Regular Flattening

Troels Henriksen (athas@sigkill.dk)

DIKU
University of Copenhagen

Agenda

Representation and Fusion

Handling nested parallelism

Basic flattening rules

Incremental flattening

Multi-level parallelism

Final words as time permits

Representation and Fusion

Handling nested parallelism

Basic flattening rules

Incremental flattening

Multi-level parallelism

Final words as time permits

”Unzipped” SOACs

Representation

An array `[](t1, t2, t3...)` is represented in memory as `([]t1, []t2, []t3...)`, i.e. as *multiple arrays*, each containing only primitive values.

Instead of

```
let tmp = map (\(x,y) -> (x-1, y+1))
              (zip xs ys)
let (xs, ys) = unzip xs_ys’
```

we write

```
let (xs, ys) = map (\x y -> (x-1, y+1)) xs ys
```

- In the compiler, **All SOACs accept multiple inputs and produce unzipped results.**
- Arrays of tuples (or records, or sums) do not exist in the core language.
- **Isomorphic to source language**, but this form is simpler in a compiler.

Loop fusion

```
def increment [n][m] (as:  [n][m]i32) : [n]i32 =
  map (\r -> map (+2) r) a
def sum [n] (a:  [n]i32) : i32 =
  reduce (+) 0 a
def sumrows [n][m] (as: [n][m]i32) : [n]i32 =
  map sum as
```

Suppose we wish to first call `increment`, then `sumrows`:

`sumrows` (`increment` *a*)

Naively Run `increment`, then call `sumrows`.

Problem Manifests intermediate matrix in memory.

Solution *Loop fusion*, which combines loops to avoid intermediate results.

An example of a fusion rule

The expression

$$\mathbf{map}\ f\ (\mathbf{map}\ g\ a)$$

is *always* equivalent to

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

- This is an extremely powerful property that is only true in the absence of side effects.
- Fusion is *the* core optimisation that permits the efficient decomposition of a data parallel program.
- A full fusion engine has much more awkwark rules (mostly bookkeeping related to fusing only *some* of several inputs), but safety is guaranteed.

A fusion example

$$
\begin{aligned}
\mathtt{sumrows}\ (\mathtt{increment}\ a) &= && \text{(Initial expression)} \\
\mathbf{map}\ \mathtt{sum}\ (\mathtt{increment}\ a) &= && \text{(Inline }\mathtt{sumrows}\text{)} \\
\mathbf{map}\ \mathtt{sum}\ (\mathbf{map}\ (\lambda r \rightarrow \mathbf{map}\ (+2)\ r)\ a) &= && \text{(Inline }\mathtt{increment}\text{)} \\
\mathbf{map}\ (\mathtt{sum}\ \circ\ (\lambda r \rightarrow \mathbf{map}\ (+2)\ r)\ a) &= && \text{(Apply }\mathbf{map\text{-}map}\text{ fusion)} \\
\mathbf{map}\ (\lambda r \rightarrow \mathtt{sum}\ (\mathbf{map}\ (+2)\ r)\ a) &= && \text{(Apply composition)}
\end{aligned}
$$

- We have avoided the temporary matrix, but the composition of `sum` and the **`map`** also holds an opportunity for fusion – specifically, **`reduce`-`map`** fusion.
- Will not cover in detail, but a **`reduce`** can efficiently apply a function to each input element before engaging in the actual reduction operation.
- Important to remember: a **`map`** going into a **`reduce`** is an efficient pattern.

A shorthand notation for sequences

$$\overline{z}^{(n)} = z_0, \cdots, z_{(n-1)}$$

- The n may be omitted.
- A separator may be implied by context.

$$f\ \overline{v}^{(n)} \equiv f\ v_1\ \cdots\ v_n$$

or a tuple

$$(\overline{v}^{(n)}) \equiv (v_1, \ldots, v_n)$$

or a function type

$$\overline{\tau}^{(n)} \to \tau_{n+1} \equiv \tau_1 \to \cdots \to \tau_n \to \tau_{n+1}.$$

When not all terms under the bar are variant, subscript variant terms with i.

$$(\overline{[d]v_i}^{(n)}) = ([d]v_1, \ldots, [d]v_n)$$

and

$$(\overline{[d_i]v_i}^{(n)}) = ([d_1]v_1, \ldots, [d_n]v_n)$$

Fused constructs

Convenient shorthands

$$\begin{aligned}
\texttt{\textbf{redomap}}\ \odot\ f\ (\overline{d})\ \overline{xs} &\equiv \texttt{\textbf{reduce}}\ \odot\ (\overline{d})\ (\textbf{map}\ f\ \overline{xs}) \\
\texttt{\textbf{scanomap}}\ \odot\ f\ (\overline{d})\ \overline{xs} &\equiv \texttt{\textbf{scan}}\ \odot\ (\overline{d})\ (\textbf{map}\ f\ \overline{xs})
\end{aligned}$$

- Emphasises that **reduce**/**scan**-**map** compositions can be considered as a single construct.
- We will see several examples where this is useful.

