Practice 6: Scan

<u>Objective:</u> To understand how to implement a very useful parallel algorithm called "Scan".

Scan/Prefix Sum:

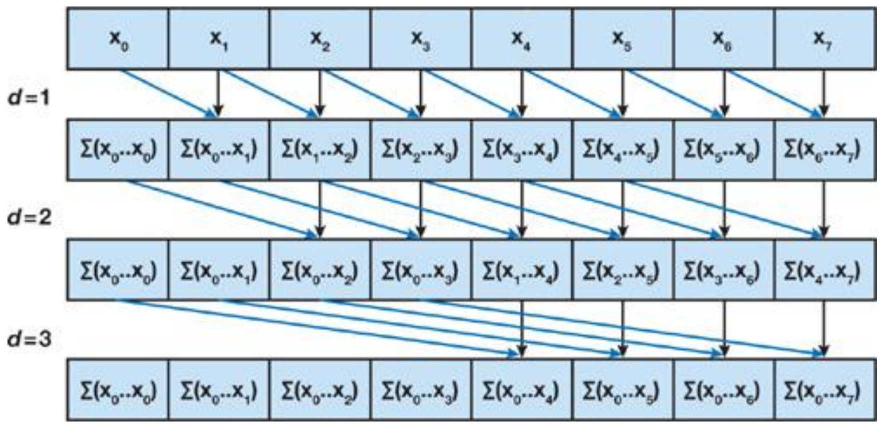
- Input:
    - A sequence of n elements $< x_o, x_1, \ldots, x_{n-1} >$.
    - A binary associative operator $\oplus$ (e.g. +, *, *max, min*).
    - An identity element I associated with the operator.

  NOTE: Here, we will use addition operator ($\oplus$ = +) with I = 0.

- Output:
    - Inclusive version: return $< x_o, (x_0 \oplus x_1), \ldots, (x_0 \oplus x_1 \oplus \cdots \oplus x_{n-1}) >$.
    - Exclusive version: return $< I, x_0, (x_0 \oplus x_1), \ldots, (x_0 \oplus x_1 \oplus \cdots \oplus x_{n-2}) >$.

- Example:
    - Input: <3, 1, 7, 0, 4, 1, 6, 3>, 0, +
    - Inclusive output: <3, 4, 11, 11, 15, 16, 22, 25>
    - Exclusive output: <0, 3, 4, 11, 11, 15, 16, 22>

Parallel Inclusive Scan (Hillis and Steele 1986):



(NVIDIA and UIUC, 2017)

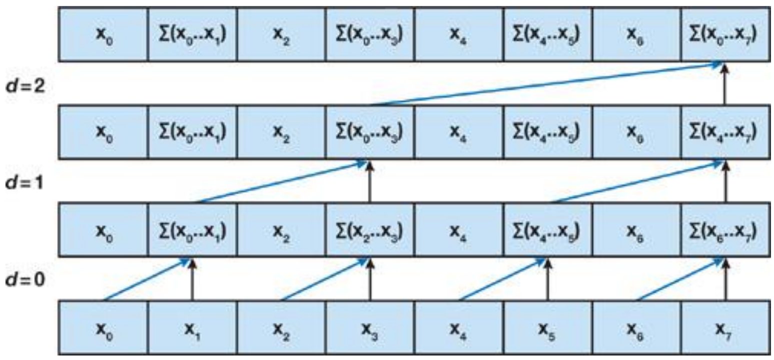
- The parallel inclusive scan by Hillis and Steel needs about O($n \log n$) additions, and so the work complexity of the algorithm is $O(n \ \log n)$.
- The step complexity of the algorithm is $\log n$.
- The parallel inclusive scan is considered work-inefficient since in sequential we need only $O(n)$ additions.
- The pseudocode of the parallel inclusive scan is given as follows:

for $d = 1$ to $\log_2 n$ **do:**
    for all k in parallel **do:**
        if $k \geq 2^d$ **then:**
            $x[k] = x[k - 2^{d-1}] + x[k]$

<u>Practice 5.1</u>: Implement a CUDA C program for the parallel inclusive scan.

Parallel Exclusive Scan (Blelloch 1990):

- The parallel exclusive scan by Blelloch comprises two phases: the *reduce* phase (a.k.a, the *up-sweep* phase) and the *down-sweep* phase.
- The reduce phase is similar to reduce operation. So, the work and step complexities of this phase are $O(n)$ and $\log n$, respectively.
- The *down-sweep* phase is also similar to reduce operation. So, the work and step complexities of this phase are $O(n)$ and $\log n$, respectively.
- In total, the work and step complexities of the algorithm are $O(n)$ and $\log n$, respectively.
- The illustration of the reduce phase is given as follows:

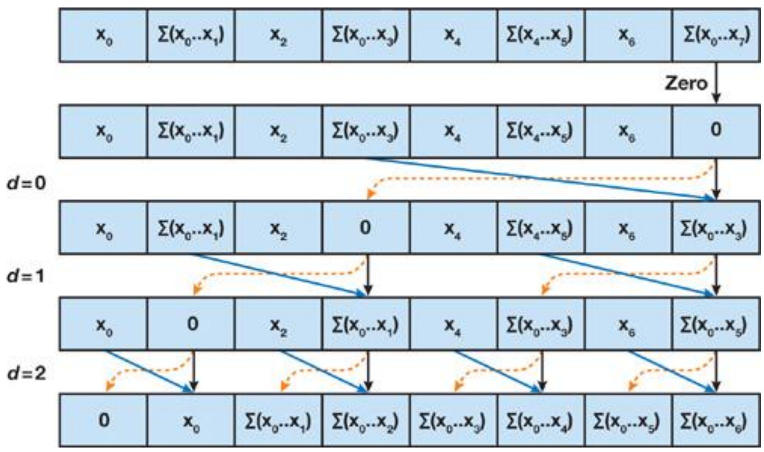


(NVIDIA and UIUC, 2017)

- The pseudocode of the reduce phase is given as follows:

for $d = 0$ to $\log_2 n - 1$ **do:**
    for all $k = 0$ to $n - 1$ by 2^{d+1} in parallel **do:**
        if $k \geq 2^d$ **then:**
            $x[k + 2^{d+1} - 1] = x[k + 2^d - 1] + x[k + 2^d + 1 - 1]$

- The illustration of the down-sweep phase is given as follows:



(NVIDIA and UIUC, 2017)

- The pseudocode of the down-sweep phase is given as follows:

```
x[n − 1] = 0
for d = log₂ n − 1 down to 0 do:
        for all k = 0 to n − 1 by 2^d + 1 in parallel do:
              t = x[k + 2^d − 1]
              x[k + 2^d − 1] = x[k + 2^d + 1 − 1]
              x[k + 2^d + 1 − 1] = t + x[k + 2^d + 1 − 1]
```

<u>Practice 5.2</u>: Implement a CUDA C program for the parallel exclusive scan.