

Flattening Irregular Nested Parallelism

Cosmin E. Oancea
`cosmin.oancea@dku.dk`

Department of Computer Science (DIKU)
University of Copenhagen

December 2024 DPP Lecture Slides

Parallel Basic Blocks Recap

Part I: Flattening Nested and Irregular Parallelism

What is "Flattening"? Recipe for Applying Flattening

Several Re-Write Rules (inefficient for replicate & iota)

Jagged (Irregular Multi-Dim) Array Representation

Revisiting the Rewrites for Replicate & Iota Nested Inside Map

Revisiting the Solution to Our Example

Part II: Flattening Nested and Irregular Parallelism

Several Applications of Flattening

More Flattening Rules

Flattening by Function Lifting

Flattening Quicksort

Flattening Prime-Number (Sieve) Computation

"To Flatten or Not To Flatten, that is the question"

Zip, Unzip, iota, replicate

- $\text{zip} : [n]\alpha_1 \rightarrow [n]\alpha_2 \rightarrow [n](\alpha_1, \alpha_2)$
- $\text{zip } [a_1, \dots, a_n] [b_1, \dots, b_n] \equiv [(a_1, b_1), \dots, (a_n, b_n)],$
- $\text{unzip} : [n](\alpha_1, \alpha_2) \rightarrow ([n]\alpha_1, [n]\alpha_2)$
- $\text{unzip } [(a_1, b_1), \dots, (a_n, b_n)] \equiv ([a_1, \dots, a_n], [b_1, \dots, b_n]),$
- In some sense `zip/unzip` are syntactic sugar
- $\text{replicate} : (n: \text{ int}) \rightarrow \alpha \rightarrow [n]\alpha$
- $\text{replicate } n \text{ a} \equiv [a, a, \dots, a],$
- $\text{iota} : (n: \text{ int}) \rightarrow [n]\text{int}$
- $\text{iota } n \equiv [0, 1, \dots, n-1]$

Note: in Haskell `zip` does not expect same-length arrays;
in Futhark it does!

Map, Reduce, and Scan Types and Semantics

- $[n]\alpha$ denotes the type of an array of n elements of type α .
- $\text{map} : (\alpha \rightarrow \beta) \rightarrow [n]\alpha \rightarrow [n]\beta$
 $\text{map } f [x_1, \dots, x_n] = [f\ x_1, \dots, f\ x_n],$
i.e., $x_i : \alpha, \forall i$, and $f : \alpha \rightarrow \beta$.
- $\text{reduce} : (\alpha \rightarrow \alpha \rightarrow \alpha) \rightarrow \alpha \rightarrow [n]\alpha \rightarrow \alpha$
 $\text{reduce } \odot e [x_1, x_2, \dots, x_n] = e \odot x_1 \odot x_2 \odot \dots \odot x_n,$
i.e., $e : \alpha, x_i : \alpha, \forall i$, and $\odot : \alpha \rightarrow \alpha \rightarrow \alpha$.
- $\text{scan}^{exc} : (\alpha \rightarrow \alpha \rightarrow \alpha) \rightarrow \alpha \rightarrow [n]\alpha \rightarrow [n]\alpha$
 $\text{scan}^{exc} \odot e [x_1, \dots, x_n] = [e, e \odot x_1, \dots, e \odot x_1 \odot \dots \odot x_{n-1}]$
i.e., $e : \alpha, x_i : \alpha, \forall i$, and $\odot : \alpha \rightarrow \alpha \rightarrow \alpha$.
- $\text{scan}^{inc} : (\alpha \rightarrow \alpha \rightarrow \alpha) \rightarrow \alpha \rightarrow [n]\alpha \rightarrow [n]\alpha$
 $\text{scan}^{inc} \odot e [x_1, \dots, x_n] = [e \odot x_1, \dots, e \odot x_1 \odot \dots \odot x_n]$
i.e., $e : \alpha, x_i : \alpha, \forall i$, and $\odot : \alpha \rightarrow \alpha \rightarrow \alpha$.

Map2, Filter

- $\text{map2} : (\alpha_1 \rightarrow \alpha_2 \rightarrow \beta) \rightarrow [\mathbf{n}]\alpha_1 \rightarrow [\mathbf{n}]\alpha_2 \rightarrow [\mathbf{n}]\beta$
- $\text{map2 } \odot [a_1, \dots, a_n] [b_1, \dots, b_n] \equiv [a_1 \odot b_1, \dots, a_n \odot b_n]$
- $\text{map3} \dots$
- $\text{filter} : (\alpha \rightarrow \text{Bool}) \rightarrow [\mathbf{n}]\alpha \rightarrow [\mathbf{m}]\alpha (\mathbf{m} \leq \mathbf{n})$
- $\text{filter } p [a_1, \dots, a_n] = [a_{k_1}, \dots, a_{k_m}]$ such that $k_1 < k_2 < \dots < k_m$, and denoting $\bar{k} = k_1, \dots, k_m$, we have $(p \ a_j == \text{true}) \forall j \in \bar{k}$, **and** $(p \ a_j == \text{false}) \forall j \notin \bar{k}$.

Note: in Haskell `map2`, `map3` do not expect same-length arrays;
in Futhark they do!

Scatter: A Parallel Write Operator

Scatter **updates in parallel** a base array with a set of values at specified indices:

scatter : *[m]α → [n]int → [n]α → *[m]α

A (data vector) = [b0, b1, b2, b3]

I (index vector) = [2, 4, 1, -1]

X (input array) = [a0, a1, a2, a3, a4, a5]

scatter X | A = [a0, b2, b0, a3, b1, a5]

Scatter: A Parallel Write Operator

Scatter **updates in parallel** a base array with a set of values at specified indices:

$\text{scatter} : *[m]\alpha \rightarrow [n]\text{int} \rightarrow [n]\alpha \rightarrow *[m]\alpha$

A (data vector) = [b₀, b₁, b₂, b₃]

I (index vector) = [2, 4, 1, -1]

X (input array) = [a₀, a₁, a₂, a₃, a₄, a₅]

scatter X | A = [a₀, b₂, b₀, a₃, b₁, a₅]

scatter has $D(n) = \Theta(1)$ and $W(n) = \Theta(n)$,

i.e., requires n update operations (n is the size of I or A, not of X!).

1 Array X is consumed by scatter; following uses of X are illegal!

2 Similarly, X can alias neither I nor A!

In Futhark, scatter check and ignores the indices that are out of bounds (no update is performed on those). This is useful for padding the iteration space in order to obtain regular parallelism.

Partition2/Filter Implementation

partition2: $(\alpha \rightarrow \text{Bool}) \rightarrow [\text{n}]\alpha \rightarrow (\text{i32}, [\text{n}]\alpha)$

In result, the elements satisfying the predicate occur before the others. Can be implemented by means of map, scan, scatter.

```
let partition2 't [n] (dummy: t)           Assume X = [5,4,2,3,7,8], and
    (cond: t -> bool) (X: [n]t) :          cond is T(rue) for even nums.
                                             
    (i64 , [n]t) =
      let cs = map cond X
      let tfs= map (\ f->if f then 1
                     else 0) cs
      let isT= scan (+) 0 tfs
      let i  = isT[n-1]

      let ffs= map (\f->if f then 0
                     else 1) cs
      let isF= map (+i) <| scan (+) 0 ffs
      let inds=map (\(c,iT,iF) ->
                     if c then iT-1
                     else iF-1
                     ) (zip3 cs isT isF)
      let tmp = replicate n dummy
      in (i, scatter tmp inds X)
```

Partition2/Filter Implementation

partition2: $(\alpha \rightarrow \text{Bool}) \rightarrow [\text{n}]\alpha \rightarrow (\text{i32}, [\text{n}]\alpha)$

In result, the elements satisfying the predicate occur before the others. Can be implemented by means of map, scan, scatter.

```
let partition2 't [n] (dummy: t)
    (cond: t -> bool) (X: [n]t) :
    (i64, [n]t) =
let cs = map cond X
let tfs= map (\ f->if f then 1
                  else 0) cs
let isT= scan (+) 0 tfs
let i  = isT[n-1]

let ffs= map (\f->if f then 0
                  else 1) cs
let isF= map (+i) <| scan (+) 0 ffs
let inds=map (\(c,iT,iF) ->
                  if c then iT-1
                  else iF-1
                  ) (zip3 cs isT isF)
let tmp = replicate n dummy
in (i, scatter tmp inds X)
```

Assume X = [5,4,2,3,7,8], and
cond is T(rue) for even nums.

n = 6

cs = [F, T, T, F, F, T]

tfs = [0, 1, 1, 0, 0, 1]

isT = [0, 1, 2, 2, 2, 3]

i = 3

ffs = [1, 0, 0, 1, 1, 0]

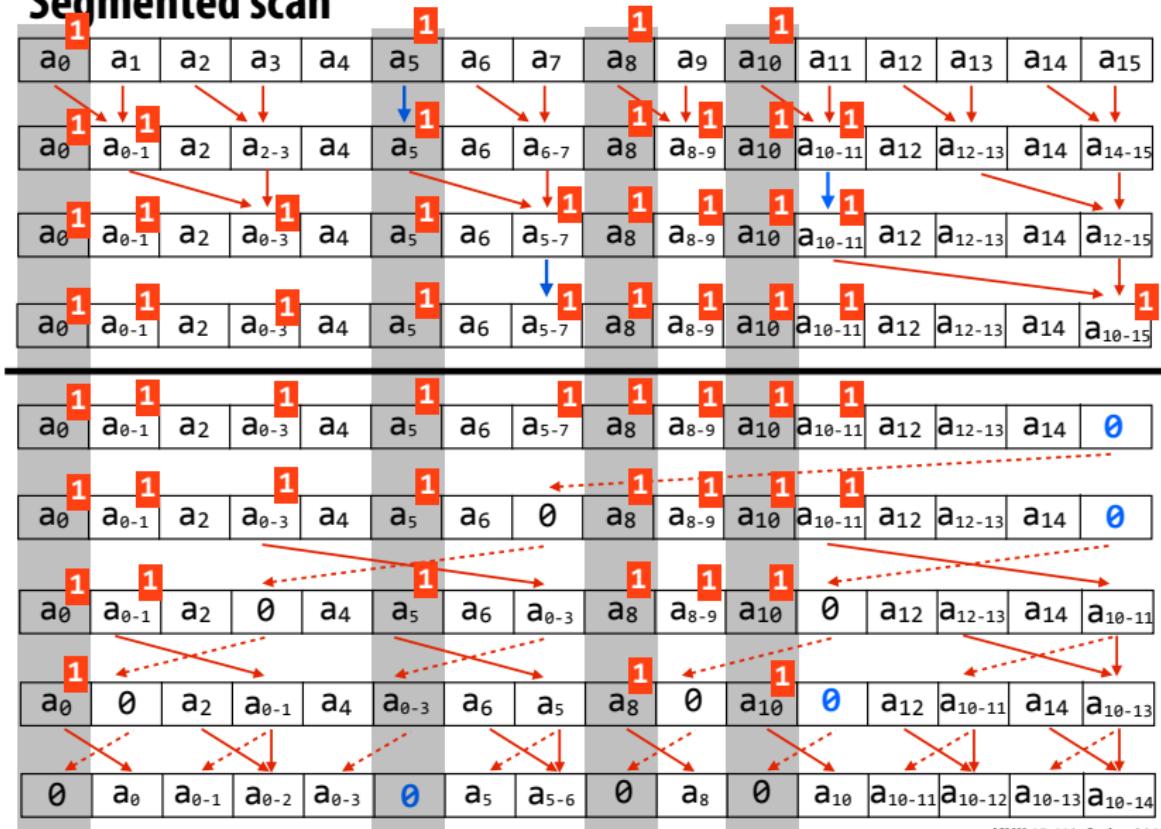
isF = [4, 4, 4, 5, 6, 6]

inds= [3, 0, 1, 4, 5, 2]

flags = [3, 0, 0, 3, 0, 0]

Result = [4, 2, 8, 5, 3, 7]

Segmented scan



Segmented Scan Is a Sort of Scan

```
def sgmscan 't [n] (op: t->t->t) (ne: t)
            (flg : [n]bool) (arr : [n]t) : [n]t =
let flgs_vals =
  zip flg arr |>
  scan (\(f1, x1) (f2, x2) ->
    let f = f1 || f2
    in if f2 then (f, x2)
       else (f, op x1 x2)
  ) (false, ne)
let (_, vals) = unzip flgs_vals
in vals
```

```
sgmscan (+) 0 [1,0,0,1,0, 0, 0]
              [1,2,3,4,5, 6, 7]
              = = = = = =
              [1,3,6,4,9,15,22]
```

```
map (\ row -> scan (+) 0 row)
      [[1,2,3], [4,5, 6, 7]]
      = = = = = =
      [[1,3,6], [4,9,15,22]]
```

Correctness Argument:

Segmented Scan Is a Sort of Scan

```
def sgmscan 't [n] (op: t->t->t) (ne: t)
            (flg : [n]bool) (arr : [n]t) : [n]t =
let flgs_vals =
  zip flg arr |>
  scan (\(f1, x1) (f2, x2) ->
    let f = f1 || f2
    in if f2 then (f, x2)
       else (f, op x1 x2)
  ) (false, ne)
let (_, vals) = unzip flgs_vals
in vals
```

```
sgmscan (+) 0 [1,0,0,1,0, 0, 0]
              [1,2,3,4,5, 6, 7]
              = = = = = =
              [1,3,6,4,9,15,22]
```

```
map (\ row -> scan (+) 0 row)
      [[1,2,3], [4,5, 6, 7]]
      = = = = = =
      [[1,3,6], [4,9,15,22]]
```

Correctness Argument:

verify sequential semantics + associative operator \Rightarrow
parallel semantics also holds

Parallel Basic Blocks Recap

Part I: Flattening Nested and Irregular Parallelism

What is "Flattening"? Recipe for Applying Flattening

Several Re-Write Rules (inefficient for replicate & iota)

Jagged (Irregular Multi-Dim) Array Representation

Revisiting the Rewrites for Replicate & Iota Nested Inside Map

Revisiting the Solution to Our Example

Part II: Flattening Nested and Irregular Parallelism

Several Applications of Flattening

More Flattening Rules

Flattening by Function Lifting

Flattening Quicksort

Flattening Prime-Number (Sieve) Computation

"To Flatten or Not To Flatten, that is the question"

What is "Flattening"?

