Parallel Algorithms Patterns: Scan

COMPUTAÇÃO DE ALTO DESEMPENHO 2018/2019

HERVÉ PAULINO

SLIDES ADAPTED FROM

"STRUCTURED PARALLEL PROGRAMMING - CIS 410/510, UNIVERSITY OF OREGON"

REZAUL A. CHOWDHURY - CSE 613: PARALLEL PROGRAMMING

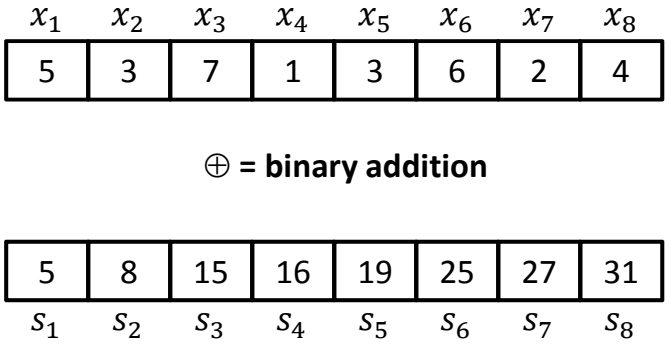
Bibliography

Section 3.3.5 of Structured Parallel Programming, Michael McCool, Arch D. Robison and James Reinders. Morgan Kaufmann (2012)

Parallel Prefix Sum with CUDA, Mark Harris, https://www.mimuw.edu.pl/~ps209291/kgkp/slides/scan.pdf

Scan or Prefix Sum

Scan: computes all partial reduction of a collection



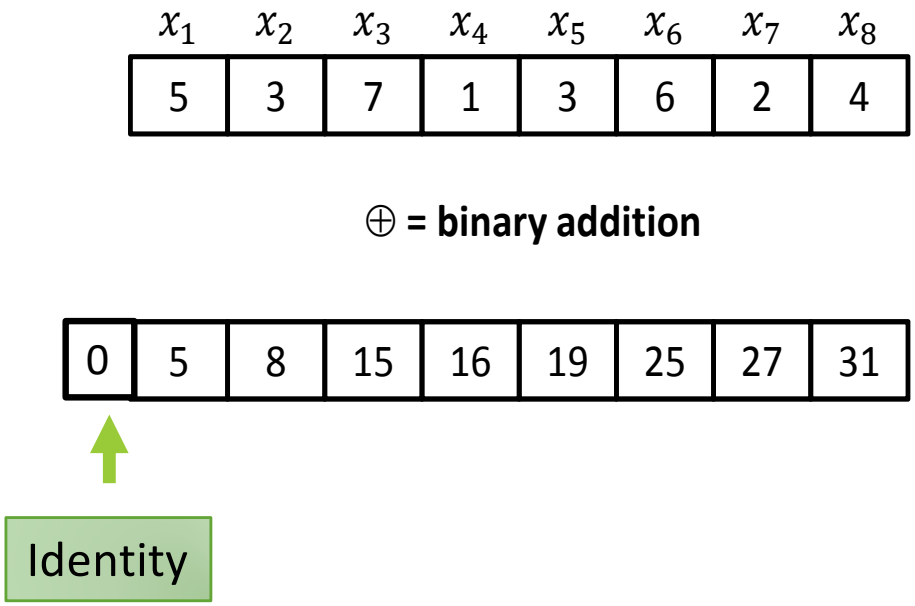
For every output in a collection, a reduction of the input up to that point is computed

If the function being used is associative, the scan can be parallelized

Parallelizing a scan is not obvious at first, because of dependencies to previous iterations in the serial loop

A parallel scan will require more operations than a serial version

Exclusive Scan



Contraction Technique

1. **Reduce**: reduce the original problem to a smaller problem
2. **Conquer**: solve the smaller problem (often recursively)
3. **Expand**: use the solution to the smaller problem to obtain a solution for the original larger problem

