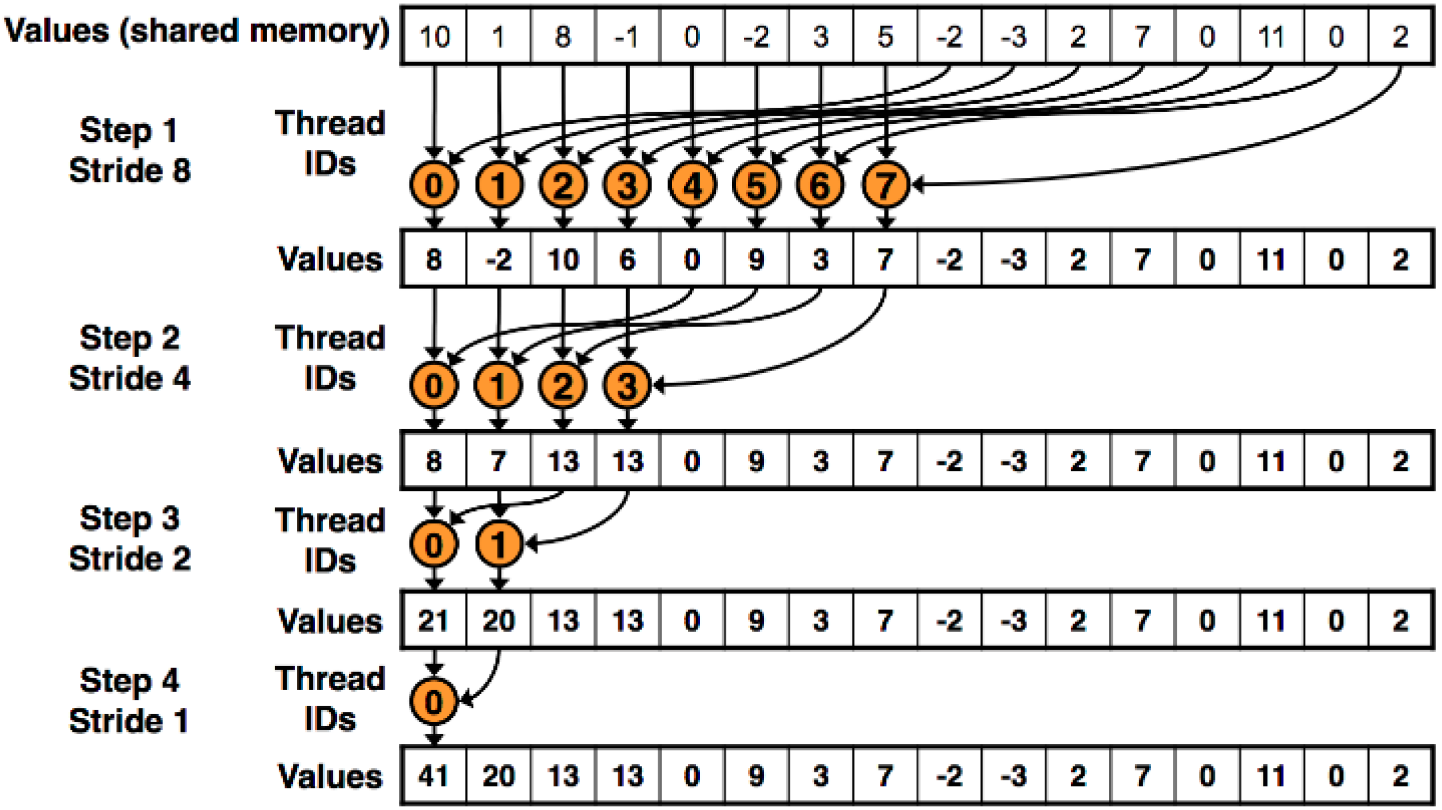
Хорошо

Пропускная способность одной ячейки shared памяти — **32 бита за инструкцию**.

Reduction

Задача: Найти сумму всех элементов массива.

Решение **без bank conflicts**:



Reduction

Микрооптимизации:

- Инструкции для shuffle данных внутри warp
  https://docs.nvidia.com/cuda/cuda-c-programming-guide/index.html#warp-shuffle-functions
- Сложение элементов при записи в shared память
- Развёрнутый цикл для повышения эффективности *скомпилированного* кода