A code transformation, attributed to Blelloch in the context of the NESL languages, that takes as input a nested parallel program—possibly involving recursion and irregular/jagged arrays—and produces a semantically-equivalent, flat-parallel programs that runs optimally on a PRAM machine.

Meaning: *it is guaranteed to preserve the work and depth of the original nested-parallel program.***

** As long as scan has $O(1)$ depth and concat has $O(1)$ work and . . .

Flattening Pros and Cons

Pros:

- + clever code transformation
- + important as a programming technique as well
(promotes parallel **thinking**)
- + perhaps the only way of mapping a set of challenging problems to capricious architectures such as GPUs
(e.g., that do not support dynamic scheduling of parallelism)

Flattening Pros and Cons

Pros:

- + clever code transformation
- + important as a programming technique as well (promotes parallel **thinking**)
- + perhaps the only way of mapping a set of challenging problems to capricious architectures such as GPUs (e.g., that do not support dynamic scheduling of parallelism)

Cons:

- does not consider communication/locality and hardware gets more and more heterogeneous
- worse, it tends to destroy the available locality and may explode memory footprint
- useful to cover datasets that fall outside the “common case”

Demonstration at the end of the second Flattening lecture

Flattening: A Bird's Eye View

Incomplete recipe for flattening a nested-parallel program consisting of maps and scan/reduce/scatters at innermost level:

I. **Normalize the program (think 3-address form).**

The easy way is to replicate free variables appearing in the current map if they are variant in an outer, enclosing map.

II. **Distribute the parallel context (perfect nest of maps) across the enclosed let-binding statements and handle recurrences by function lifting or map-loop interchange.** Systematic application results in a smallish number of code patterns.

III. **Apply a set of rewrite rules to flatten each pattern**, e.g., treating the cases of a reduce, scan, replicate, iota, scatter, array index which is perfectly nested inside the context.

Flattening: A Bird's Eye View

Incomplete recipe for flattening a nested-parallel program consisting of maps and scan/reduce/scatters at innermost level:

I. **Normalize the program (think 3-address form).**

The easy way is to replicate free variables appearing in the current map if they are variant in an outer, enclosing map.

II. **Distribute the parallel context (perfect nest of maps) across the enclosed let-binding statements and handle recurrences by function lifting or map-loop interchange.** Systematic application results in a smallish number of code patterns.

III. **Apply a set of rewrite rules to flatten each pattern**, e.g., treating the cases of a reduce, scan, replicate, iota, scatter, array index which is perfectly nested inside the context.

Differences w.r.t. the PMPH material:

- 1 **also optimize the number of accesses to global memory**
flattening results in memory-bound performance behavior.
- 2 **cover more rewrite rules and more challenging problems**
(e.g., divide-and-conquer recursion)

PMPH Recap: A Simple Demonstration of How to Flatten

Contrived Example:

```
let arr = [1, 2, 3, 4] in
map (\i -> map (+(i+1)) (iota i)) arr
-- Result: [[2],[3,4],[4,5,6],[5,6,7,8]]
```

PMPH Recap: A Simple Demonstration of How to Flatten

Contrived Example:

```
let arr = [1, 2, 3, 4] in
map (\i -> map (+(i+1)) (iota i)) arr
-- Result: [[2],[3,4],[4,5,6],[5,6,7,8]]
```

I. Normalize the code:

```
map (\i -> let ip1 = i+1 in
        let iot = (iota i) in
        let ip1r = (replicate i ip1)
        in map2 (+) ip1r iot
              ) arr
```

PMPH Recap: A Simple Demonstration of How to Flatten

Contrived Example:

```
let arr = [1, 2, 3, 4] in
map (\i -> map (+(i+1)) (iota i)) arr
-- Result: [[2],[3,4],[4,5,6],[5,6,7,8]]
```

I. Normalize the code:

```
map (\i -> let ip1 = i+1 in
        let iot = (iota i) in
        let ip1r = (replicate i ip1)
        in map2 (+) ip1r iot
              ) arr
```

II. Distribute the map across every statement in the body

and adjust the inputs accordingly (\mathcal{F} denotes the transformation)

```
 $\mathcal{F}(\text{map } (\backslash i \rightarrow \text{map } (+(\text{i}+1)) (\text{iota i})) \text{ arr}) \equiv$ 
1. let ip1s = map (\i -> i+1) arr in -- [2, 3, 4, 5]
2. let iots =  $\mathcal{F}(\text{map } (\backslash i \rightarrow (\text{iota i})) \text{ arr})$  in
3. let ip1rs=  $\mathcal{F}(\text{map2 } (\backslash i \text{ ip1} \rightarrow (\text{replicate i ip1})) \text{ arr ip1s})$ 
4. in  $\mathcal{F}(\text{map2 } (\backslash \text{ip1r iot} \rightarrow \text{map2 } (+) \text{ ip1r iot}) \text{ ip1rs iots})$ 
```

For simplicity we assume arr contains strictly-positive integers.

PMPH Recap: A Simple Demonstration of How to Flatten

According to inefficient rule “iota nested inside a map”

(assuming arr = [1,2,3,4]):

2. let iots = $\mathcal{F}(\text{map } (\lambda i \rightarrow \text{iota } i) \text{ arr})$

\equiv

```
inds = scanexc (+) 0 arr          -- [0,1,3,6]
size = (last inds) + (last arr)    -- 6 + 4 = 10
flag = scatter (replicate size 0)  -- [1, 2, 0, 3, 0, 0, 4, 0, 0, 0]
                                    inds arr
tmp  = replicate size 1
iots = sgmScanexc (+) 0 flag tmp   -- [0, 0, 1, 0, 1, 2, 0, 1, 2, 3]
```

PMPH Recap: A Simple Demonstration of How to Flatten

According to inefficient rule “replicate nested inside a map”

(assuming arr = [1,2,3,4]):

```
3. let ip1rs=  $\mathcal{F}(\text{map2 } (\backslash \text{ i ip1} \rightarrow \text{replicate i ip1}) \text{ arr ip1s})$ 
 $\equiv$ 
vals = scatter (replicate size 0) inds ip1s -- [2,3,0,4,0,0,5,0,0,0]
ip1rs= sgmScaninc (+) 0 flag vals -- [2,3,3,4,4,4,5,5,5,5]
```

According to rule “map nested inside a map”

```
 $\mathcal{F}(\text{map2 } (\backslash \text{ ip1r iot} \rightarrow \text{map2 } (+) \text{ ip1r iot}) \text{ ip1rs iots})$ 
 $\equiv$ 
4. result = map2 (+) ip1rs iots
-- [2, 3, 3, 4, 4, 4, 5, 5, 5]
-- [0, 0, 1, 0, 1, 2, 0, 1, 2, 3]
-- + + + + + + + +
-----
-- [2, 3, 4, 4, 5, 6, 5, 6, 7, 8] values
%-- [1, 2, 0, 3, 0, 0, 4, 0, 0, 0] flags
```

**At each step we also reason about the shape of the resulting array.
The shape of the 2D jagged arrays iots, ip1rs, result is arr.**

Parallel Basic Blocks Recap

Part I: Flattening Nested and Irregular Parallelism

What is "Flattening"? Recipe for Applying Flattening

Several Re-Write Rules (**inefficient for replicate & iota**)

Jagged (Irregular Multi-Dim) Array Representation

Revisiting the Rewrites for Replicate & Iota Nested Inside Map

Revisiting the Solution to Our Example

Part II: Flattening Nested and Irregular Parallelism

Several Applications of Flattening

More Flattening Rules

Flattening by Function Lifting

Flattening Quicksort

Flattening Prime-Number (Sieve) Computation

"To Flatten or Not To Flatten, that is the question"

Nested vs Flattened Parallelism: Scan inside a Map

(1) Scan nested inside a map:

```
res = map (\row->scaninc (+) 0 row) [[1,3], [2,4,6]]  
≡  
res = [ scaninc (+) 0 [1,3],      scaninc (+) 0 [2,4,6] ]  
≡  
res = [ [ 1, 4],                  [2, 6, 12] ]
```

Nested vs Flattened Parallelism: Scan inside a Map

(1) Scan nested inside a map:

```
res = map (\row->scaninc (+) 0 row) [[1,3], [2,4,6]]  
≡  
res = [ scaninc (+) 0 [1,3],      scaninc (+) 0 [2,4,6] ]  
≡  
res = [ [ 1, 4],                  [2, 6, 12] ]
```

becomes a segmented scan, which requires a flag array as arg:

```
sgmScaninc (+) 0 [1, 0, 1, 0, 0] [1, 3, 2, 4, 6] ≡ [ 1, 4, 2, 6, 12 ]
```

Flattening a scan directly nested inside a map:

- $S_{arr}^1, F_{arr}, D_{arr}$ denote the shape, flag & flat data of input arr.
- The flat-data result is obtained by a segmented scan.
- The shape of the result array is the same as the input array.

$\mathcal{F}(res = \text{map } (\lambda \text{row} \rightarrow \text{scan } (\odot) 0 \odot \text{row}) \text{ arr}) \Rightarrow$
 $S_{res}^1 = S_{arr}^1$
 $D_{res} = \text{sgmScan } (\odot) 0 \odot F_{arr} D_{arr}$

Nested vs Flattened Parallelism: Map inside a Map

(2) Map nested inside a map:

```
res = map (\row->map f row) [[1,3], [2,4,6]]  
≡  
res = [ map f [1, 3],      map f [2, 4, 6] ]  
≡  
res = [ [f(1),f(3)], [f(2),f(4),f(6)] ]
```

Nested vs Flattened Parallelism: Map inside a Map

(2) Map nested inside a map:

```
res = map (\row->map f row) [[1,3], [2,4,6]]  
≡  
res = [ map f [1, 3],      map f [2, 4, 6] ]  
≡  
res = [ [f(1),f(3)], [f(2),f(4),f(6)] ]
```

Flattening a map directly nested inside a map:

- the flat-data array is obtained by a map on the flat input;
- the shape of the result array is the same as the input array.