Fused constructs

Convenient shorthands

$$\texttt{redomap} \odot f\ (\overline{d})\ \overline{xs} \equiv \texttt{reduce} \odot (\overline{d})\ (\textbf{map}\ f\ \overline{xs})$$

$$\texttt{scanomap} \odot f\ (\overline{d})\ \overline{xs} \equiv \texttt{scan} \odot (\overline{d})\ (\textbf{map}\ f\ \overline{xs})$$

- Emphasises that **reduce/scan-map** compositions can be considered as a single construct.
- We will see several examples where this is useful.

Note:

$$\texttt{reduce} \odot (\overline{d})\ \overline{xs} \equiv \texttt{reduce} \odot (\overline{d})\ (\textbf{map id}\ \overline{xs}) \equiv \texttt{redomap} \odot \textbf{id}\ (\overline{d})\ \overline{xs}$$

$$\texttt{scan} \odot (\overline{d})\ \overline{xs} \equiv \texttt{scan} \odot (\overline{d})\ (\textbf{map id}\ \overline{xs}) \equiv \texttt{scanomap} \odot \textbf{id}\ (\overline{d})\ \overline{xs}$$

Representation and Fusion

Handling nested parallelism

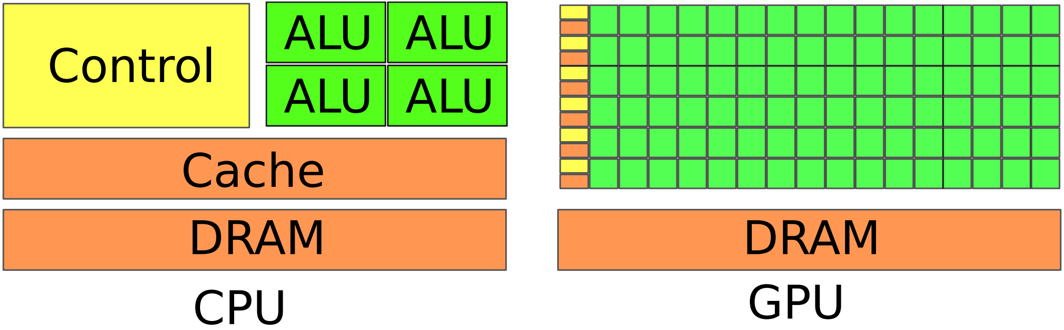
Basic flattening rules

Incremental flattening

Multi-level parallelism

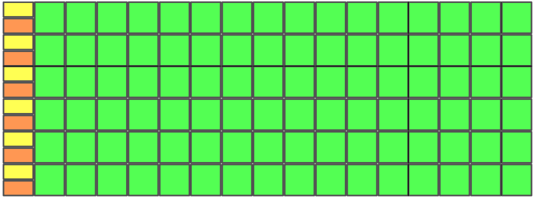
Final words as time permits

GPUs vs CPUs



- GPUs have *thousands* of simple cores and taking full advantage of their compute power requires *tens of thousands* of threads.
- GPU threads are very *restricted* in what they can do: no stack, no allocation, limited control flow, etc.
- Potential *very high performance* and *lower power usage* compared to CPUs, but programming them is *hard*.

The SIMT Programming Model



- GPUs are programmed using the SIMT model (*Single Instruction Multiple Thread*).
- Similar to SIMD (*Single Instruction Multiple Data*), but while SIMD has explicit vectors, we provide *sequential scalar per-thread* code in SIMT.

Each thread has its own registers, but they all execute the same instructions at the same time (i.e. they share their instruction pointer).

SIMT example

For example, to increment every element in an array a, we might use this code:

```
increment(a) {
  tid = get_thread_id();
  x = a[tid];
  a[tid] = x + 1;
}
```

- If a has n elements, we launch n threads, with `get_thread_id()` returning *i* for thread *i*.
- This is *data-parallel programming*: applying the same operation to different data.
- When we launch a GPU program (*kernel*), we say how many threads should be launched, *all running the same code*.