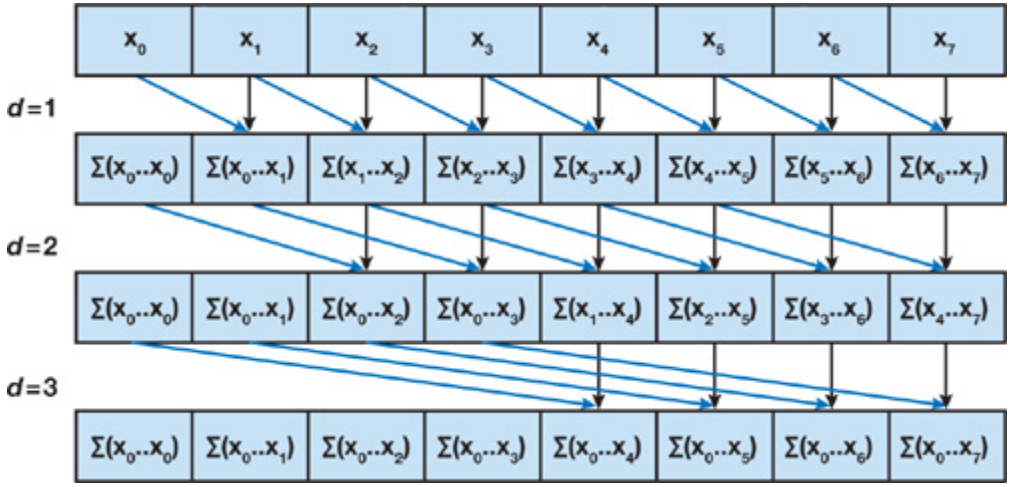
Scan

Задача: Подсчитать все префиксные суммы массива

Scan

Задача: Подсчитать все префиксные суммы массива

Наивное решение:

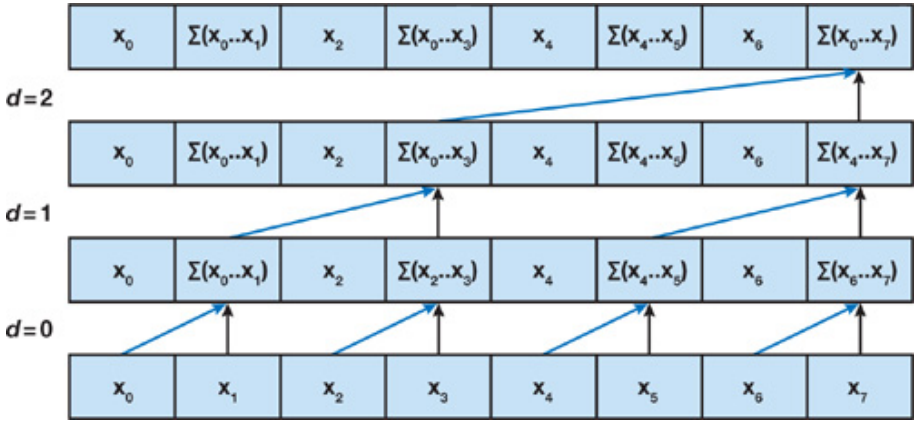


Scan

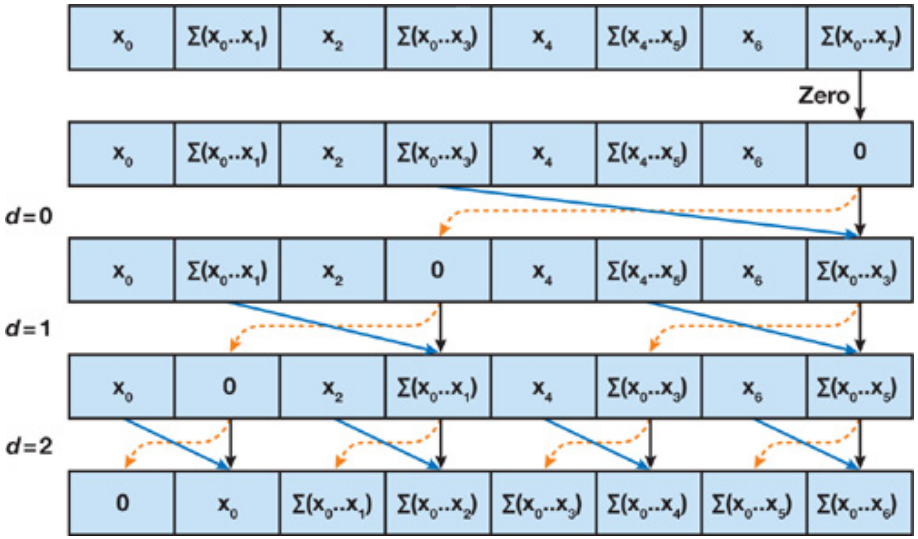
Задача: Подсчитать все префиксные суммы массива

Эффективное решение: **Дерево Фенвика**

Первый этап



Второй этап



Полезные материалы

CUDA GPU gems
https://developer.nvidia.com/gpugems/gpugems3/contributors

CUDA GPU gems: Префиксные суммы
https://developer.nvidia.com/gpugems/gpugems3/part-vi-gpu-computing/chapter-39-parallel-prefix-sum-scan-cuda