$\mathcal{F}(\text{res} = \text{map } (\lambda \text{row} \rightarrow \text{map } f \text{ row}) \text{ arr}) \Rightarrow$

$$S_{\text{res}}^1 = S_{\text{arr}}^1$$

$$D_{\text{res}} = \text{map } f \text{ D}_{\text{arr}}$$

Nested vs Flattened Parallelism: Replicate inside Map

(3) Replicate nested inside a map:

```
res = map2 (\ n m -> replicate n m) [1,0,3,2] [7,3,8,9] ≡  
res = [ replicate 1 7, replicate 0 3, replicate 3 8, replicate 2 9 ]  
res = [ [7], [], [8,8,8], [9,9] ]
```

Nested vs Flattened Parallelism: Replicate inside Map

(3) Replicate nested inside a map:

```
res = map2 (\ n m -> replicate n m) [1,0,3,2] [7,3,8,9] ≡  
res = [ replicate 1 7, replicate 0 3, replicate 3 8, replicate 2 9 ]  
res = [ [7], [], [8,8,8], [9,9] ]
```

```
res = map2(\n m-> replicate n m) ns ms
```

becomes a scan-scatter composition:

Nested vs Flattened Parallelism: Replicate inside Map

(3) Replicate nested inside a map:

```
res = map2 (\ n m -> replicate n m) [1,0,3,2] [7,3,8,9] ≡  
res = [ replicate 1 7, replicate 0 3, replicate 3 8, replicate 2 9 ] ≡  
res = [ [7], [], [8,8,8], [9,9] ]
```

```
res = map2(\n m-> replicate n m) ns ms
```

becomes a scan-scatter composition:

1. the shape of the result array is ns
 - 2-3. builds the indices at which segment start (-1 for null shape)
 4. get the size of the flat array (summing ns)
 - 5-6. write the ms and ns values at the start of their segments
 7. propagate the ms values throughout their segments.

```

F(res = map2 (\n m -> replicate n m) ns ms) =>      -- ms = [7,3,8,9]
1. Sres1 = ns                                -- ns = [1,0,3,2]
2. inds' = scanexc (+) 0 ns                  -- [0,1,1,4]
3. inds = map2 (\n i->if n>0 then i else -1) ns inds' -- [0,-1,1,4]
4. size = (last inds') + (last ns)           -- 4 + 2 = 6
5. vls = scatter (replicate size 0) inds ms    -- [7, 8, 0, 0, 9, 0]
6. Fres = scatter (replicate size false) inds   -- [1, 1, 0, 0, 1, 0]
               (replicate size true)
7. Dres = sgmScaninc (+) 0 Fres vls          -- [7, 8, 8, 8, 9, 9]

```

Nested vs Flattened Parallelism: Iota inside Map

(4) Iota nested inside a map ($(\text{iota } n) \equiv [0, \dots, n-1]$):

```
res = map (\i -> iota i) [1,3,2] ≡  
res = [ iota 1, iota 3, iota 2 ] ≡ [ [0], [0,1,2], [0,1] ]
```

Nested vs Flattened Parallelism: Iota inside Map

(4) Iota nested inside a map ($(\text{iota } n) \equiv [0, \dots, n-1]$):

```
res = map (\i -> iota i) [1,3,2] ≡  
res = [ iota 1, iota 3, iota 2 ] ≡ [ [0], [0,1,2], [0,1] ]
```

boils down to a segmented scan applied to an array of ones:

1. by definition of iota, ns contains the size of each subarray, hence the shape of the result is ns;
- 2-3. the flag-array of the result, F_{res} , is constructed from ns; (we will introduce function mkFlagArray a bit later).
4. the result is obtained by an exclusive segmented scan operation applied to an array of ones.

```
F(res = map (\n -> iota n) ns) ⇒  
1. S1res = ns -- ns = [1, 3, 2]  
2. trues = replicate (length ns) true  
3. (., Fres) = mkFlagArray ns false trues -- Fres = [1, 1, 0, 0, 1, 0]  
4. Dres = sgmScanexc (+) 0 Fres (replicateflenres 1) -- [0, 0, 1, 2, 0, 1]
```

Note 1: $\text{iota } n \equiv \text{scan}^{\text{exc}} (+) 0 (\text{replicate } n \ 1)$.

Note 2: 1 and 0 denote true and false; flen_{res} is the sum of ns.

Parallel Basic Blocks Recap

Part I: Flattening Nested and Irregular Parallelism

What is "Flattening"? Recipe for Applying Flattening

Several Re-Write Rules (inefficient for replicate & iota)

Jagged (Irregular Multi-Dim) Array Representation

Revisiting the Rewrites for Replicate & Iota Nested Inside Map

Revisiting the Solution to Our Example

Part II: Flattening Nested and Irregular Parallelism

Several Applications of Flattening

More Flattening Rules

Flattening by Function Lifting

Flattening Quicksort

Flattening Prime-Number (Sieve) Computation

"To Flatten or Not To Flatten, that is the question"

Shape-Based Representation

- Two dimensional arrays:

```
arr = [ [1,2,3], [4], [], [5,6] ]
```

⇒

```
S0arr = [4]
```

```
S1arr = [3, 1, 0, 2]
```

```
Darr = [1, 2, 3, 4, 5, 6]
```

Shape-Based Representation

- Two dimensional arrays:

```
arr = [ [1,2,3], [4], [], [5,6] ]
```

⇒

$S_{arr}^0 = [4]$

$S_{arr}^1 = [3, 1, 0, 2]$

$D_{arr} = [1, 2, 3, 4, 5, 6]$

- Three dimensional arrays:

```
arr = [ [], [ [1,2,3], [4], [], [5,6] ], [ [7], [], [8,9,10] ] ]
```

⇒

Shape-Based Representation

- Two dimensional arrays:

```
arr = [ [1,2,3], [4], [], [5,6] ]
```

⇒

```
S0arr = [4]
```

```
S1arr = [3, 1, 0, 2]
```

```
Darr = [1, 2, 3, 4, 5, 6]
```

- Three dimensional arrays:

```
arr = [ [], [ [1,2,3], [4], [], [5,6] ], [ [7], [], [8,9,10] ] ]
```

⇒

```
S0arr = [3]
```

```
S1arr = [0, 4, 3]
```

```
S2arr = [3, 1, 0, 2, 1, 0, 3]
```

```
flenarr = 10
```

```
Darr = [1, 2, 3, 4, 5, 6, 7, 8, 9, 10]
```

Assume a n-dimensional array; The following invariant holds:

```
length Siarr = reduce (+) 0 Si-1arr, ∀1 ≤ i < n
```

```
length Darr = reduce (+) 0 Sn-1arr
```

Flat Representation: Auxiliary Structures

```
arr = [ [], [ 1, 2, 3], [4], [], [5,6] ], [ [7], [], [8,9,10] ]  
⇒  
Sarr0 = [3]  
Sarr1 = [0, 4, 3]  
Sarr2 = [3, 1, 0, 2, 1, 0, 3]  
Darr = [1, 2, 3, 4, 5, 6, 7, 8, 9, 10]
```

- **Offset Indices (B)**: segment-start offset in the flat data:
 $B_{arr}^1 = [0, 0, 6]$
 $B_{arr}^2 = [0, 3, 4, 4, 6, 7, 7]$
- **Flag Array (F)**: start of a segment indicated by a true value (could also use !=0 integrals), e.g., used for segmented scans:

```
Farr1 = [1, 0, 0, 0, 0, 0, 1, 0, 0, 0]  
Farr2 = [1, 0, 0, 1, 1, 0, 1, 1, 0, 0]
```

- **Segment and Inner indices (II)**:

```
IIarr1 = [1, 1, 1, 1, 1, 1, 2, 2, 2, 2]  
IIarr2 = [0, 0, 0, 1, 3, 3, 0, 2, 2, 2]  
IIarr3 = [0, 1, 2, 0, 0, 1, 0, 0, 1, 2]
```

Auxiliary Structures: Intuitive Motivation

Auxiliary structures are useful to optimize the replication of values.

Nested-Execution Example:

```
let xss = [ [1,2,3], [], [5,7] ]
let ys = [ 4, 2, 1 ]
let rss = map2 (\ xs y -> map (+y) xs ) xss ys
⇒
rss = [ map (+4) [1,2,3], map (+2) [], map (+1) [5,7] ]
rss = [ [5,6,7], [], [6,8] ]
```

Traditional flattening would replicate the values of y:

```
let (S1yss,Dyss) = F(map2 (\ xs y -> replicate (length xs) y) xss ys)
let Drss = map2 (\ x y -> x + y) Dxss Dyss
⇒
Dxss = [1, 2, 3, 5, 7]
        + + + +
Dyss = [4, 4, 4, 1, 1]
        = = = =
Drss = [5, 6, 7, 6, 8]
```

Auxiliary Structures: Intuitive Motivation

Auxiliary structures are useful to optimize the replication of values.

Nested-Execution Example:

```
let XSS = [ [1,2,3], [], [5,7] ]
let YS = [ 4, 2, 1 ]
let RSS = map2 (\ xs y -> map (+y) xs ) XSS YS
⇒
RSS = [ map (+4) [1,2,3], map (+2) [], map (+1) [5,7] ]
RSS = [ [5,6,7], [], [6,8] ]
```

Using the auxiliary structures we indirectly access other arrays:

```
let DRSS = map2 (\ x sgmind -> x + ys[sgmind]) XSS II1RSS
⇒
S1RSS = [3, 0, 2]
II1RSS = [0, 0, 0, 2, 2]
DXSS = [1, 2, 3, 5, 7]
DRSS = [1+4, 2+4, 3+4, 5+1, 7+1] = [5, 6, 7, 6, 8]
```

But what have we gained? Creating Π_{RSS}^1 is as expensive as XSS (or better said the expanded YS from the other slide) ...

Auxiliary Structures: Intuitive Motivation

Auxiliary structures are useful to optimize replication:

- they depend only on the shape of the result (created once)
- can indirectly access several lower-dimensional arrays, sharing parallel dimensions!

Nested-Execution Example:

```
let XSS = [ [1,2,3], [], [5,7] ]
let ys  = [ 4, 2, 1 ]
let zs  = [ 1, 2, 3 ]
let RSS = map3 (\ xs y z -> map (\x -> x*y + z) xs ) XSS ys zs
⇒
RSS = [ [5,9,13], [], [8,10] ]
```

Using the auxiliary structures we indirectly access other arrays:

```
let D RSS = map2 (\ y sgmind -> x*ys[sgmind] + zs[sgmind]) D XSS ||1 RSS
⇒
||1 RSS = [0, 0, 0, 2, 2]
D XSS = [1, 2, 3, 5, 7]
D RSS = [1*4+1, 2*4+1, 3*4+1, 5*1+3, 7*1+3] = [5, 9, 13, 8, 10]
```

**We build Π_{RSS}^1 once and reuse it twice. Also improves locality:
ys and zs are much smaller than XSS, hence reused from L1/2\$.**

Auxiliary Structures: Intuitive Motivation

Nested-Execution Example:

```
let xss = [ [1,3], [2] ]
let yss = [ [2], [4,5] ]
let rss = map2 (\xs ys -> map (\x -> map (+x) ys ) xs ) xss yss
⇒
rss = [ [[3],[5]], [[6,7]] ]
```

Using the auxiliary structures we indirectly access other arrays:

```
let Drss = map3(\ s1 s2 s3 -> let ind_x = B1xss[s1] + s2
                           let ind_y = B1yss[s1] + s3
                           in X[ind_x] + Y[ind_y]
                  ) ||1rss ||2rss ||3rss
⇒
B1xss = [0, 2]
B1yss = [0, 1]
||1rss = [0, 0, 1, 1]
||2rss = [0, 1, 0, 0]
||3rss = [0, 0, 0, 1]
Drss = [ Dxss[0+0]+Dyss[0+0] , Dxss[0+1]+Dyss[0+0]
        , Dxss[2+0]+Dyss[1+0] , Dxss[2+0]+Dyss[1+1] ]
Drss = [ 1+2, 3+2, 2+4, 2+5] = [ 3, 5, 6, 7]
```

Constructing the Offset Indices (B)

```
arr = [ [], [ 1,2,3 ], [4], [], [5,6] ], [ [7], [], [8,9,10] ]  
⇒  
Sarr0 = [3]  
Sarr1 = [0, 4, 3]  
Sarr2 = [3, 1, 0, 2, 1, 0, 3]  
Darr = [1, 2, 3, 4, 5, 6, 7, 8, 9, 10]
```

Offset Indices (B): segment-start offset in the flat data:

```
Barr1 = [0, 0, 6]  
Barr2 = [0, 3, 4, 4, 6, 7, 7]
```

How to construct Offset Indices (B)?

Constructing the Offset Indices (B)

```
arr = [ [], [ 1,2,3], [4], [], [5,6] ], [ [7], [], [8,9,10] ]  
⇒  
S0arr = [3]  
S1arr = [0, 4, 3]  
S2arr = [3, 1, 0, 2, 1, 0, 3]  
Darr = [1, 2, 3, 4, 5, 6, 7, 8, 9, 10]
```

Offset Indices (B): segment-start offset in the flat data:

```
B1arr = [0, 0, 6]  
B2arr = [0, 3, 4, 4, 6, 7, 7]
```

How to construct Offset Indices (B)?