Contraction
Scan or Prefix Sum in CPU

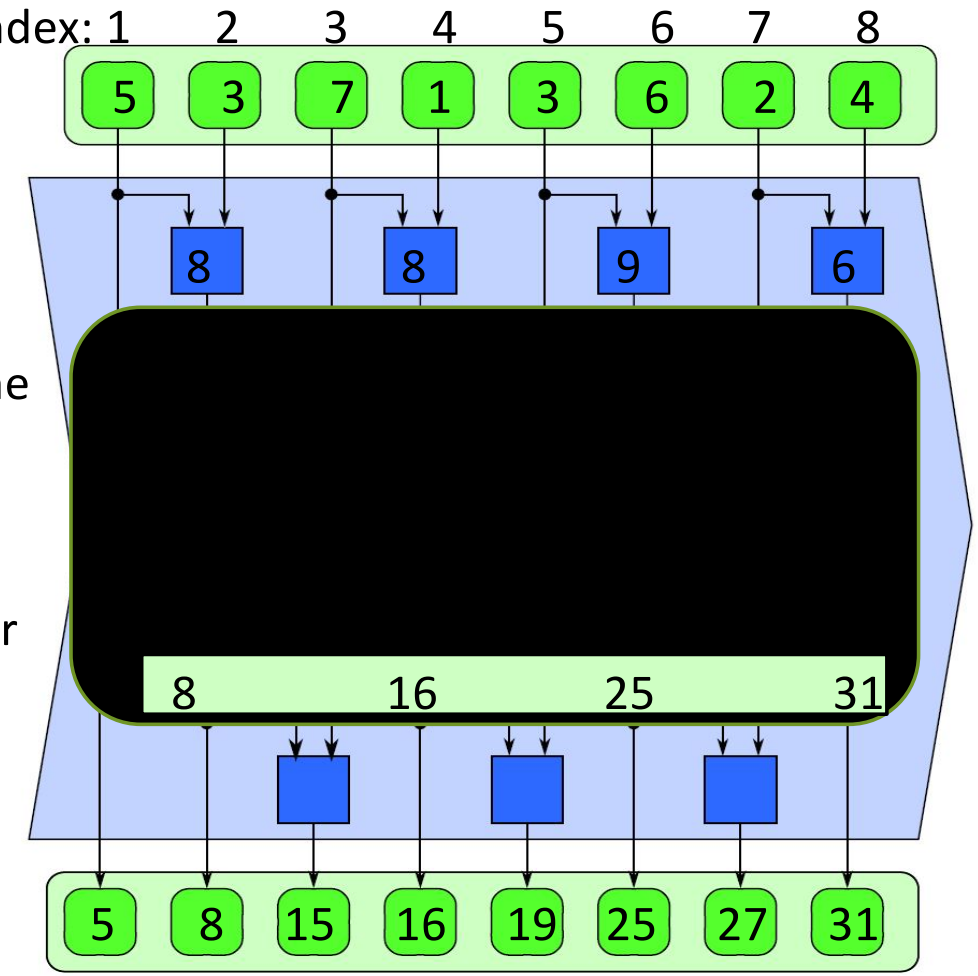
```
prefix_sum(array) {
 if  array.size() == 1 then result[1] = array[1]
  else
    parallel for  i = 1 to array.size()/2 do
      aux[i] = array[2*i-1] + array[2*i]
    aux = prefix_sum(aux)
    parallel for  i = 1 to array.size() do
      if i = 1 then result[1] = array[1]
     else if i = even then result[i] = aux[i/2]
             else result[i] = aux[(i-1)/2] + array[i]
  return result
}
```

Solve problem by processing the result of solving a smaller instance of the problem

