Казанский федеральный университет

ВЫСШАЯ ШКОЛА информационных технологий и информационных систем

Оптимизация алгоритмов

Произведение матриц

- Пример: оптимизация перемножения матриц
- Одномерная область вычислений размером n
- Каждый рабочий элемент обсчитывает свою строку матрицы C
- Использование частной памяти – массив размером n для хранения строки матрицы A
- Использование локальной памяти – копирование столбца матрицы B в локальную память

Произведение матриц

- Пример: оптимизация перемножения матриц
- Двумерная область вычислений
- Блочное произведение матриц по аналогии с CUDA-версией

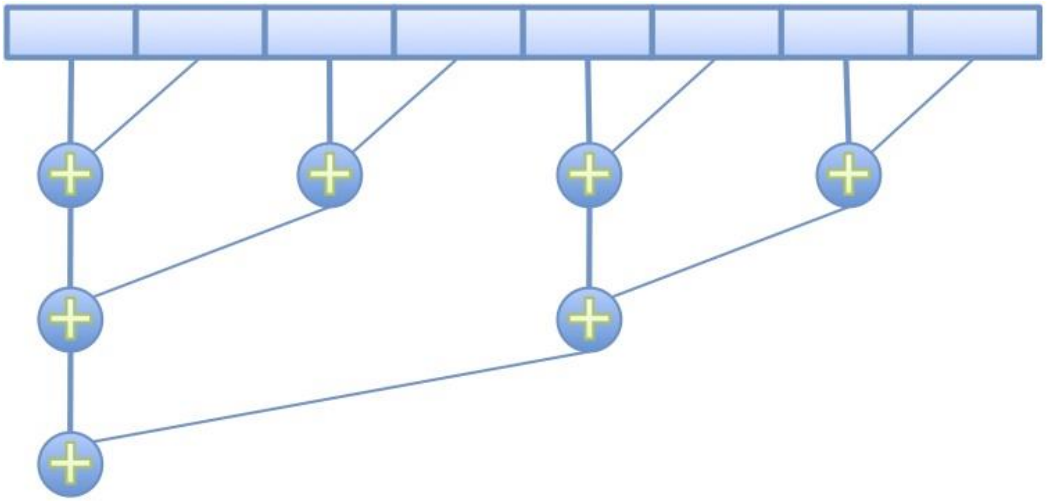
Редукция на OpenCL

- Классический алгоритм для демонстрации техник оптимизации
- Простая последовательная реализация O(n)
- Параллельный алгоритм нетривиален
- Параллельные версии сильно различаются по производительности

Наивная реализация

- Каждая рабочая группа вычисляет свою частичную сумму
- Условие `(local_index & mask) == 0` проверяет делимость на 2/4/8 etc.
- Частичные суммы обрабатываются на процессоре
- Недостаток – плохое дерево редукции
  - На каждом этапе рабочая группа становится все более разреженной