By exclusive scanning the corresponding shape and reindexing!

```
B2arr = scanexc (+) 0 S2arr -- [0, 3, 4, 4, 6, 7, 7]
```

```
B1arr = scanexc (+) 0 S1arr -- [0, 0, 4]  
|> map (\ i -> B2arr [ i ]) -- [0, 0, 6]
```

Constructing the Flag Array

From now on, we discuss only TWO-dimensional irregular arrays!

```
def mkFlagArray 't [m]
    (aoa_shp: [m]u32) (zero: t)          -- aoa_shp=[0,3,1,0,4,2,0]
    (aoa_val: [m]t) : ([m]u32, []t) =   -- aoa_val=[1,1,1,1,1,1,1]
let shp_rot = map (\i->if i==0 then 0      -- shp_rot=[0,0,3,1,0,4,2]
                    else aoa_shp[i-1]
                    ) (iota m)
let shp_scn = scan (+) 0 shp_rot           -- shp_scn=[0,0,3,4,4,8,10]
let aoa_len = if m == 0 then 0i64          -- aoa_len = 10
              else i64.u32 <|
                  shp_scn[m-1]+aoa_shp[m-1]
let shp_ind = map2 (\shp ind ->          -- shp_ind=
                     if shp==0 then -1i64 -- [-1,0,3,-1,4,8,-1]
                     else i64.u32 ind -- scatter
                     ) aoa_shp shp_scn        -- [0,0,0,0,0,0,0,0,0,0]
let r = scatter (replicate aoa_len zero)  -- [-1,0,3,-1,4,8,-1]
                                         shp_ind aoa_val
                                         -- [1,1,1, 1,1,1, 1]
in (shp_scn, r)                         -- r=[1,0,0,1,1,0,0,0,1,0]
```

Versatile: computes B^1 and F^1 of a 2D jagged array of shape `aoa_shp`, with the start-segment values taken from `aoa_val`.

**Unless you have a good reason, F should be a bool array
(to reduce memory traffic).**

Constructing the Segment and Inner Indices

From now on, we discuss only TWO-dimensional irregular arrays!

```
arr = [ [1,2,3], [4], [], [5,6], [7], [], [8,9,10] ]
```

⇒

```
Sarr0 = [7]
```

```
Sarr1 = [3, 1, 0, 2, 1, 0, 3]
```

```
flenarr = reduce (+) 0 Sarr1 = 10
```

```
Darr = [1, 2, 3, 4, 5, 6, 7, 8, 9, 10]
```

Segment and Inner indices (II):

```
IIarr1 = [0, 0, 0, 1, 3, 3, 4, 6, 6, 6]
```

```
IIarr2 = [0, 1, 2, 0, 0, 1, 0, 0, 1, 2]
```

Constructing Segment and Inner indices (II):

```
(Barr1, Farr) = mkFlagArray Sarr1 0 (iota (length Sarr1))  
-- ([0, 3, 4, 4, 6, 7, 7], [0, 0, 0, 1, 3, 0, 4, 6, 0, 0])
```

```
IIarr1 = ???
```

```
IIarr2 = ???
```

Constructing the Segment and Inner Indices

I need to get this:

$II_{arr}^1 = [0, 0, 0, 1, 3, 3, 4, 6, 6, 6]$

from this:

```
(_, Farr) = mkFlagArray Sarr1 0 (iota (length Sarr1))  
-- [0, 0, 0, 1, 3, 0, 4, 6, 0, 0]
```

How ?

Constructing the Segment and Inner Indices

I need to get this:

$$\Pi_{arr}^2 = [0, 1, 2, 0, 0, 1, 0, 0, 1, 2]$$

from these:

$$B_{arr} = [0, 3, 4, 4, 6, 7, 7]$$

$$\Pi_{arr}^1 = [0, 0, 0, 1, 3, 3, 4, 6, 6, 6]$$

$$\text{iota } 10 = [0, 1, 2, 3, 4, 5, 6, 7, 8, 9]$$

Π_{arr}^1 and Π_{arr}^2 have the same length as flat arr, in our case 10.

How?

We can also construct it by binary searching B_{arr} or by means of a segmented scan.

Constructing the Segment and Inner Indices

From now on, we discuss only TWO-dimensional irregular arrays!

```
arr = [ [1,2,3], [4], [], [5,6], [7], [], [8,9,10] ]
```

⇒

```
Sarr0 = [7]
```

```
Sarr1 = [3, 1, 0, 2, 1, 0, 3]
```

```
flenarr = reduce (+) 0 Sarr1 = 10
```

```
Darr = [1, 2, 3, 4, 5, 6, 7, 8, 9, 10]
```

Segment and Inner indices (II):

```
IIarr1 = [0, 0, 0, 1, 3, 3, 4, 6, 6, 6]
```

```
IIarr2 = [0, 1, 2, 0, 0, 1, 0, 0, 1, 2]
```

Constructing Segment and Inner indices (II):

```
(Barr1, Farr) = mkFlagArray Sarr1 0 (iota (length Sarr1))  
-- ([0, 3, 4, 4, 6, 7, 7], [0, 0, 0, 1, 3, 0, 4, 6, 0, 0])
```

```
Farr = map bool.u32 Farr
```

```
IIarr1 = sgmScaninc (+) 0 Farr Farr
```

```
IIarr2 = map2 (\ i sgm -> i - Barr1[sgm] ) (iotaflenarr) IIarr1
```

-- ^ this fuses better & performs less memory traffic than the below:

```
IIarr2 = sgmScaninc (+) 0 Farr (replicateflen 1) |> map (-1)
```

B^{inc} and II^1 Are the Important Ones

Because you can deduce the other arrays by means of simple maps, that fuse better and generate less traffic.

```
arr = [ [1,2,3], [4], [], [5,6], [7], [], [8,9,10] ]  
⇒  
 $S_{arr}^0 = [7]$   
 $S_{arr}^1 = [3, 1, 0, 2, 1, 0, 3]$   
 $flen_{arr} = \text{reduce } (+) \ 0 \ S_{arr}^1 = 10$   
 $D_{arr} = [1, 2, 3, 4, 5, 6, 7, 8, 9, 10]$ 
```

If we know the Segment Offsets (B_{arr}^{inc}) and Indices (II_{arr}^1):

```
 $B_{arr}^{inc} = [3, 4, 4, 6, 7, 7, 10] \text{ -- inclusive scan of } S_{arr}^1 \text{ and}$   
 $II_{arr}^1 = [0, 0, 0, 1, 3, 3, 4, 6, 6, 6]$ 
```

We can efficiently compute the S_{arr}^1 and F_{arr} arrays (also II_{arr}^2) by:

```
 $S_{arr}^1 = \text{iota } (\text{length } B_{arr}^{inc})$   
|> map (\i -> if i == 0 then  $B_{arr}^{inc}[i]$  else  $B_{arr}^{inc}[i] - B_{arr}^{inc}[i-1]$ )
```

```
 $F_{arr} = \text{iota } flen_{arr} \text{ -- } flen_{arr} \text{ is the length of } II_{arr}^1$   
|> map (\i -> if i == 0 then true else  $II_{arr}^1[i] != II_{arr}^1[i-1]$ )
```

Note: B_{arr}^{inc} different than B_{arr}^1 : inclusive vs exclusive scan of shape!

Parallel Basic Blocks Recap

Part I: Flattening Nested and Irregular Parallelism

What is "Flattening"? Recipe for Applying Flattening

Several Re-Write Rules (inefficient for replicate & iota)

Jagged (Irregular Multi-Dim) Array Representation

Revisiting the Rewrites for Replicate & Iota Nested Inside Map

Revisiting the Solution to Our Example

Part II: Flattening Nested and Irregular Parallelism

Several Applications of Flattening

More Flattening Rules

Flattening by Function Lifting

Flattening Quicksort

Flattening Prime-Number (Sieve) Computation

"To Flatten or Not To Flatten, that is the question"

Revisiting Replicate inside Map

(3) Replicate nested inside a map:

```
res = map2 (\ n m -> replicate n m) [1,0,3,2] [7,3,8,9] ≡  
res = [ replicate 1 7, replicate 0 3, replicate 3 8, replicate 2 9 ] ≡  
res = [ [7], [], [8,8,8], [9,9] ]
```

```
res = map2(\n m-> replicate n m) ns ms
```

becomes a very simple gather operation:

1. the shape of the result array is ns
2. build Π^1 of a jagged array of shape ns

Revisiting Replicate inside Map

(3) Replicate nested inside a map:

```
res = map2 (\ n m -> replicate n m) [1,0,3,2] [7,3,8,9] ≡  
res = [ replicate 1 7, replicate 0 3, replicate 3 8, replicate 2 9 ] ≡  
res = [ [7], [], [8,8,8], [9,9] ]
```

```
res = map2(\n m-> replicate n m) ns ms
```

becomes a very simple gather operation:

1. the shape of the result array is ns
2. build II^1 of a jagged array of shape ns
3. gather the corresponding values from ms by indexing through II^1

$\mathcal{F}(\text{res} = \text{map2 } (\backslash n m -> \text{replicate } n m) \text{ ns ms}) \Rightarrow$ -- ms = [7,3,8,9]
1. $S_{\text{res}}^1 = \text{ns}$ -- ns = [1,0,3,2]
2. $\text{II}_{\text{res}}^1 = \dots$ -- construct II^1 for a jagged array of shape ns
3. $D_{\text{res}} = \text{map } (\backslash \text{sgm} -> \text{ms}[\text{sgm}]) \text{ II}_{\text{res}}^1$ -- [7, 8,8,8, 9,9]

Nested vs Flattened Parallelism: iota inside Map

(4) **iota nested inside a map** ($(\text{iota } n) \equiv [0, \dots, n-1]$):

```
res = map (\i -> iota i) [1,3,2] ≡  
res = [ iota 1, iota 3, iota 2 ] ≡ [ [0], [0,1,2], [0,1] ]
```

```
res = map (\n-> iota n) ns
```

The result is exactly the II^2 array of a jagged array of shape ns

$\mathcal{F}(\text{res} = \text{map } (\backslash n \rightarrow \text{iota } n) \text{ ns}) \Rightarrow$

1. $S_{\text{res}}^1 = \text{ns}$ -- $\text{ns} = [1, 3, 2]$
2. $II_{\text{res}}^2 = \dots$ -- construct II^2 of a jagged array of shape ns
4. $D_{\text{res}} = II_{\text{res}}^2$

Parallel Basic Blocks Recap

Part I: Flattening Nested and Irregular Parallelism

What is "Flattening"? Recipe for Applying Flattening

Several Re-Write Rules (inefficient for replicate & iota)

Jagged (Irregular Multi-Dim) Array Representation

Revisiting the Rewrites for Replicate & Iota Nested Inside Map

Revisiting the Solution to Our Example

Part II: Flattening Nested and Irregular Parallelism

Several Applications of Flattening

More Flattening Rules

Flattening by Function Lifting

Flattening Quicksort

Flattening Prime-Number (Sieve) Computation

"To Flatten or Not To Flatten, that is the question"

Revisiting Our Demonstration of How to Flatten

Contrived Example:

```
let arr = [1, 2, 3, 4] in
map (\i -> map (+(i+1)) (iota i)) arr
-- Result: [[2],[3,4],[4,5,6],[5,6,7,8]]
```

I. Normalize the code:

```
map (\i -> let ip1 = i+1 in
      let iot = (iota i) in
      let ip1r = (replicate i ip1)
      in map2 (+) ip1r iot
           ) arr
```

II. Distribute the map across every statement in the body

and adjust the inputs accordingly (\mathcal{F} denotes the transformation)

```
 $\mathcal{F}(\text{map } (\backslash i \rightarrow \text{map } (+(\text{i}+1)) (\text{iota i})) \text{ arr}) \equiv$ 
1. let ip1s = map (\i -> i+1) arr in -- [2, 3, 4, 5]
2. let iots =  $\mathcal{F}(\text{map } (\backslash i \rightarrow (\text{iota i})) \text{ arr})$  in
3. let ip1rs=  $\mathcal{F}(\text{map2 } (\backslash i \text{ ip1} \rightarrow (\text{replicate i ip1})) \text{ arr ip1s})$ 
4. in  $\mathcal{F}(\text{map2 } (\backslash \text{ip1r iot} \rightarrow \text{map2 } (+) \text{ ip1r iot}) \text{ ip1rs iots})$ 
```

We do **not** assume that arr contains strictly-positive integers.