Branching

If all threads share an instruction pointer, what about branches?

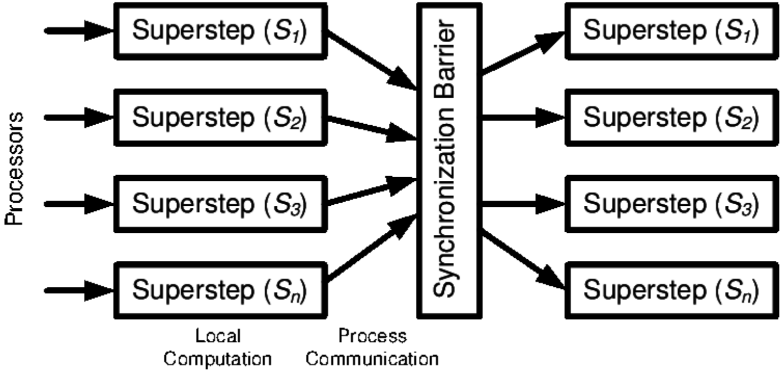
```
mapabs(a) {
  tid = get_thread_id();
  x = a[tid];
  if (x < 0) {
    a[tid] = -x;
  }
}
```

Masked Execution

Both branches are executed in all threads, but in those threads where the condition is false, a mask bit is set to treat the instructions inside the branch as no-ops.

Do GPUs exist in theory as well?

GPU programming is a close fit to the *bulk synchronous parallelism* model:



[1]

- Supersteps are *threads*, which cannot talk to each other.
- The synchronisation barriers are kernel launches.

[1] Illustration by Aftab A. Chandio.

A SOAC-kernel correspondence

The compiler *knows*[2] that certain nestings of **map**s correspond to GPU basic blocks.

- **map**s containing scalar code is a kernel with one thread per iteration of the **map**s.
- **map**s containing a single **`reduce`** is a *segmented reduction*.
- **map**s containing a single **`scan`** is a *segmented scan*.
- **map**s containing a single **`scatter`** is a *segmented scatter*.
- ...see the pattern?

Crucial: the **map**s must be *perfectly nested* around the operation.

Perfect nesting of an operation e_o

An expression e is a *perfect nesting* of e_o if e has form

$$\mathbf{map}\ (\lambda \overline{p} \rightarrow e_f)\ \overline{x}$$

where either $e_f = e_o$ or e_f is a perfect nesting of e_o.

[2] Because it was taught it by Cosmin in PMPH.

Example

```
map (\xs y ->
       map (\x ->
              x + y)
    xs)
  xss ys
```

- Suppose `xss` is of shape `[n][m]`.
- This could be compiled to a kernel with $n \times m$ threads, each doing a single $x + y$ operation.

Handling nested parallelism

Problem

Futhark permits *nested* parallelism, but GPUs need *flat* parallel *kernels*.

Handling nested parallelism

Problem

Futhark permits *nested* parallelism, but GPUs need *flat* parallel *kernels*.

Solution

Have the compiler rewrite program to perfectly nested **map**s containing sequential operations, or known parallel patterns such as segmented reduction.

Handling nested parallelism

Problem

Futhark permits *nested* parallelism, but GPUs need *flat* parallel *kernels*.

Solution

Have the compiler rewrite program to perfectly nested **map**s containing sequential operations, or known parallel patterns such as segmented reduction.

```
map (\xs -> let y = reduce (+) 0 xs
            in map (\x -> x + y) xs)
    xss
```

⇓

```
let ys = map (\xs -> reduce (+) 0 xs) xss
in map (\xs y -> map (\x -> x + y) xs) xss ys
```

Flattening via loop fission

The classic map fusion rule:

$$\text{map } f \circ \text{map } g \Rightarrow \text{map } (f \circ g)$$

[3] *Futhark: Purely Functional GPU-Programming with Nested Parallelism and In-Place Array Updates*, PLDI 2017

Flattening via loop fission

The classic map fusion rule:

$$\text{map } f \circ \text{map } g \Rightarrow \text{map } (f \circ g)$$

We can also apply it backwards to obtain *fission*:

$$\text{map } (f \circ g) \Rightarrow \text{map } f \circ \text{map } g$$

This, along with other fission rules (see paper[3]), are applied by the compiler to extract perfect map nests.

