Practice 5: Reduce

Objective: To understand how to implement a basic parallel algorithm called "Reduce."

Reduce:

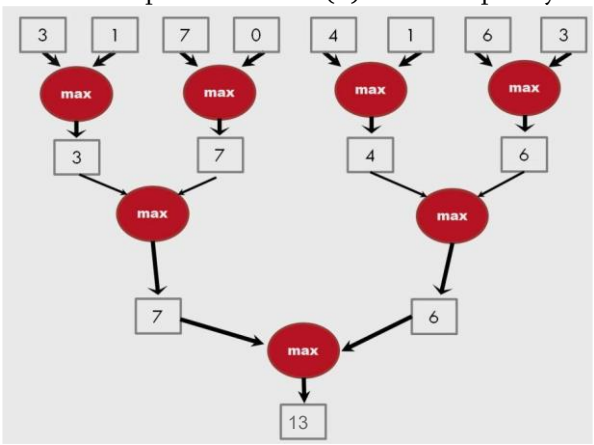
- Input:
  - A set $A = \{a_1, a_2, \ldots, a_n\}$ of n elements ($2 \leq n \leq 1024^2$).
  - A binary associative operator $\oplus$ (e.g. +, *, *max, min*).

  NOTE: Here, we will use addition operator ($\oplus$ = +). So, our parallel reduce will just be parallel sum.

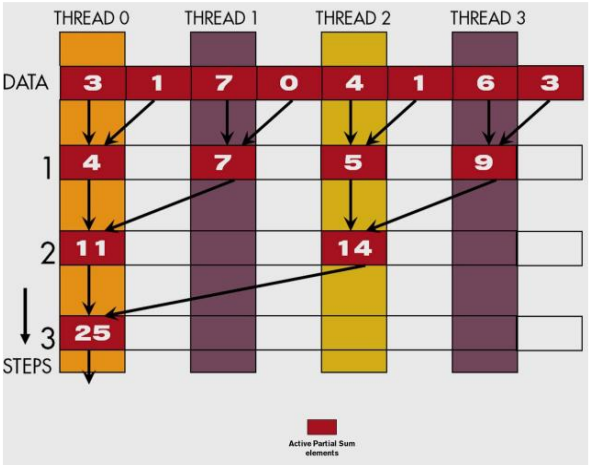
- Output:
  - Return $a_1 \oplus a_2 \oplus \cdots \oplus a_n$.

Parallel Reduce:

- Perform balanced tree-like computation with $O(n)$ work complexity and $\log n$ step complexity.



(NVIDIA and UIUC, 2017)



(NVIDIA and UIUC, 2017)

Practice 4.1: Implement a CUDA C program for parallel sum by using just global device memory.

Practice 4.2: Implement a CUDA C program for parallel sum by using per-block shared memory.