Revisiting Our Example: Think Like a Compiler

```
 $\mathcal{F}(\text{map } (\lambda i \rightarrow \text{map } (+(i+1)) (\text{iota } i)) \text{ arr}) \equiv$ 
1. let ip1s = map (\iota → i+1) arr in -- [2, 3, 4, 5]
2. let iots =  $\mathcal{F}(\text{map } (\lambda i \rightarrow (\text{iota } i)) \text{ arr})$  in
3. let ip1rs =  $\mathcal{F}(\text{map2 } (\lambda i \text{ ip1} \rightarrow (\text{replicate } i \text{ ip1})) \text{ arr} \text{ ip1s})$ 
4. in  $\mathcal{F}(\text{map2 } (\lambda \text{ ip1r } \text{ iot} \rightarrow \text{map2 } (+) \text{ ip1r } \text{ iot}) \text{ ip1rs } \text{ iots})$ 
```

Applying the new rules results in:

1. $S_{\text{res}}^1 = \text{arr}$ -- $\text{arr} = [1, 2, 3, 4]$
2. $(B_{\text{res}}^1, F'_{\text{res}}) = \text{mkFlagArray } \text{arr} \ 0 \ (\text{iota } (\text{length } \text{arr}))$ -- $B_{\text{res}}^1 = [0, 1, 3, 6]$
3. $F_{\text{res}} = \text{map } \text{bool}.u32 \ F'_{\text{res}}$
4. $II_{\text{res}}^1 = \text{sgmScan}^{inc} \ (+) \ 0 \ F_{\text{res}} \ F'_{\text{res}}$ -- $[0, 1, 1, 2, 2, 2, 3, 3, 3, 3]$
5. $\text{ip1s} = \text{map } (\lambda i \rightarrow i+1) \text{ arr}$ -- $[2, 3, 4, 5]$
6. $\text{iots} = \text{map2 } (\lambda \text{ ind } \text{ sgm} \rightarrow \text{ind} - B_{\text{res}}^1[\text{sgm}]) \ (\text{iota } \text{flen}_{\text{res}}) \ II_{\text{res}}^1$
--- = $II_{\text{arr}}^2 = [0, 0, 1, 0, 1, 2, 0, 1, 2, 3]$
7. $\text{ip1rs} = \text{map } (\lambda \text{ sgm} \rightarrow \text{ip1s}[\text{sgm}]) \ II_{\text{res}}^1$ -- $[2, 3, 3, 4, 4, 4, 5, 5, 5, 5]$
8. in $\text{map2 } (+) \ \text{ip1rs } \text{iots}$ -- $[2, 3, 4, 4, 5, 6, 5, 6, 7, 8]$

Lines 6 – 8 are trivially fusible ⇒ the iots and ip1rs arrays are not manifested in memory.

Revisiting Our Example: Think Like a Human

I. Normalize the code:

```
map (\i -> let ip1 = i+1 in
        let iot = (iota i) in
        let ip1r = (replicate i ip1)
        in map2 (+) ip1r iot
            ) arr
```

Using the new intuition results in:

1. $S_{res}^1 = arr$ $-- arr = [1, 2, 3, 4]$
2. $(B_{res}^1, F'_{res}) = \text{mkFlagArray } arr\ 0\ (\text{iota}\ (\text{length}\ arr))$
3. $F_{res} = \text{map bool.u32 } F'_{res}$
4. $II_{res}^1 = \text{sgmScan}^{inc}\ (+)\ 0\ F_{res}\ F'_{res}$
5. $\text{in map2}\ (\backslash\ \text{sgm}\ \text{ind} -> \text{let ip1} = arr[\text{sgm}] + 1$
6. $\text{let iot_el} = \text{ind} - B_{res}^1[\text{sgm}]$
7. $\text{in ip1} + \text{iot_el}$
8. $)\ II_{res}^1\ (\text{iota}\ (\text{length}\ II_{res}^1))$

Have done a tiny bit better job than the compiler, as array ip1s is not manifested either.

Fusion in Futhark

Map fusion:

$$(\mathbf{map} \ g) \circ (\mathbf{map} \ f) \equiv \mathbf{map} \ (g \circ f)$$

$$\begin{aligned} x &= \mathbf{map} \ f \ [\ a_1, & a_2, & \dots, & a_n \] \\ && \downarrow & \downarrow & \downarrow \\ x &\equiv [\ f \ a_1, & f \ a_2, & \dots, & f \ a_n \] \\ && \downarrow & \downarrow & \downarrow \\ \mathbf{map} \ g \ x &= [\ g(f \ a_1), & g(f \ a_2), & \dots, & g(f \ a_n) \] \\ &\equiv & = & = & = \\ \mathbf{map} \ (g \circ f) \ x &= [\ g(f \ a_1), & g(f \ a_2), & \dots, & g(f \ a_n) \] \end{aligned}$$

All other SOACs (reduce, scan, reduce-by-index, scatter) fuse with a map producer, if the mapped array is not used elsewhere.

Direct indexing in the map-produced array prevents fusion.

E.g., assuming array xs of length n the following will **not fuse**:

```
let xs = map f as
let ys = map (\i -> if i == 0 || i == n-1 then 0
                  else xs[i-1] + xs[i] + xs[i+1] ) (iota n)
```

Demonstrating Performance of New vs Old Rules

PERFORMANCE DEMONSTRATION

Demo on Prime Numbers: Haskell Implementation

If we have all primes from 2 to \sqrt{n} we could generate all multiples of these primes (up to n) at once: $\{[2*p:n:p] : p \text{ in } \text{sqr_primes}\}$ in NESL. Also call algorithm recursively on \sqrt{n} \Rightarrow Depth: $O(\lg \lg n)$ (solution of $n^{(1/2)^{\text{depth}}} = 2$). Work: $O(n \lg \lg n)$

```
primesOpt :: Int -> [Int]
primesOpt n =
  if n <= 2 then [2]
  else
    let sqrtN = floor (sqrt (fromIntegral n))
        sqrt_primes = primesOpt sqrtN
        nested = map (\p->let m = (n `div` p)
                      in map (\j-> j*p)
                             [2..m])
                     ) sqrt_primes
    not_primes = reduce (++) [] nested
    mm = length not_primes
    zeros = replicate mm False
    prime_flags=scatter(replicate (n+1) True)
                not_primes zeros
    (primes,_)= unzip $ filter (\(i,f)->f)
                $ (zip [0..n] prime_flags)
in drop 2 primes
```

Demo on Prime Numbers: Haskell Implementation

If we have all primes from 2 to \sqrt{n} we could generate all multiples of these primes (up to n) at once: $\{[2*p:n:p] : p \text{ in } \text{sqr_primes}\}$ in NESL. Also call algorithm recursively on \sqrt{n} \Rightarrow Depth: $O(\lg \lg n)$ (solution of $n^{(1/2)^{\text{depth}}} = 2$). Work: $O(n \lg \lg n)$

```
primesOpt :: Int -> [Int]                                Assume n = 9, sqrtN = 3
primesOpt n =
  if n <= 2 then [2]                                     call primesOpt 3
  else
    let sqrtN = floor (sqrt (fromIntegral n))           n = 3, sqrtN = 1, sqrt_primes = []
      nested = []                                         not_primes = []
      sqrt_primes = primesOpt sqrtN                      mm = 0; zeros = []
      nested = map (\p->let m = (n `div` p)             prime_flags = [T,T,T,T]
                    in map (\j-> j*p)                   mm = 0; zeros = []
                    prime_flags = [0,1,2,3]; returns [2..m])    not_primes = []
      ) sqrt_primes                                     in primesOpt 9, after
      not_primes = reduce (++) [] nested return from primesOpt3,
      mm = length not_primes                           sqrt_primes = [2,3]
      zeros = replicate mm False                      nested = [[4,6,8],[6,9]]
      prime_flags=scatter(replicate (n+1)nTruePrimes = [4,6,8,6,9]
                           not_primes zeros=5;zeros= [F,F,F,F,F]
                           (primes,_)= unzip $ filter ((i,f)->prime_flags= [T,T,T,T,F,T,F,T,
                           $ (zip [0..n] prime_flags)not_primes = [0,1,2,3,5,7]
                           in drop 2 primes                               returns [2,3,5,7]
```

Demonstrating Performance of New vs Old Rules

PERFORMANCE DEMONSTRATION

Parallel Basic Blocks Recap

Part I: Flattening Nested and Irregular Parallelism

What is "Flattening"? Recipe for Applying Flattening

Several Re-Write Rules (inefficient for replicate & iota)

Jagged (Irregular Multi-Dim) Array Representation

Revisiting the Rewrites for Replicate & Iota Nested Inside Map

Revisiting the Solution to Our Example

Part II: Flattening Nested and Irregular Parallelism

Several Applications of Flattening

More Flattening Rules

Flattening by Function Lifting

Flattening Quicksort

Flattening Prime-Number (Sieve) Computation

"To Flatten or Not To Flatten, that is the question"

Eratosthenes Alg. for Computing Prime Numbers up To n

See also "Scan as Primitive Parallel Operation" [Blelloch].

Start with an array of size n filled initially with 1, i.e., all are primes, and iteratively zero out all multiples of numbers up to \sqrt{n} .

```
int res[n] = {0, 0, 1, 1, 1, ..., 1}
for(i = 2; i <= sqrt(n); i++) { //sequential
    if (res[i] != 0) {
        forall m ∈ multiples of i ≤ n do {
            res[m] = 0;
        }
    }
}
```

Work: $O(n \lg \lg n)$ but Depth: $O(\sqrt{n})$ (Not Good Enough!)

Eratosthenes Algorithm Improved for Parallel Execution

If we have all primes from 2 to \sqrt{n} we could generate all multiples of these primes (up to n) at once: `{[2*p:n:p]: p in sqr_primes}` in NESL. Also call algorithm recursively on \sqrt{n} \Rightarrow Depth: $O(\lg \lg n)$ (solution of $n^{(1/2)^{\text{depth}}} = 2$). Work: $O(n \lg \lg n)$

```
primesOpt :: Int -> [Int]
primesOpt n =
  if n <= 2 then [2]
  else
    let sqrtN = floor (sqrt (fromIntegral n))
        sqrt_primes = primesOpt sqrtN
        nested = map (\p->let m = (n `div` p)
                      in map (\j-> j*p)
                             [2..m])
                     ) sqrt_primes
    not_primes = reduce (++) [] nested
    mm = length not_primes
    zeros = replicate mm False
    prime_flags=scatter(replicate (n+1) True)
                not_primes zeros
    (primes,_)= unzip $ filter (\(i,f)->f)
                $ (zip [0..n] prime_flags)
  in drop 2 primes
```

Batch of Rank-Search K Problems

Rank-Search k: finds the k^{th} smallest element of a vector.

Typically used for median computation.

```
let rankSearch (k: i64) (A: []f32) : f32 =
    let p = random_element A
    let A_lth_p = filter (< p) A
    let A_eqt_p = filter (==p) A
    let A_gth_p = filter (> p) A

    if (k <= A_lth_p.length)
    then rankSearch k A_lth_p
    else if (k <= A_lth_p.length + A_eqt_p.length)
        then p
        else rankSearch (k - A_lth_p.length - A_eqt_p.length) A_gth_p

let main [m] (ks: [m]i64) (As: [m][]f32) : [m]f32 =
    map2 rankSearch ks As
```

Quicksort with Nested Parallelism

```
nestedQuicksort :: [a] -> [a]
nestedQuicksort arr =
  if (length arr) <= 1 then arr else
    let i = getRand (0, (length arr) - 1)
      a = arr !! i
      s_< = filter (\ x -> x < a) arr
      s_= = filter (\ x -> x == a) arr
      s_> = filter (\ x -> x > a) arr
      rs = map nestedQuicksort [s_<, s_>]
    in (rs !! 0) ++ s_= ++ (rs !! 1)
```

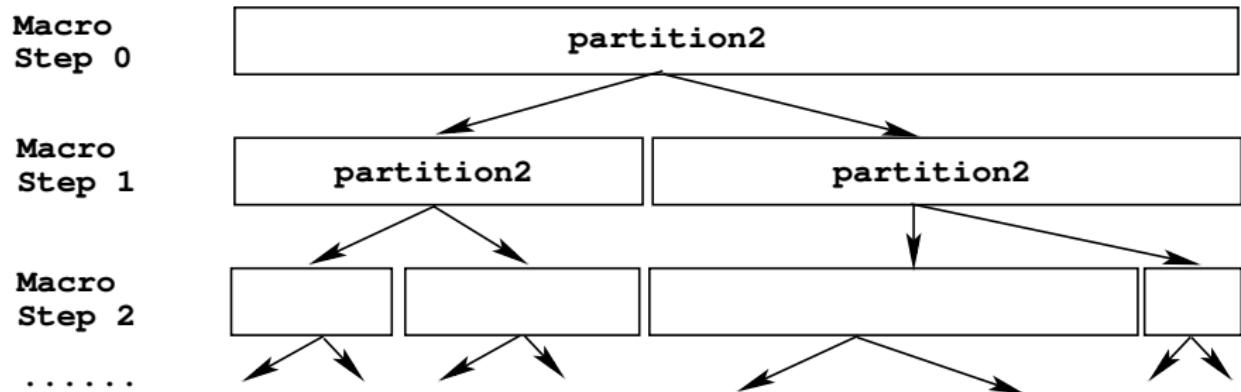
Using n for the input's length: Average Work is $O(n \lg N)$.