[3] *Futhark: Purely Functional GPU-Programming with Nested Parallelism and In-Place Array Updates*, PLDI 2017

Example: (a) Initial program, we inspect the map-nest

```
let (asss, bss) =
  map (\(ps: [m]i32) ->
        let ass = map (\(p: i32): [m]i32 ->
                        let cs = scan (+) 0 (iota p)
                        let r = reduce (+) 0 cs
                        in map (+r) ps) ps
        let bs = loop ws=ps for i < n do
                    map (\as w: i32 ->
                          let d = reduce (+) 0 as
                          let e = d + w
                          in 2 * e) ass ws
        in (ass, bs)) pss
```

We assume the type of `pss : [m][m]i32`.

(b) Distribution

```
let asss: [m][m][m]i32 =
  map (\(ps: [m]i32) ->
        let ass = map (\(p: i32): [m]i32 ->
                        let cs = scan (+) 0 (iota p)
                        let r = reduce (+) 0 cs
                        in map (+r) ps) ps
        in ass) pss
let bss: [m][m]i32 =
  map (\ps ass ->
        let bs = loop ws=ps for i < n do
                    map (\as w ->
                          let d = reduce (+) 0 as
                          let e = d + w
                          in 2 * e) ass ws
        in bs) pss asss
```

(c) Interchanging outermost map inwards

```
let asss: [m][m][m]i32 =
  map (\(ps: [m]i32) ->
        let ass = map (\(p: i32): [m]i32 ->
                        let cs = scan (+) 0 (iota p)
                        let r = reduce (+) 0 cs
                        in map (+r) ps) ps
        in ass) pss
let bss: [m][m]i32 =
  map (\ps ass ->
        let bs = loop ws=ps for i < n do
                    map (\as w ->
                          let d = reduce (+) 0 as
                          let e = d + w
                          in 2 * e) ass ws
        in bs) pss asss
```

(c) Interchanging outermost map inwards

```
let asss: [m][m][m]i32 =
  map (\(ps: [m]i32) ->
        let ass = map (\(p: i32): [m]i32 ->
                        let cs = scan (+) 0 (iota p)
                        let r = reduce (+) 0 cs
                        in map (+r) ps) ps
        in ass) pss
let bss: [m][m]i32 =
  loop wss=pss for i < n do
    map (\ass ws ->
          let ws' = map (\as w ->
                          let d = reduce (+) 0 as
                          let e = d + w
                          in 2 * e) ass ws
          in ws') asss wss
```

(d) Skipping scalar computation

```
let asss: [m][m][m]i32 =
  map (\(ps: [m]i32) ->
        let ass = map (\(p: i32): [m]i32 ->
                        let cs = scan (+) 0 (iota p)
                        let r = reduce (+) 0 cs
                        in map (+r) ps) ps
        in ass) pss
let bss: [m][m]i32 =
  loop wss=pss for i < n do
    map (\ass ws ->
          let ws' = map (\as w ->
                          let d = reduce (+) 0 as
                          let e = d + w
                          in 2 * e) ass ws
          in ws') asss wss
```

(d) Skipping scalar computation

```
let asss: [m][m][m]i32 =
  map (\(ps: [m]i32) ->
        let ass = map (\(p: i32): [m]i32 ->
                        let cs = scan (+) 0 (iota p)
                        let r = reduce (+) 0 cs
                        in map (+r) ps) ps
        in ass) pss
let bss: [m][m]i32 =
  loop wss=pss for i < n do
    map (\ass ws ->
          let ws' = map (\as w ->
                          let d = reduce (+) 0 as
                          let e = d + w
                          in 2 * e) ass ws
          in ws') asss wss
```

(e) Distributing reduction

```
let asss: [m][m][m]i32 =
  map (\(ps: [m]i32) ->
        let ass = map (\(p: i32): [m]i32 ->
                        let cs = scan (+) 0 (iota p)
                        let r = reduce (+) 0 cs
                        in map (+r) ps) ps
        in ass) pss
let bss: [m][m]i32 =
  loop wss=pss for i < n do
    map (\ass ws ->
          let ws' = map (\as w ->
                          let d = reduce (+) 0 as
                          let e = d + w
                          in 2 * e) ass ws
          in ws') asss wss
```

(e) Distributing reduction

```
let asss: [m][m][m]i32 =
  map (\(ps: [m]i32) ->
        let ass = map (\(p: i32): [m]i32 ->
                        let cs = scan (+) 0 (iota p)
                        let r = reduce (+) 0 cs
                        in map (+r) ps) ps
        in ass) pss
let bss: [m][m]i32 =
  loop wss=pss for i < n do
    let dss: [m][m]i32 =
      map (\ass ->
            map (\as ->
                  reduce (+) 0 as) ass)
          asss
    in map (\ws ds ->
             let ws’ =
                map (\w d -> let e  = d + w
                            in 2 * e) ws ds
             in ws’) asss dss
```