Contraction
Scan or Prefix Sum

Black box solves the smaller problem

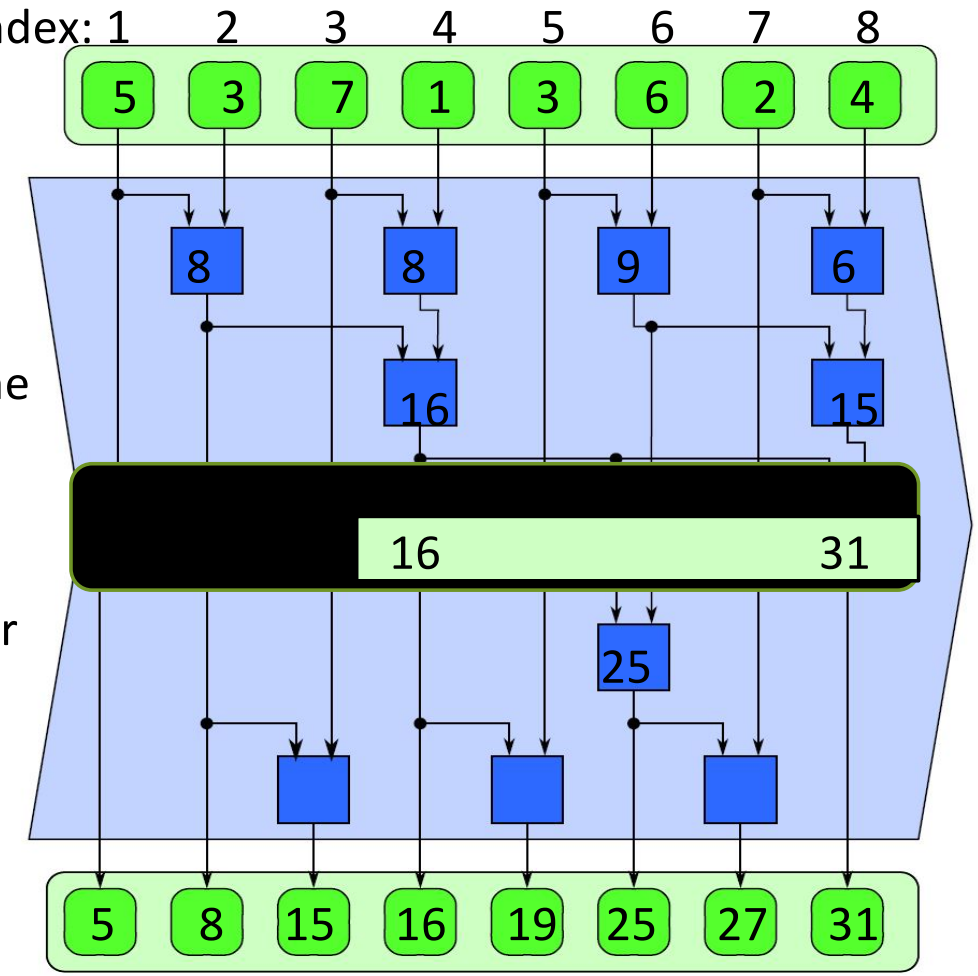
Result is the aux vector of the upper function