If filter would have depth 1, then Average Depth: $O(\lg n)$.

In practice we have depth: $O(\lg^2 n)$.

In principle, the implementation can be re-structured to use one `partition2` instead of three filters.

Quicksort: Illustrating Flat-Parallel Execution



QuickHull with Nested Parallelism

Algorithm 1 QuickHull

Require: S : a set of $n \geq 2$ two-dimensional points

Ensure: CH : the convex-hull set of S

$(A, B) =$ the leftmost
and rightmost
points of S

$S_{1,2} =$ points of S above
and below line AB

$CH = \{A, B\} \cup$
 $findHull(S_1, A, B) \cup$
 $findHull(S_2, A, B)$

QuickHull with Nested Parallelism

Algorithm 2 Divide-And-Conquer Helper

```
1: procedure findHull( $S, P, Q$ )
2:    $Hull = \emptyset$ 
3:   if  $S \neq \emptyset$  then
4:      $C =$  furthest point of  $S$  from line  $PQ$ 
5:      $(S_l, S_r) =$  the points of  $S$  on the left-
6:                   and right-hand side of lines
7:                    $CP$  and  $CQ$ , respectively
8:                   (and not inside  $\Delta PCQ$ )
9:      $Hull = \{C\} \cup$ 
10:     $findHull(S_l, P, C) \cup$ 
11:     $findHull(S_r, C, Q)$ 
12:   return  $Hull$ 
```

Parallel Basic Blocks Recap

Part I: Flattening Nested and Irregular Parallelism

What is "Flattening"? Recipe for Applying Flattening

Several Re-Write Rules (inefficient for replicate & iota)

Jagged (Irregular Multi-Dim) Array Representation

Revisiting the Rewrites for Replicate & Iota Nested Inside Map

Revisiting the Solution to Our Example

Part II: Flattening Nested and Irregular Parallelism

Several Applications of Flattening

More Flattening Rules

Flattening by Function Lifting

Flattening Quicksort

Flattening Prime-Number (Sieve) Computation

"To Flatten or Not To Flatten, that is the question"

Nested vs Flattened Parallelism: Reduce Inside Map

(5) Reduce Inside a Map or Segmented Reduce:

```
let arr = [[1, 3, 4], [6, 7]] in
let res = map (\x -> reduce (+) 0 x) arr
-- should result in [8, 13]
```

Nested vs Flattened Parallelism: Reduce Inside Map

(5) Reduce Inside a Map or Segmented Reduce:

```
let arr = [[1, 3, 4], [6, 7]] in
let res = map (\x -> reduce (+) 0 x) arr
-- should result in [8, 13]
```

translates to a **scan-pack** composition:

1. the length of `res` equals the number of subarrays of `arr`;
2. the shape of `arr` is scanned: the result records the position of the last element in a segment plus one;
3. segmented scan is applied on the input array: the last elem in a segment holds the reduced value of the segment;
4. segment's last element is extracted by a map operation.

```
F(res = map (\row -> reduce ⊕ 0 ⊙ row) arr) ⇒
-- S0arr = [2], S1arr = [3,2], Farr = [1,0,0,1,0], Darr = [1,3,4,6,7]
1. S0res = S0arr                                     -- S0res = [2]
2. indsp1 = scan (+) 0 S1arr                  -- indsp1 = [3, 5]
3. tmp = sgmScan (⊕) 0 ⊙ Farr Darr          -- tmp = [1, 4, 8, 6, 13]
4. Dres = map2(\s ip1 -> if s ≤ 0 then 0 ⊙
                           else tmp[ip1-1]) S1arr indsp1 -- Dres = [8, 13]
```

Nested vs Flattened Parallelism: Reduce Inside Map

(5) Reduce Inside a Map or Segmented Reduce:

```
let arr = [[1, 3, 4], [6, 7]] in
let res = map (\x -> reduce (+) 0 x) arr
-- should result in [8, 13]
```

Nested vs Flattened Parallelism: Reduce Inside Map

(5) Reduce Inside a Map or Segmented Reduce:

```
let arr = [[1, 3, 4], [6, 7]] in  
let res = map (\x -> reduce (+) 0 x) arr  
-- should result in [8, 13]
```

We can also “cheat” and use a histogram-like computation

Nested vs Flattened Parallelism: Reduce Inside Map

(5) Reduce Inside a Map or Segmented Reduce:

```
let arr = [[1, 3, 4], [6, 7]] in
let res = map (\x -> reduce (+) 0 x) arr
-- should result in [8, 13]
```

We can also “cheat” and use a histogram-like computation

```
F(res = map (\row -> reduce ⊕ 0 ⊙ row) arr) ⇒
-- Sarr0 = [2], Sarr1 = [3, 2], Farr = [1, 0, 0, 1, 0], Darr = [1, 3, 4, 6, 7]
1. Sres0 = Sarr0 -- Sres0 = [2]
2. Dres = hist (⊕) 0 ⊙ (Sarr0[0]) IIarr1 Darr
```

How else can one try to optimize this code by hand?

- practical performance refers to how many global-memory accesses you perform
- accessing II_{arr}^1 from memory has significant cost
- in some practical cases, it might be more efficient to not manifest II_{arr}^1 , but instead to compute its elements by binary searching the B_{arr}^1 array.

Flattening Scatter and Histogram

How does one flattens a scatter perfectly nested inside a map?

How does one flattens a histogram perfectly nested inside a map?

Flattening Scatter and Histogram

How does one flattens a scatter perfectly nested inside a map?

How does one flattens a histogram perfectly nested inside a map?

You will have to answer it yourselves as part of the third weekly assignment :)

Treating a Scalar Variant to the Outer Map

(6) The inner construct uses a scalar variant to the outer map:

```
let res = map2 (\x ys -> map (+x) ys) [1,3] [[4,5,6], [9,7]] ≡  
let res = [map (+1) [4,5,6], map (+3) [9,7]]  
let res = [ [5,6,7], [12,10] ]
```

Treating a Scalar Variant to the Outer Map

(6) The inner construct uses a scalar variant to the outer map:

```
let res = map2 (\x ys -> map (+x) ys) [1,3] [[4,5,6], [9,7]] ≡  
let res = [map (+1) [4,5,6], map (+3) [9,7]]  
let res = [ [5,6,7], [12,10] ]
```

Traditionally, this is handled by expanding (replicating) each x across the whole segment

```
let Dxss = [1, 1, 1, 3, 3]  
let res = map2 (+) [1, 1, 1, 3, 3]  
          [4, 5, 6, 9, 7]  
          = = = = =  
          [5, 6, 7, 12, 10]
```

Instead, we use II_{arr}^1 to indirectly access in the xs array:

```
F(res = map2 (\x ys -> map (f x) ys) xs yss) ⇒  
-- xs = [1,3], Syss1 = [3,2], Fyss = [1,0,0,1,0], Dyss = [4,5,6,9,7]  
1. Sres1 = Syss1  
2. Dres = map2 (\y sgmind -> f xs[sgmind] y ) Dyss IIyss1  
-- IIyss1 = [0,0,0,1,1], Dres = [5,6,7,12,10]
```

Treating Indexing Variant to the Outer Map

(7) Indexing Operations Variant to the Outer Map:

```
let res = map2 (\i xs -> xs[i]) [2,0] [[4,5,6], [9,7]] ≡  
let res = [ 6, 9 ]
```

Treating Indexing Variant to the Outer Map

(7) Indexing Operations Variant to the Outer Map:

```
let res = map2 (\i xs -> xs[i]) [2,0] [[4,5,6], [9,7]] ≡  
let res = [ 6, 9 ]
```

To corresponding flat index in D_{yss} is obtained by summing up

- the start offset of every segment, which we get from B_{yss}^1 , and
- the index inside the segment, which we get from i

$\mathcal{F}(res = \text{map2 } (\backslash i \ xs \ -> \ xs[i]) \ i \ s \ xss) \Rightarrow$
-- $i \ s = [2,0]$, $S_{xss}^1 = [3,2]$, $B_{xss}^1 = [0,3]$, $D_{xss} = [4,5,6,9,7]$
1. $S_{res}^0 = S_{is}^0$ -- $= S_{is}^0 = [2]$
2. $D_{res} = \text{map2 } (\backslash \ off \ i \ -> \ D_{xss}[\text{off} + i]) \ B_{xss}^1 \ i \ s \ -- \ D_{res} = [6, 9]$

Nested vs Flattened Parallelism: If Inside a Map 2D Case

(8) If-Then-Else with inner parallelism nested inside a map:

```
bs  = [F,T,F,T]
xss = [[1,2,3],[4,5,6,7],[8,9],[10]]
res = map(\b xs -> if b  then map (+1) xs  else map (*2) xs) bs xss
res = [ map(*2)[1,2,3], map(+1)[4,5,6,7], map(*2)[8,9], map(+1)[10] ]
res = [ [2,4,6], [5,6,7,8], [16,18], [11] ]
```

Nested vs Flattened Parallelism: If Inside a Map 2D Case

(8) If-Then-Else with inner parallelism nested inside a map:

```
bs  = [F,T,F,T]
xss = [[1,2,3],[4,5,6,7],[8,9],[10]]
res = map(\b xs -> if b  then map (+1) xs  else map (*2) xs) bs xss
res = [ map(*2)[1,2,3], map(+1)[4,5,6,7], map(*2)[8,9], map(+1)[10] ]
res = [ [2,4,6], [5,6,7,8], [16,18], [11] ]
```

translates to a scatter-map-gather composition. Intuition:

1. compute *iinds*, the permutation of segments w.r.t. *bs*;
 - 2-3. partition the *xss* array based on *bs*;
 - 4-5. **flatten outer map and/on top of the parallel code of the then and else branches;**
 6. inverse permute the resulted segments according to *iinds*.
- ```
1. iinds = partition2 (\i -> bs[i]) (iota (length b)) -- [1,3,0,2]
2. xssthen = gatherThen iinds xss -- ([4,1], [4,5,6,7, 10])
3. xsselse = gatherElse iinds xss -- ([3,2], [1,2,3, 8,9])
-- Recursively Flatten the Then and Else Branches!
4. resthen = F(map (map (+1)) xssthen) -- ([4,1], [5,6,7,8, 11])
5. reselse = F(map (map (*2)) xsselse) -- ([3,2], [2,4,6,16,18])
6. res = inversePermute iinds (resthen++reselse)
-- ([3,4,2,1], [2,4,6, 5,6,7,8, 16,18, 11])
```

# Nested vs Flattened Parallelism: If Inside a Map 2D Case

## (8) If-Then-Else with inner parallelism nested inside a map:

```
bs = [F,T,F,T], xss = [[1,2,3],[4,5,6,7],[8,9],[10]], Sxss1=[3,4,2,1], f=map (+1), g=map (*2)

F(res = map2 (\b xs -> if b then f xs else g xs) bs xss) =>
(spl, iinds) = partition2 bs (iota (length bs)) -- (2, [1,3,0,2])
(Sxss1then, Sxss1else) = split spl (map (\ii -> Sxss1[ii]) iinds)--([4,1],[3,2])
maskxss = map (\sgmind -> bs[sgmind]) ||xss1 -- [F,F,F,T,T,T,T,F,F,T]
(brk, DxssP) = partition2 maskxss Dxss
(Dxssthen, Dxsselse) = split brk DxssP -- ([4,5,6,7,10],[1,2,3,8,9])
(Sres1then, Dresthen) = F(map f) (Sxss1then, Dxssthen) -- ([4,1], [5,6,7,8,11])
(Sres1else, Dreselse) = F(map g) (Sxss1else, Dxsselse) -- ([3,2], [2,4,6,16,18])
Sres1P = Sres1then ++ Sres1else -- [4,1,3,2]
Sres1 = scatter (replicate (length bs) 0) iinds Sres1P -- [3,4,2,1]
Bres1 = scanexc (+) 0 Sres1 -- [0,3,7,9]
FresP = mkFlagArray Sres1P 0 (map (+1) iinds) -- [2,0,0,0,4,1,0,0,3,0]
||res1P = sgmscan (+) 0 FresP FresP>> map(\x -> x-1) -- [1,1,1,1,3,0,0,0,2,2]
||res2P = ||resthen2 ++ ||reselse2 -- [0,1,2,3,0, 0,1,2,0,1]
sindsres = map2 (\sgm iin -> Bres1[sgm] + iin) ||res1P ||res2P
-- [3+0,3+1,3+2,3+3, 9+0, 0+0,0+1,0+2, 7+0,7+1]=[3,4,5,6,9,0,1,2,7,8]
Dres = scatter (replicate flenres 0) sindsres (Dresthen ++ Dreselse)
-- [2,4,6, 5,6,7,8, 16,18, 11]
(Sres1, Dres)
```