(f) Distributing inner map

```
let asss =
  map (\(ps: [m]i32) ->
        let ass = map (\(p: i32): [m]i32 ->
                        let cs = scan (+) 0 (iota p)
                        let r = reduce (+) 0 cs
                        in map (+r) ps) ps
        in ass) pss
let bss: [m][m]i32 = ...
```

(f) Distributing inner map

```
let rss: [m][m]i32 =
  map (\(ps: [m]i32) ->
        let rs = map (\(p: i32): i32 ->
                       let cs = scan (+) 0 (iota p)
                       let r = reduce (+) 0 cs
                       in r) ps
        in rs) pss
let asss: [m][m][m]i32 =
  map (\(ps: [m]i32) (rs: [m]i32) ->
        map (\(r: i32): [m]i32 ->
              map (+r) ps) rs
      ) pss rss
let bss: [m][m]i32 = ...
```

(g) Cannot distribute as it would create irregular array

```
let rss: [m][m]i32 =
  map (\(ps: [m]i32) ->
        let rs = map (\(p: i32): i32 ->
                        let cs = scan (+) 0 (iota p)
                        let r = reduce (+) 0 cs
                        in r) ps
        in rs) pss
let asss: [m][m][m]i32 = ...
let bss: [m][m]i32 = ...
```

Array `cs` has type `[p]i32`, and p is variant to the innermost map nest.

(h) These statements are sequentialised

```
let rss: [m][m]i32 =
  map (\(ps: [m]i32) ->
        let rs = map (\(p: i32): i32 ->
                       let cs = scan (+) 0 (iota p)
                       let r = reduce (+) 0 cs
                       in r) ps
        in rs) pss
let asss: [m][m][m]i32 = ...
let bss: [m][m]i32 = ...
```

Array `cs` has type `[p]i32`, and p is variant to the innermost map nest.

Result

```
let rss: [m][m]i32 = map (\ps -> map (...) ps) pss
let asss: [m][m][m]i32 =
  map (\ps rs -> map (\r -> map (...) ps) rs) pss rss
let bss: [m][m]i32 =
  loop wss=pss for i < n do
    let dss: [m][m]i32 = map (\ass -> map (reduce ...) ass)
                             asss
    in map (\ws ds -> map (...) ws ds ) asss dss
```

- From a single kernel with parallelism *m* to four kernels of parallelism m^2, m^3, m^3, and m^2.
- The last two kernels are executed *n* times each.

Representation and Fusion

Handling nested parallelism

Basic flattening rules

Incremental flattening

Multi-level parallelism

Final words as time permits

Notation for flat parallelism

Instead of

```
map (\ps rs ->
  map (\r ->
    map (\p -> e)
      ps)
    rs)
  pss rss
```

we write

$$\begin{array}{l}\textbf{segmap}\ (\langle \texttt{ps}, \texttt{rs} \in \texttt{pss}, \texttt{rss}\rangle,\ \langle \texttt{r} \in \texttt{rs}\rangle,\ \langle \texttt{p} \in \texttt{ps}\rangle) \\ \quad e\end{array}$$

Segmented flat parallel constructs

$$\Sigma = \Sigma', \langle \overline{x} \in \overline{y} \rangle$$

$$\texttt{\textbf{segmap}}\ \Sigma\ e \equiv \quad \begin{array}{l} \texttt{\textbf{map}}\ (\lambda \overline{x_p} \rightarrow \\ \quad \texttt{\textbf{map}}\ (\lambda \overline{x_{p-1}} \rightarrow \ldots \\ \quad\quad \texttt{\textbf{map}}\ (\lambda \overline{x_1} \rightarrow e)\ \overline{y_1}) \\ \quad\quad \overline{y_{p-1}}) \\ \quad \overline{y_p} \end{array}$$

- Conceptually a perfect nest of **map**s with some operation inside.
- *These* are what trigger GPU code generation.
- Any SOACs left in *e* will be executed sequentially.