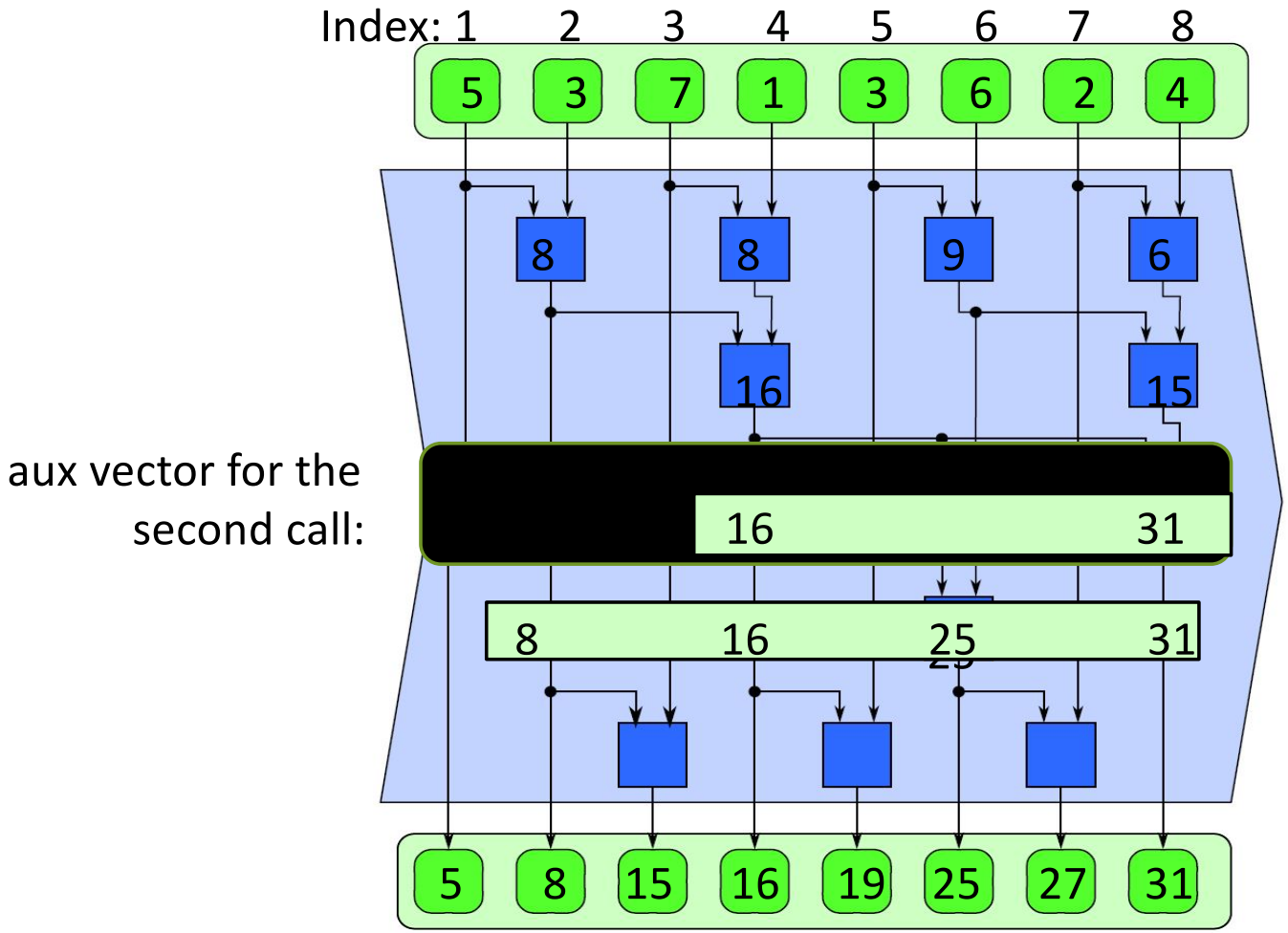
Contraction
Scan or Prefix Sum

Black box solves the smaller problem

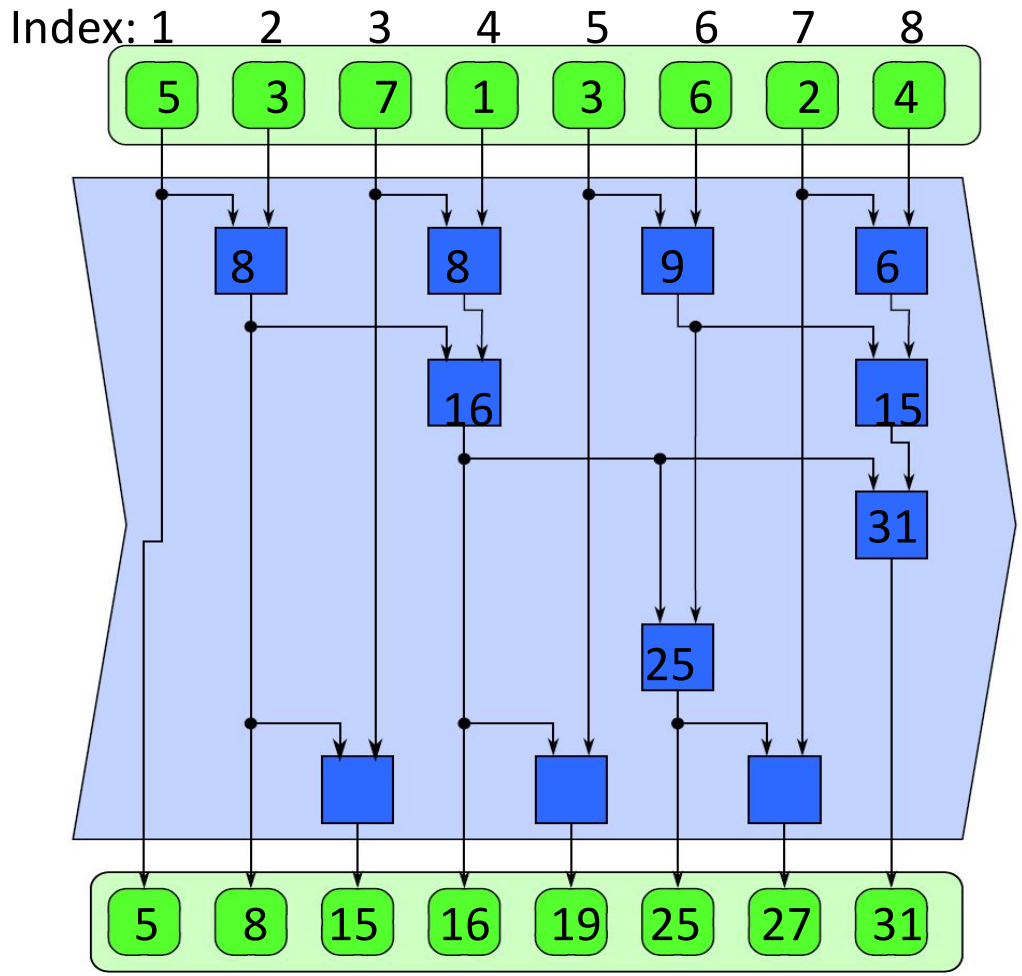
Result is the aux vector of the upper function



Contraction
Scan or Prefix Sum



Contraction
Scan or Prefix Sum



A GPU Implementation in a Single Block

Two phases:

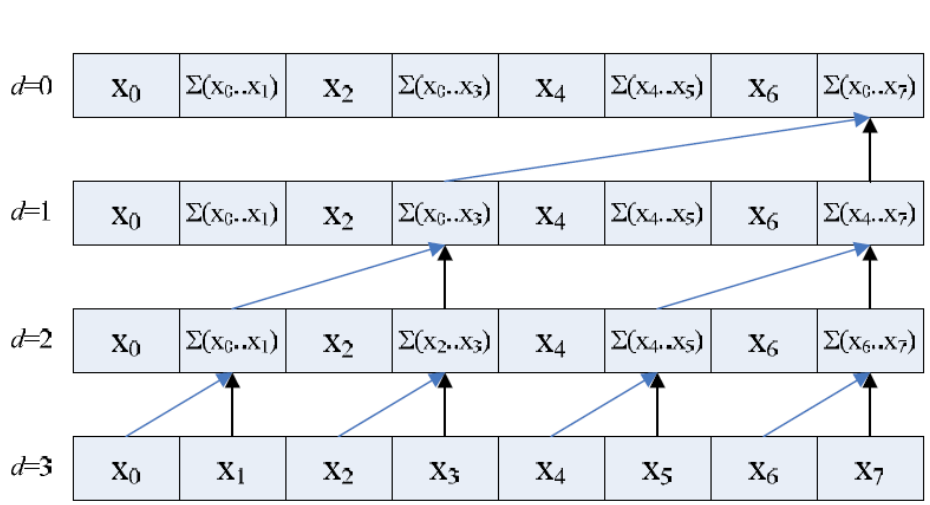
- up-sweep (reduce)
- down-sweep

Thread number: NUM_ELEMENTS /2

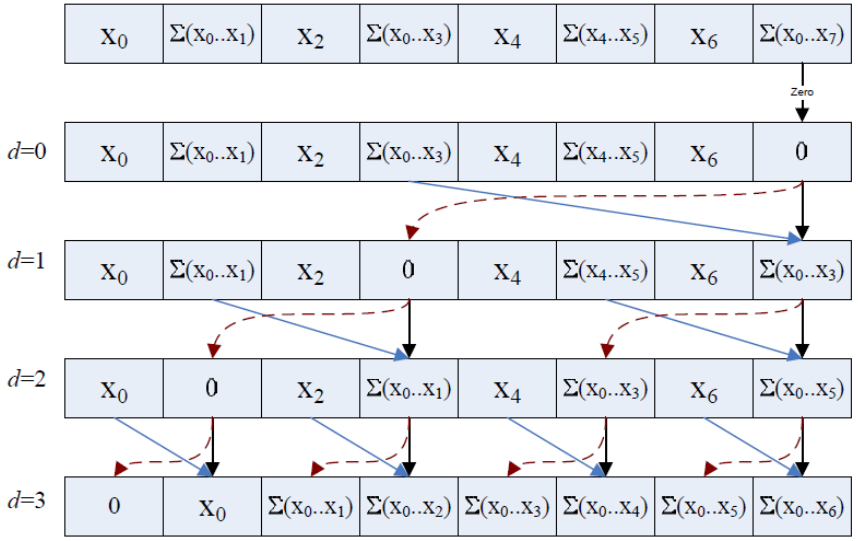
Shared memory size: NUM_ELEMENTS

A GPU Implementation in a Single Block