Наивная реализация



Parallel Reduction Tree for Associative Operator

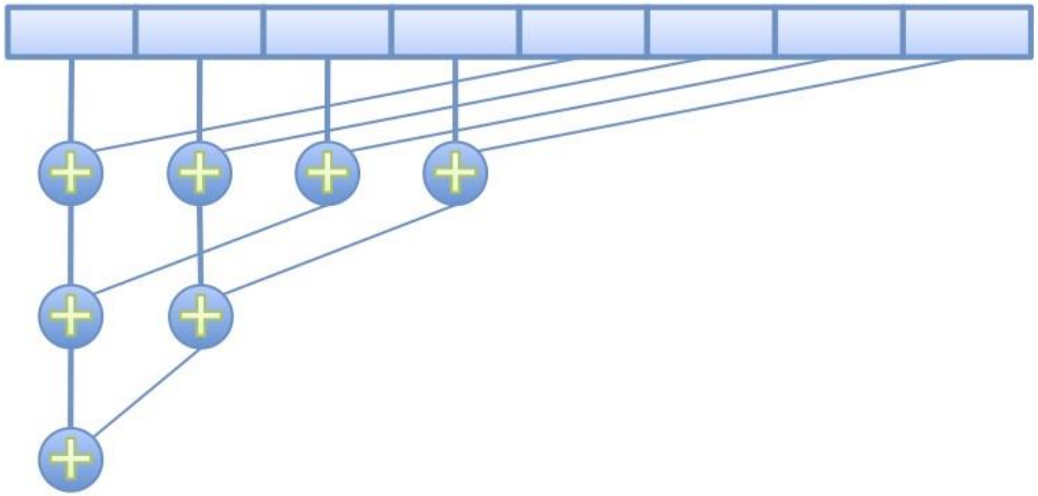
SIMD Utilization for Reduction Tree

http://developer.amd.com

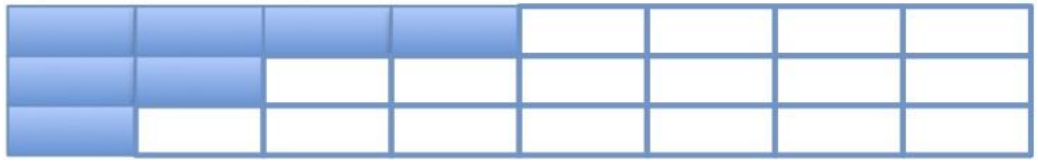
Реорганизация дерева

- Будем не увеличивать а уменьшать расстояние между элементами
- Это позволит эффективнее нагружать вычислительные ресурсы
- Количество элементов в 2 раза меньше *n*
- Это позволит сложить 2 половины массива при загрузке в локальную память

Реорганизация дерева



Parallel Reduction Tree for Commutative Operator



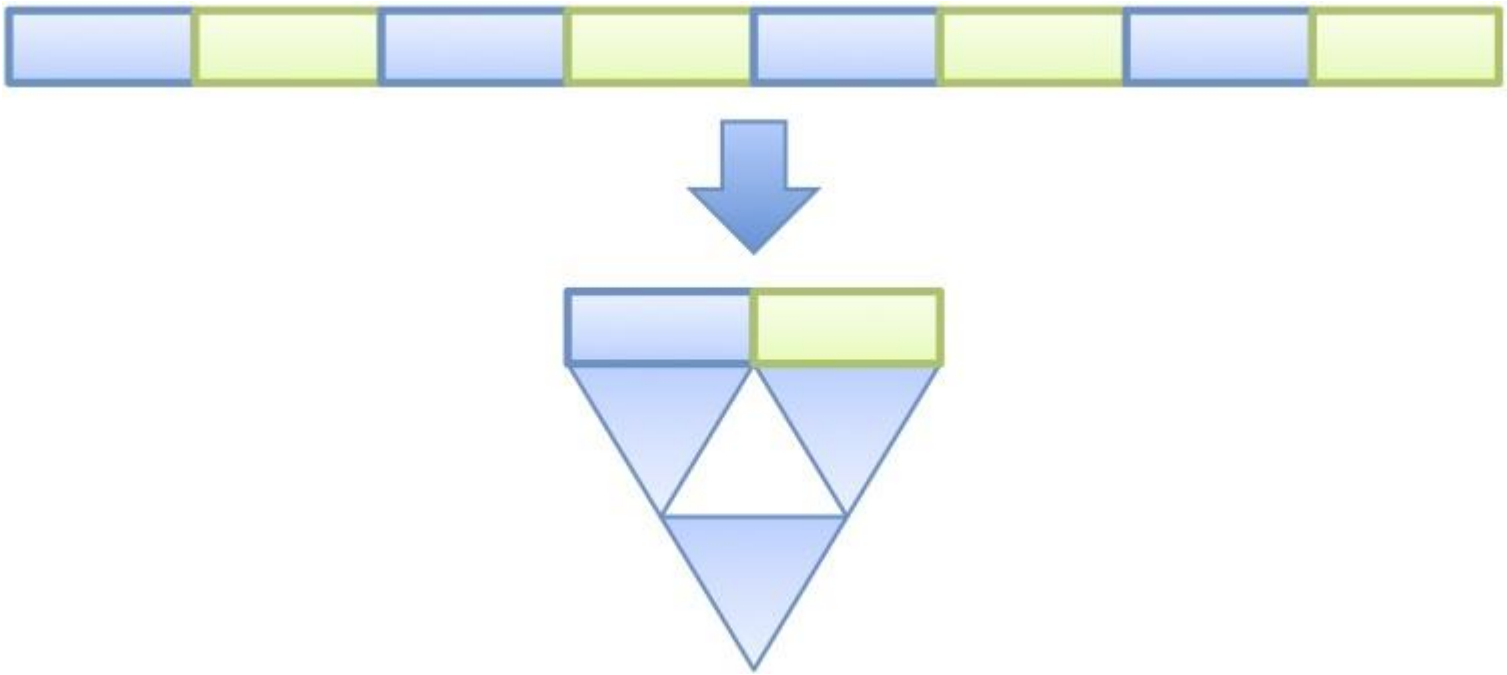
SIMD Utilization for Reduction Tree

http://developer.amd.com

Двухэтапная редукция

- Развитие идеи со сложением элементов при загрузке
- Количество элементов кратно меньше *n*
- Последовательный цикл для каждого элемента по массиву
- Параллельная редукция по группам

Двухэтапная редукция



http://developer.amd.com

Казанский федеральный
УНИВЕРСИТЕТ
ВЫСШАЯ ШКОЛА
информационных технологий
и информационных систем

Задание на практику

- Оптимизация умножение матриц – умножение на транспонированную B (1-d)