Similarly for reductions and scans

$$
\begin{aligned}
\texttt{\textbf{segred}}\ \Sigma\ \odot\ \overline{d}\ e \equiv\ & \texttt{\textbf{map}}\ (\lambda\overline{x_p} \rightarrow \\
& \quad \texttt{\textbf{map}}\ (\lambda\overline{x_{p-1}} \rightarrow \ldots \\
& \qquad \texttt{\textbf{redomap}}\ \odot\ (\lambda\overline{x_1} \rightarrow e)\ (\overline{d})\ \ \overline{y_1}) \\
& \qquad \overline{y_{p-1}}) \\
& \quad \overline{y_p}
\end{aligned}
$$

$$
\begin{aligned}
\texttt{\textbf{segscan}}\ \Sigma\ \odot\ \overline{d}\ e \equiv\ & \texttt{\textbf{map}}\ (\lambda\overline{x_p} \rightarrow \\
& \quad \texttt{\textbf{map}}\ (\lambda\overline{x_{p-1}} \rightarrow \ldots \\
& \qquad \texttt{\textbf{scanomap}}\ \odot\ (\lambda\overline{x_1} \rightarrow e)\ \overline{d}\ \overline{y_1}) \\
& \qquad \overline{y_{p-1}}) \\
& \quad \overline{y_p}
\end{aligned}
$$

Let us look at how one can rewrite SOAC nests to these segmented operations.

Functional flattening rules

A flattening rewrite

$$\mathcal{G}(\Sigma, e) \Rightarrow e'$$

says that flattening an operation e nested inside a map nest Σ produces the expression e'.

- *Rewrite rules* define which rewrites are valid.
- Not an algorithm: multiple rewrites are often possible.
- Form in is a bit different and have more details; I have tried to keep it more high level.

Basic rules

$$\mathcal{G}(\bullet, e) \Rightarrow e \tag{G0}$$

$$\mathcal{G}(\Sigma, e) \Rightarrow \texttt{\textbf{segmap}}\ \Sigma\ e \tag{G1}$$

$$\mathcal{G}(\Sigma, \textbf{map}\ (\lambda \overline{x} \rightarrow e)\ \overline{xs}) \Rightarrow \mathcal{G}(\Sigma\ \langle \overline{x} \rangle \in \overline{xs}, e) \tag{G2}$$

$$\mathcal{G}(\Sigma, \texttt{\textbf{redomap}}\ \odot\ (\lambda \overline{x} \rightarrow e)\ (\overline{d})\ \overline{xs}) \Rightarrow \texttt{\textbf{segred}}\ (\Sigma\ \overline{x} \in \overline{xs})\ \odot\ \overline{d}\ e \tag{Gr}$$

G0: at top level, choose not to flatten.
G1: inside a nest, emit the corresponding **segmap**.
G2: descend into a **map**.
Gr: for a nested **redomap**, emit a **segred**.

Distribution rule

$$\mathcal{G}(\Sigma, \mathtt{\textbf{let}}\ \overline{a_0} = e_1\ \mathtt{\textbf{in}}\ e_2) \Rightarrow \mathtt{\textbf{let}}\ \overline{a_p} = \mathcal{G}(\Sigma, e_1)\ \mathtt{\textbf{in}}\ \mathcal{G}(\Sigma', e_2) \tag{G6}$$

where

$$\begin{array}{rcl} \Sigma & = & \langle \overline{x_p} \in \overline{y_p} \rangle, \ldots, \langle \overline{x_1} \in \overline{y_1} \rangle \\ \Sigma' & = & \langle \overline{x_p}\ \overline{a_{p-1}} \in \overline{y_p}\ \overline{a_p} \rangle, \ldots, \langle \overline{x_1}\ \overline{a_0} \in \overline{y_1}\ \overline{a_1} \rangle \end{array}$$

and $\overline{a_p}, \ldots, \overline{a_1}$ are fresh names.

Note: only applicable when each array in $\overline{a_0}$ is invariant to Σ.

Example

We are flattening

$$
\begin{aligned}
e = \mathbf{map}\ (\lambda \mathtt{xs} \rightarrow\ & \mathbf{let}\ \mathtt{y} = \mathbf{redomap}\ (+)\ (\lambda \mathtt{x} \rightarrow \mathtt{x})\ 0\ xs \\
& \mathbf{in}\ \mathbf{map}\ (\lambda \mathtt{x} \rightarrow \mathtt{x} + \mathtt{y})\ \mathtt{xs}) \\
\mathtt{xss}\quad &
\end{aligned}
$$

Example

We are flattening

$$
\begin{aligned}
e = \textbf{map}\ (\lambda \mathtt{xs} \rightarrow\ & \texttt{\textbf{let}}\ \mathtt{y} = \texttt{\textbf{redomap}}\ (+)\ (\lambda \mathtt{x} \rightarrow \mathtt{x})\ 0\ \mathit{xs} \\
& \texttt{\textbf{in}}\ \textbf{map}\ (\lambda \mathtt{x} \rightarrow \mathtt{x} + \mathtt{y})\ \mathtt{xs}) \\
\mathtt{xss} &
\end{aligned}
$$