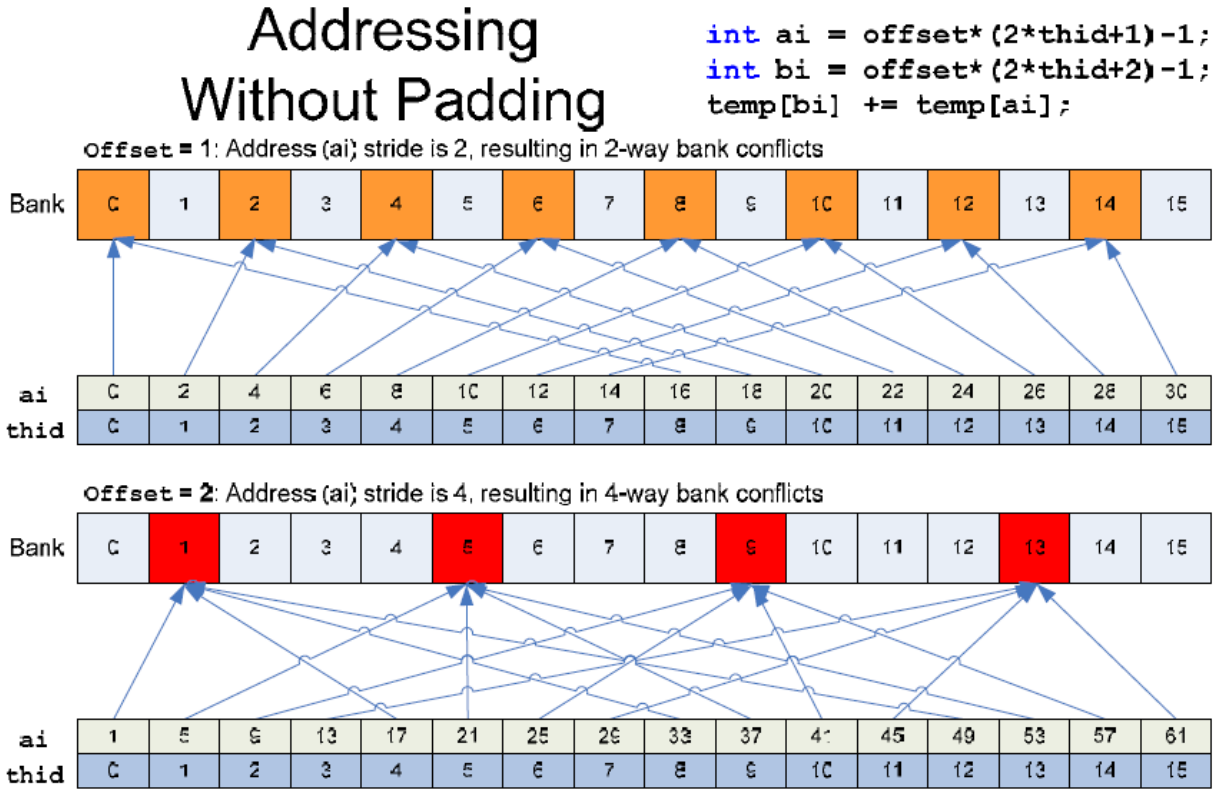
UP-SWEEP



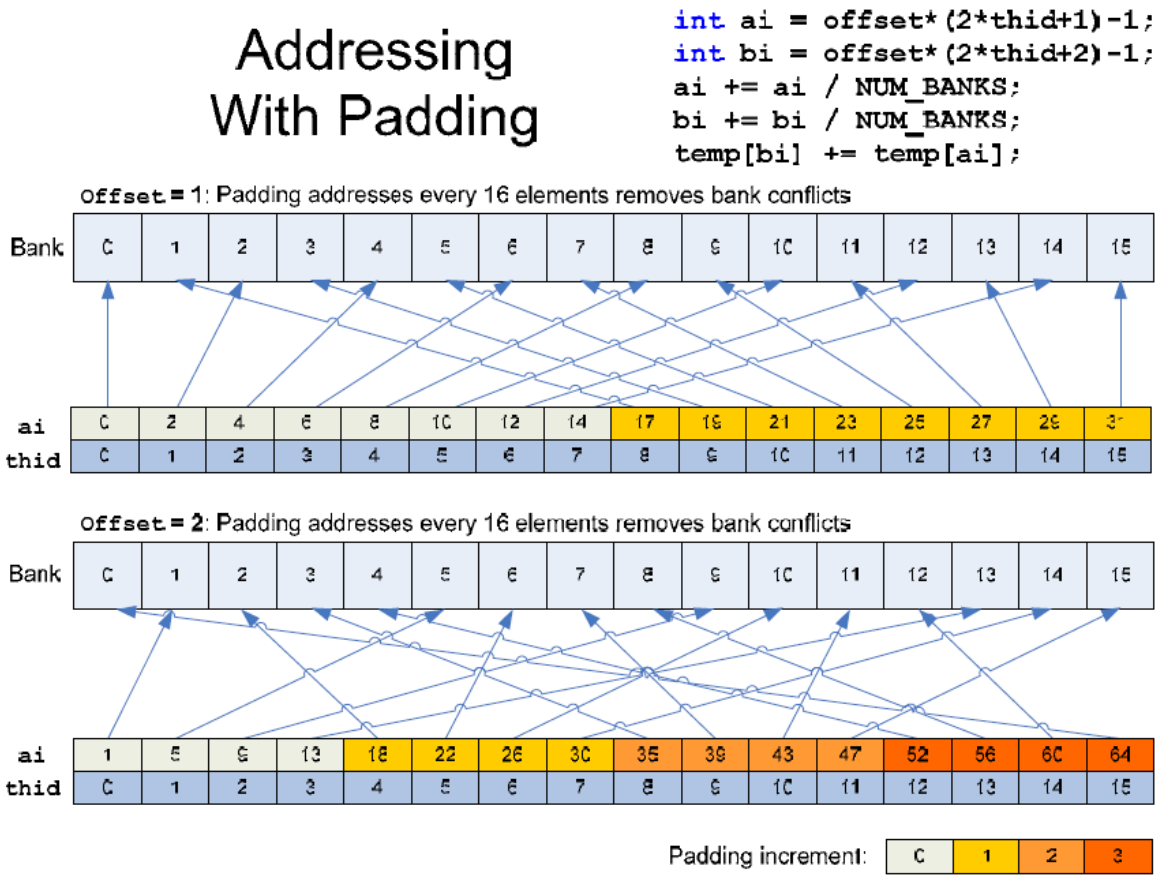
DOWN SWEEP



Avoiding Bank Conflicts



Avoiding Bank Conflicts



Arrays of Arbitrary Size