# Nested vs Flattened Parallelism: Do Loop Inside a Map

## (9) Flattening a Do Loop Nested Inside a Map:

- compute the maximal loop count  $n_{max}$
- interchange the loop and the map:
  - ▶ loop count becomes  $n_{max}$
  - ▶ the loop body is wrapped inside a `if i < n` condition, and
  - ▶ the new loop body is flattened!

```
 $\mathcal{F}(\text{res} = \text{map2 } (\backslash n \text{ xs} \rightarrow \text{loop (xs) for } i < n \text{ do } f \text{ xs}) \text{ ns } xss) \Rightarrow$
1. $n_{max} = \text{reduce max } 0i32 \text{ ns}$
2. $g \text{ i m arr} = \text{if } i < m \text{ then } f \text{ arr else arr}$
3. $\text{loop}(S_{xss}^1, D_{xss}) \text{ for } i < n_{max} \text{ do}$
4. $\mathcal{F}(\text{map2 } (g \text{ i})) \text{ ns } (S_{xss}^1, D_{xss})$
5. $-- (g i)^L \text{ ns } (S_{xss}^1, D_{xss})$
```

But this treatment does not necessarily preserve the work asymptotic ... what to do?

# Nested vs Flattened Parallelism: Do Loop Inside a Map

## (9) Flattening a Do Loop Nested Inside a Map:

- compute the maximal loop count  $n_{max}$
- interchange the loop and the map:
  - ▶ loop count becomes  $n_{max}$
  - ▶ the loop body is wrapped inside a `if i < n` condition, and
  - ▶ the new loop body is flattened!

```
F(res = map2 (\n xs -> loop (xs) for i < n do f xs) ns xss) =>
1. nmax = reduce max 0i32 ns
2. g i m arr = if i < m then f arr else arr
3. loop (Sxss1, Dxss) for i < nmax do
4. F(map2 (g i)) ns (Sxss1, Dxss)
5. -- (g i)L ns (Sxss1, Dxss)
```

But this treatment does not necessarily preserve the work asymptotic ... what to do?

If the size of the result can be deduced/inferred:

- allocate the flat-result array before the loop
- filter out the empty segments, and make the loop iterate until the shape is empty
- each time a segment finish execution (i) it is scattered into the result, and (2) it is filtered out from the running set of segments.

# Nested vs Flattened Parallelism: Do Loop Inside a Map

## (9) Flattening a Do Loop Nested Inside a Map:

The general case can be solved by the technique of  
“reducing it to a more challenging/general problem” :))

A loop such as

```
loop (xs) = (xs0) while goOn xs do f xs
```

is equivalent with a call to a tail recursive function

```
fTailRec xs0
 where
 fTailRec xs = if goOn xs
 then f xs
 else xs -- base case
```

We already know how to flatten an if-then-else; if we figure out how to flatten a function called directly inside a map, we are done

```
F(res = map (\xs0 -> loop (xs) = (xs0) while goOn xs do f xs) xss0)
≡
F(res = map (\xs0 -> fTailRec xs0) xss0)
```

## Parallel Basic Blocks Recap

### Part I: Flattening Nested and Irregular Parallelism

What is "Flattening"? Recipe for Applying Flattening

Several Re-Write Rules (inefficient for replicate & iota)

Jagged (Irregular Multi-Dim) Array Representation

Revisiting the Rewrites for Replicate & Iota Nested Inside Map

Revisiting the Solution to Our Example

### Part II: Flattening Nested and Irregular Parallelism

Several Applications of Flattening

More Flattening Rules

**Flattening by Function Lifting**

Flattening Quicksort

Flattening Prime-Number (Sieve) Computation

"To Flatten or Not To Flatten, that is the question"

## Flattening by Function Lifting: Basic Idea

Assume a simple function f:

```
let f (x: i32) : i32 = x + 1
```

f lifted, denoted  $f^L$  semantically corresponds to map f, where the arguments have been expanded to an extra array dimension, and the inner operators/functions have also been lifted:

```
let +L [n] (as: [n]i32) (bs: [n]i32) : [n]i32 =
 map2 (+) as bs
```

```
let fL [n] (xs: [n]i32) : [n]i32 =
 xs +L (replicate n 1)
```

# Flattening by Function Lifting: Basic Idea

Assume a simple function  $f$ :

```
let f (x: i32) : i32 = x + 1
```

$f$  lifted, denoted  $f^L$  semantically corresponds to map  $f$ , where the arguments have been expanded to an extra array dimension, and the inner operators/functions have also been lifted:

```
let +L [n] (as: [n]i32) (bs: [n]i32) : [n]i32 =
 map2 (+) as bs
```

```
let fL [n] (xs: [n]i32) : [n]i32 =
 xs +L (replicate n 1)
```

- Locals such as  $x \Rightarrow$  left alone
- Global such as  $+$   $\Rightarrow$  lifted ( $+^L$ )
- Constants such as  $k \Rightarrow$  replicate (length  $xs$ )  $k$ 
  - ▶ good for vectorization, bad for locality, asymptotics
  - ▶ for GPU better to indirectly index into a smaller array, rather than replicate.

# Flattening by Function Lifting: Key Insight!

```
let f (xs: []f32) : [][]f32 = map g xs -- = gL xs
let fL (xss: [][]f32) : [][][]i32 = (gL)L -- ???
```

How do we stop lifting?  $g$  and  $g^L$  are enough: no need for  $(g^L)^L$ !

# Flattening by Function Lifting: Key Insight!

```
let f (xs: []f32) : [][]f32 = map g xs -- = gL xs
let fL (xss: [][][]f32) : [][][]i32 = (gL)L -- ???
```

How do we stop lifting?  $g$  and  $g^L$  are enough: no need for  $(g^L)^L$ !

```
let f (xs: []f32) : [][]f32 = map g xs -- = gL xs
-- in nested parallel form
let fL (xss: [][][]f32) : [][][]i32 =
 segment xss (gL (concat xss))
```

In Haskell Notation:

```
concat :: [[a]] -> [a]
segment :: [[a]] -> [b] -> [[b]]
 shape flat data nested data
```

# Flattening by Function Lifting: General Case!

```
let f (xs: []f32) : []...[]f32 = map g xs -- = gL xs
let fL (xss: [][]f32) : [][]...[]f32 = (gL)L -- ???
```

How do we stop lifting?  $g$  and  $g^L$  are enough: no need for  $(g^L)^L$ !

A 3D array  $rsss$ :  $[][][]f32$  has the representation  $(S_{rsss}^0, S_{rsss}^1, S_{rsss}^2, D_{rsss})$

```
let f (xs: []f32) : []...[]f32 = map g xs -- = gL xs
```

-- in nested parallel form

```
let fL (xss: [][]f32) : [][]...[]f32 =
 segment xss (gL (concat xss))
```

-- in flatten form

```
let fL (Sxss0: i64, Sxss1: []i64, Dxss: []f32)
 : (i64, []i64, ..., []i64, []f32) =
let (Sfrsss0, ..., Sfrsssq, Drss) = gL (Sxss0, Sxss1, Dxss)
let (Srsss0, Srsss1, ..., Srsssq) = (Sxss0, Sxss1, ..., Sfrsssq)
in (Srsss0, Srsss1, ..., Srsssq, Drss)
```

In Haskell Notation:

```
concat :: [[a]] -> [a]
segment :: [[a]] -> [b] -> [[b]]
 shape flat data nested data
```

## Parallel Basic Blocks Recap

### Part I: Flattening Nested and Irregular Parallelism

What is "Flattening"? Recipe for Applying Flattening

Several Re-Write Rules (inefficient for replicate & iota)

Jagged (Irregular Multi-Dim) Array Representation

Revisiting the Rewrites for Replicate & Iota Nested Inside Map

Revisiting the Solution to Our Example

### Part II: Flattening Nested and Irregular Parallelism

Several Applications of Flattening

More Flattening Rules

Flattening by Function Lifting

**Flattening Quicksort**

Flattening Prime-Number (Sieve) Computation

"To Flatten or Not To Flatten, that is the question"

# Recounting Quicksort

Recount the classic nested-parallel definition:

```
let quicksort [n] (arr : [n]f32) : [n]f32 =
 if n < 2 then arr else
 let i = getRand (0, (length arr) - 1)
 let a = arr[i]
 let s1 = filter (< a) arr
 let s2 = filter (== a) arr
 let s3 = filter (> a) arr
 in (quicksort s1) ++ s2 ++ (quicksort s3)
-- can be re-written as:
-- rs = map nestedQuicksort [s1, s3]
-- in (rs[0]) ++ s2 ++ (rs[1])
```

Note: Futhark does not support recursive calls, hence not valid code!

# Nested-Parallel Quicksort Simplified

For simplicity we will rewrite it in terms of partition2:

```
let isSorted [n] (as: [n]f32) : bool =
 map (\i -> if i==0 then true else as[i-1] < as[i])) (iota n)
 |> reduce (&&) true

let quicksort [n] (arr: [n]f32) : [n]f32 =
 if isSorted arr then arr else
 let i = getRand (0, (length arr) - 1)
 let a = arr[i]
 let bs = map (< a) arr
 let (q, arr') = partition2 bs 0.0f32 arr
 let (arr<, arr≥) = split q arr'
 in concat <| map quicksort [arr<, arr≥]
```

Note: Futhark does not support recursive calls, irregular map operation, or concat!

# Partition2

Reorders the elements of an array such that those that correspond to a true mask come before those corresponding to false.

```
let partition2 [n] 't (conds: [n]bool) (dummy: t) (arr: [n]t)
 : (i32, [n]t) =
let tflgs = map (\ c -> if c then 1 else 0) conds
let fflgs = map (\ b -> 1 - b) tflgs

let indsT = scan (+) 0 tflgs
let tmp = scan (+) 0 fflgs
let lst = if n > 0 then indsT[n-1] else -1
let indsF = map (+ lst) tmp

let inds = map3 (\ c indT indF -> if c then indT-1 else indF-1)
 conds indsT indsF

let fltarr= scatter (replicate n dummy) inds arr
in (lst , fltarr)
```

For example:

```
conds = [F,T,F,T,F,F,T]
xss = [1,2,3,4,5,6,7]
partition2 conds 0 xss => (3, [2,4,7,1,3,5,6])
```

# Lifting Quicksort

**Key Idea: write a function with the semantics of  
map nestedQuicksort, i.e., it operates on array of arrays.**

```
let isSorted [n] (as: [n]f32) : bool =
 map (\i -> if i==0 then true else as[i-1] < as[i])) (iota n)
 |> reduce (&&) true

let quicksortL (xss: [][]f32) : [][]f32 =
 map (\xs ->
 if isSorted xs then xs else -- oh, nooooo, an if-then-else!
 let i = getRand (0, (length xs) - 1)
 let a = xs[i]
 let bs = map (< a) xs
 let (q, xsP) = partition2 bs 0.0f32 xs
 let (xs<, xs≥) = split q xsP
 in concat <| map quicksort [xs<, xs≥]
) xss
```

Important observations:

- $\text{map quicksort} \equiv \text{quicksort}^L$
- the flat data of  $[\text{xs}_<, \text{xs}_{\geq}] \equiv \text{xs}^P$ , the result of  $\text{partition2}$
- $\text{map}(\text{map quicksort}) \equiv \text{quicksort}^{L^L} \equiv \text{segment o quicksort}^{L^0} \text{o concat}$

# Lifting Quicksort

Let us treat the last three lines from the previous implem.:

```
let quicksortL (S1xss :[] i32 , Dxss :[] f32): ([] i32 ,[] f32) = --(xss: [][])
 if isSorted Dxss then (S1xss :[] i32 , Dxss :[] f32) else -- big cheat!
 let (S1bss , Dbss) = F (
 map (\ xs ->
 let i = getRand (0 , (length xs) - 1)
 let a = xs[i]
 let bs = map (< a) xs
 in bs
) xss
)
 let (ps , (S1xssp ,Dxssp)) = partition2L Dbss 0.0f32 (S1xss , Dxss)
 -- Invariant: S1xssp == S1bss == S1xss
 let S1[xss<,xss≥] = filter (!=0) <| flatten <|
 map2 (λ p s -> if s==0 then [0,0] else [p,s-p]) ps S1xss
 in quicksortL (S1[xss<,xss≥], Dxssp)
```

- $S_{[xss<,xss\geq]}^1$  is the shape of  $[xss<, xss\geq]$
- $(\text{concat } <| \text{quicksort}^L)^L xsss \equiv \text{concat } <| \text{segment } xsss <| \text{quicksort}^L (\text{concat } xsss) \equiv \text{quicksort}^L (\text{concat } xsss)$
- The function looks tail recursive now: let's replace it with a loop!