By applying rule G1 we get

$$
\begin{aligned}
\mathcal{G}(\bullet, e) = \mathcal{G}(\langle \mathtt{xs} \in \mathtt{xss} \rangle,\ & \texttt{\textbf{let}}\ \mathtt{y} = \texttt{\textbf{redomap}}\ (+)\ (\lambda \mathtt{x} \rightarrow \mathtt{x})\ 0\ \mathit{xs}\ \ ) \\
& \texttt{\textbf{in}}\ \textbf{map}\ (\lambda \mathtt{x} \rightarrow \mathtt{x} + \mathtt{y})\ \mathtt{xs})
\end{aligned}
$$

$$\begin{aligned}\mathcal{G}(\langle \mathsf{xs} \in \mathsf{xss}\rangle,\ & \mathbf{let}\ \mathsf{y} = \mathbf{redomap}\ (+)\ (\lambda \mathsf{x} \rightarrow \mathsf{x})\ 0\ \mathit{xs}\ \ ) \\ & \mathbf{in}\ \mathbf{map}\ (\lambda \mathsf{x} \rightarrow \mathsf{x} + \mathsf{y})\ \mathsf{xs})\end{aligned}$$

Apply distribution rule by:

$$\begin{aligned}
\overline{a_0} &= \mathsf{y} \\
e_1 &= \mathbf{redomap}\ (+)\ (\lambda \mathsf{x} \rightarrow \mathsf{x})\ 0\ \mathsf{xs} \\
e_2 &= \mathbf{map}\ (\lambda \mathsf{x} \rightarrow \mathsf{x} + \mathsf{y})\ \mathsf{xs} \\
\Sigma &= \langle \mathsf{xs} \in \mathsf{xss}\rangle \\
\Sigma' &= \langle \mathsf{xs}, \mathsf{y} \in \mathsf{xss}, \mathsf{ys}\rangle \\
\mathcal{G}(\Sigma, e) &= \mathbf{let}\ \mathsf{ys} = \mathcal{G}(\Sigma, e_1)\ \mathbf{in}\ \mathcal{G}(\Sigma', e_2) \\
&= \mathbf{let}\ \mathsf{ys} = \mathbf{segred}\ (\langle \mathsf{xs} \in \mathsf{xss}\rangle, \langle \mathsf{x} \in \mathsf{xs}\rangle)\ (+)\ 0\ \mathsf{x} \\
&\ \mathbf{in}\ \mathbf{segmap}\ \langle \mathsf{xs}, \mathsf{y} \in \mathsf{xss}, \mathsf{ys}\rangle\ (\mathsf{x} + \mathsf{y})
\end{aligned}$$

Reminder: distribution rule

$$\mathcal{G}(\Sigma, \mathbf{let}\ \overline{a_0} = e_1\ \mathbf{in}\ e_2) \Rightarrow \mathbf{let}\ \overline{a_p} = \mathcal{G}(\Sigma, e_1)\ \mathbf{in}\ \mathcal{G}(\Sigma', e_2) \qquad \text{(G6)}$$

Flattening transposition

$\texttt{\textbf{rearrange}}\ (d_1, \cdots, d_n)\ x$ is a generalization of $\texttt{\textbf{transpose}}$ in that it rearranges the dimensions of d-dimensional array based on a permutation defined by the integer sequence $d_1, \cdots, d_n$. E.g:

$$\texttt{\textbf{transpose}} \equiv \texttt{\textbf{rearrange}}\ (1, 0)$$

Flattening transposition

$\mathbf{\texttt{rearrange}}\,(d_1, \cdots, d_n)\ x$ is a generalization of $\mathbf{\texttt{transpose}}$ in that it rearranges the dimensions of d-dimensional array based on a permutation defined by the integer sequence $d_1, \cdots, d_n$. E.g:

$$\mathbf{\texttt{transpose}} \equiv \mathbf{\texttt{rearrange}}\,(1, 0)$$

Flattening rule

$$\mathcal{G}(\Sigma, \langle x \in y \rangle, \mathbf{\texttt{rearrange}}\,(k_1, \ldots, k_n)\ x) \Rightarrow \mathcal{G}(\Sigma, \mathbf{\texttt{rearrange}}\,(0, 1 + k_1, \ldots, 1 + k_n)\ y)$$