# Lifting Quicksort: Final Implementation

```
let quicksortL [m][n] (S1xss:[m]i32 , Dxss:[n]f32): [n]f32 =
 let (stop, count) = (isSorted Dxss, 0i32)
 let (_,res,_,_) =
 loop(S1xss, Dxss, stop, count) while (!stop) do
 -- compute helper-representation structures
 let B1xss = scanexc (+) 0 S1xss
 let F1xss = mkFlagArray S1xss 0i32 <| map (+1) <| iota m
 let II1xss = sgmscan (+) 0 F1xss <|
 map (\f -> if f==0 then 0 else f-1) F1xss
 -- flattening quicksort:
 let rL = map (\u -> randomInd (0,u-1) count) S1xss
 let aL = map3(\r l i-> if l <= 0 then 0.0 else Dxss[B1xss[i]+r]
) rL S1xss (iota m)
 let Dbss= map2 (\x sgmind -> aL[sgmind] > x) Dxss II1xss
 let (ps, (S1xssp,Dperxss)) = partition2L Dbss 0.0f32 (S1xss, Dxss)
 let S1[xss<,xss≥] = filter (!=0) <| flatten <|
 map2 (\ p s -> if s==0 then [0,0] else [p,s-p]) ps S1xss
 in (S1[xss<,xss≥], Dperxss, isSorted Dperxss, count+1)
 in res
```

PPF Weekly 2 Exercise: Implement partition2<sup>L</sup>

## Parallel Basic Blocks Recap

### Part I: Flattening Nested and Irregular Parallelism

What is "Flattening"? Recipe for Applying Flattening

Several Re-Write Rules (inefficient for replicate & iota)

Jagged (Irregular Multi-Dim) Array Representation

Revisiting the Rewrites for Replicate & Iota Nested Inside Map

Revisiting the Solution to Our Example

### Part II: Flattening Nested and Irregular Parallelism

Several Applications of Flattening

More Flattening Rules

Flattening by Function Lifting

Flattening Quicksort

Flattening Prime-Number (Sieve) Computation

"To Flatten or Not To Flatten, that is the question"

# How Does One Flattens Prime Numbers?

**The important bit with nested parallelism:**

```
sqrt_primes = primesOpt (sqrt (fromIntegral n))
nested = map (\p -> let m = (n `div` p)
 in map (\j -> j*p) [2..m]
) sqrt_primes
not_primes = reduce (++) [] nested
```

# How Does One Flattens Prime Numbers?

## The important bit with nested parallelism:

```
sqrt_primes = primes0pt (sqrt (fromIntegral n))
nested = map (\p -> let m = (n `div` p)
 in map (\j -> j*p) [2..m]
) sqrt_primes
not_primes = reduce (++) [] nested
```

## Normalize the nested map:

```
sqrt_primes = primes0pt (sqrt (fromIntegral n))
nested = map (\p ->
 let m = n `div` p in -- distribute map
 let mm1 = m - 1 in -- distribute map
 let iot = iota mm1 in -- F rule 4
 let twom= map (+2) iot in -- F rule 2
 let rp = replicate mm1 p in -- F rule 3
 in map (\(j,p) -> j*p) (zip twom rp) -- F rule 2
) sqrt_primes
not_primes = reduce (++) [] nested -- ignore, already flat
```

Flattening PrimeOpt was part of PMPH's Weekly Assignment 2!

## Parallel Basic Blocks Recap

### Part I: Flattening Nested and Irregular Parallelism

What is "Flattening"? Recipe for Applying Flattening

Several Re-Write Rules (inefficient for replicate & iota)

Jagged (Irregular Multi-Dim) Array Representation

Revisiting the Rewrites for Replicate & Iota Nested Inside Map

Revisiting the Solution to Our Example

### Part II: Flattening Nested and Irregular Parallelism

Several Applications of Flattening

More Flattening Rules

Flattening by Function Lifting

Flattening Quicksort

Flattening Prime-Number (Sieve) Computation

"To Flatten or Not To Flatten, that is the question"

# The Good: QuickHull and the Like

Flattening is perhaps the only way to go for achieving decent GPU performance for a set of challenging problems such as Quickhull:

|            | Circle |      |       | Rectangle |       |       | Quadratic |      |      |
|------------|--------|------|-------|-----------|-------|-------|-----------|------|------|
|            | CPU    |      | GPU   | CPU       |       | GPU   | CPU       |      | GPU  |
|            | 1C     | 32C  |       | 1C        | 32C   |       | 1C        | 32C  |      |
| Baseline   | 4.42   | 0.20 | —     | 3.36      | 0.11  | —     | 35.1      | 2.92 | —    |
| Accelerate | 7.39   | 1.57 | 0.160 | 3.60      | 1.175 | 0.114 | 48.4      | 12.5 | 4.28 |
| APL        | 22.2   | —    | 1.22  | 14.9      | —     | 0.690 | 113       | —    | 7.57 |
| DaCe       | —      | —    | —     | —         | —     | —     | —         | —    | —    |
| Futhark    | 5.56   | 1.28 | 0.064 | 3.81      | 1.151 | 0.047 | 37.6      | 4.03 | 0.68 |
| SaC        | 13.3   | —    | —     | 13.2      | —     | —     | 18.3      | —    | —    |

The languages that do not support scan or parallel write as primitives (DaCe and SAC) could not express it.

# Sparse Matrix Vector Multiplication: Flattened Kernel

Dense-Matrix ( $A \in \mathbb{R}^{m \times q}$ ) - Vector ( $V \in \mathbb{R}^q$ ) Multiplication:

$$X_i = \sum_{j=0 \dots q-1} A_{i,j} \times V_j$$

Sparse-matrix uses a CSR representation:

- $B$  vector records the start of each row
- $A$  flat array that tuples each non-zero element with its corresponding column index
- For simplicity we assume that all rows are non empty

$$X_i = \sum_{j=B[i] \dots B[i+1]} A_j.value \times V_{A_j.colidx}$$

Using the flattening rule for reduce nested inside of map results in:

```
def spMatVecMulFlats [m][flen][q] (B: [m]u32, spmat: [flen](u32,f32))
 (vec: [q]f32) : [m]f32 = #[unsafe]
 let flags = scatter (replicateflen false) (map i64.u32 B)
 (replicatem true)
 let scn_mat = map (\(ind,elm) -> elm * vec[i64.u32 ind]) spmat
 |> sgmScan (+) 0f32 flags prods
 in tabulate m
 (\i -> if i == m-1 then last scn_mat
 else scn_mat[i64.u32 (B[i+1]-1)])
```

# Sparse Matrix Vector Multiplication: Other Heuristics

Sparse-matrix Vector Multiplication:  $X_i = \sum_{j=B[i]\dots B[i+1]} A_j.value \times V_{A_j.colidx}$

A kernel that exploit only the outer parallelism, i.e., each thread process a matrix row:

```
def spMatVecMulOuter [m][flen][q] (B: [m]u32, spmat: [flen](u32,f32))
 (vec: [q]f32) : [m]f32 = #[unsafe]
let f i = let beg = i64.u32 B[i]
 let end = if i == m-1 thenflen else i64.u32 B[i+1]
 in loop sum = 0 for i < end - beg do
 let (j,v) = spmat[i + beg]
 in sum + v*vec[i64.u32 j]
in map f (iota m)
```

# Sparse Matrix Vector Multiplication: Other Heuristics

Sparse-matrix Vector Multiplication:  $X_i = \sum_{j=B[i]\dots B[i+1]} A_j.value \times V_{A_j.colidx}$

A kernel that exploit only the outer parallelism, i.e., each thread process a matrix row:

```
def spMatVecMulOuter [m][flen][q] (B: [m]u32, spmat: [flen](u32,f32))
 (vec: [q]f32) : [m]f32 = #[unsafe]
 let f i = let beg = i64.u32 B[i]
 let end = if i == m-1 thenflen else i64.u32 B[i+1]
 in loop sum = 0 for i < end - beg do
 let (j,v) = spmat[i + beg]
 in sum + v*vec[i64.u32 j]
 in map f (iota m)
```

A kernel that utilizes  $m \cdot 64$  parallelism: each CUDA block of 64 threads processes a row:

```
def spMatVecMulMidpt [m][flen][q] (B: [m]u32, spmat: [flen](u32,f32))
 (vec: [q]f32) : [m]f32 = #[unsafe]
 let block = 64i64 -- should use a better heuristic based on B
 let f i = let beg = i64.u32 B[i]
 let end = if i == m-1 thenflen else i64.u32 B[i+1]
 let g tid = loop sum=0 for i < (end-beg+block-1)/block do
 let ind = beg + i*block + tid in
 if ind >= end then sum
 else sum + spmat[ind].1 *
 vec[i64.u32 spmat[ind].0]
 in iota block |> map g |> reduce (+) 0f32 sums
 in #[incremental_flattening(only_intra)] map f (iota m)
```

# Performance of Sparse-Matrix Vector Multiplication

## PERFORMANCE DEMONSTRATION

# Performance of Sparse-Matrix Vector Multiplication

## PERFORMANCE DEMONSTRATION

The midpoint regular kernel commonly offers best performance,  
i.e., the one processing a row in a CUDA block of 64 threads!

# Sparse-Matrix Dense-Matrix Multiplication: Kernels

Dense Matrix ( $A \in \mathbb{R}^{m \times q}$ ) - Matrix ( $B \in \mathbb{R}^{q \times n}$ ) & Sparse-Dense Matrix Multiplication:

$$X_{i,j} = \sum_{k=0 \dots q-1} A_{i,k} \times B_{k,j}$$

$$X_{i,j} = \sum_{k=B[i] \dots B[i+1]} A_k.value \times B_{A_k.colidx,j}$$

**Dense-Dense Matrix Multiplication follows the classical implementation:**

```
def denseMMM [m][n][q] (ass: [m][q]f32) (bss: [q][n]f32) : [m][n]f32=
 let dotprod xs ys = map2 (*) xs ys |> reduce (+) 0
 in map (\as -> map (dotprod as) (transpose bss)) ass
```

**Sparse-Dense utilizing the kernel obtained by flattening:**

```
def spMMFlatS [m][flen][q][n] (B: [m]u32) (spmat: [flen](u32,f32)
 (dense: [q][n]f32) : [m][n]f32 =
 transpose dense |> map (spMatVecMulFlatS (B, spmat)) |> transpose
```

**Sparse-Dense utilizing the kernel exploiting the outer parallelism:**

```
def spMMOuter [m][flen][q][n] (B: [m]u32) (spmat: [flen](u32,f32)
 (dense: [q][n]f32) : [m][n]f32 =
 transpose dense |> map (spMatVecMulOuter (B, spmat)) |> transpose
```

**Sparse-Dense utilizing the kernel exploiting midpoint parallelism:**

```
def spMMOuter [m][flen][q][n] (B: [m]u32) (spmat: [flen](u32,f32)
 (dense: [q][n]f32) : [m][n]f32 =
 transpose dense |> map (spMatVecMulMidpt (B, spmat)) |> transpose
```

# Performance of Sparse-Dense Matrix Multiplication

## PERFORMANCE DEMONSTRATION

## PERFORMANCE DEMONSTRATION

- Flattened version performs the worst but it is useful to protect against degenerate cases, e.g., few rows (tiny  $m$  &  $n$ ) and a huge common dimension  $q$ .
- Futhark runs dense-dense at 13 out of 19 Tflops of the A100
- At sparsity factor  $64\times$  dense becomes better
- At sparsity factor  $128\times$  dense is still better than flattened
- Cublas + tensor cores will increase the threshold to (tens of) thousands, i.e., use sparse when one in thousands of elements is non zero