Eventually reaches base case where $\Sigma = \bullet$.

map-loop interchange

$$\begin{aligned}&\mathcal{G}(\Sigma\ \langle \overline{x}\,\overline{y} \in \overline{xs}\,\overline{ys}\rangle, \mathbf{\texttt{loop}}\ \overline{z'}\ \overline{y'}\ \mathbf{\texttt{=}}\ \overline{z}\ \overline{y}\ \mathbf{\texttt{for}}\ i < n\ \mathbf{do}\ f\ i\ \overline{q}\ \overline{x}\ \overline{y}\ \overline{y'}\ \overline{z'}) \Rightarrow\\ &\quad\mathcal{G}(\Sigma, \mathbf{\texttt{loop}}\ \overline{zs'}\ \overline{ys'}\ =\ \overline{z^r}\ \overline{ys}\ \mathbf{\texttt{for}}\ i\ <\ n\ \mathbf{do}\ \mathbf{map}\ (f\ i\ \overline{q})\ \overline{xs}\ \overline{ys}\ \overline{ys'}\ \overline{zs'})\end{aligned}$$

where

$$\begin{aligned} m &= \text{outer size of each of } \overline{xs} \text{ and } \overline{ys} \\ \overline{z^r} &= \overline{\mathbf{\texttt{replicate}}\ m\ z_i} \\ \{n, \overline{q}, \overline{z}\} \cap \{\overline{x}, \overline{y}\} &= \emptyset \end{aligned}$$

and $\overline{zs'}$ and $\overline{ys'}$ are fresh names.

Informal example of interchange

```
map (\xs -> loop (xs', j) = (xs, 0) for i < n do
              (map (+j) xs', j + i))
    xss
```

Becomes after interchange

```
loop (xss', js) = (xss, replicate m 0) for i < n do
  map (\xs' j -> (map (+j) xs', j + i))
      xss' js
```

Validity of interchange

The simple intuition is that

```
map (\x -> loop x’ = x for i < n do f x’) xs
```

is equivalent to

```
loop xs’ = xs for i < n do map f xs’
```

because they both produce

$$[\mathtt{f}^n\,\mathtt{xs[0]}, \ldots, \mathtt{f}^n\,\mathtt{xs[m-1]}]$$

Representation and Fusion

Handling nested parallelism

Basic flattening rules

Incremental flattening

Multi-level parallelism

Final words as time permits

Consider Matrix Multiplication

```
for i < n:
  for j < m:
    acc = 0
    for l < p:
      acc += xss[i,l] * yss[l,j]
    res[i,j] = acc
```

Turning it Functional

```
map (\xs ->
     map (\ys ->
          let zs = map (*) xs ys
          in reduce (+) 0 zs)
         (transpose yss))
    xss
```

Using `redomap` notation

```
map (\xs ->
     map (\ys ->
          redomap (+) (*) 0 xs ys)
         (transpose yss))
    xss
```

$$\texttt{\textbf{redomap}}\ \odot\ f\ 0_{\odot}\ x \ \equiv\ \texttt{\textbf{reduce}}\ \odot\ 0_{\odot}\ (\textbf{map}\ f\ x)$$

Emphasises that a **map-reduce** composition can be turned into a fused tight sequential loop, or into a parallel reduction.

So how should we parallelise this on GPU?

So how should we parallelise this on GPU?

Full flattening

```
map (\xs ->
  map (\ys ->
    redomap (+) (*) 0 xs ys)
    (transpose yss))
  xss
```

- **All parallelism exploited**
- Some communication overhead
- *Best if outer* **map***s don't saturate GPU*

So how should we parallelise this on GPU?

Full flattening

```
map (\xs ->
  map (\ys ->
    redomap (+) (*) 0 xs ys)
    (transpose yss))
  xss
```

- **All parallelism exploited**
- Some communication overhead
- *Best if outer* **maps** *don't saturate GPU*

Moderate flattening

```
map (\xs ->
  map (\ys ->
    redomap (+) (*) 0 xs ys)
    (transpose yss))
  xss
```

- **Only outer parallelism**
- The **redomap** can be block tiled
- *Best if outer* **maps** *saturate GPU*

- There is no *one size fits all.*
- Both situations may be encountered at program runtime.

The essence of *incremental flattening*

From a single source program, for each parallel construct generate multiple *semantically equivalent* parallelisations, and generate a *single program* that at runtime picks the *least parallel* that still saturates the hardware.

- Implemented in the Futhark compiler.
- ...but technique is applicable to any (regular) nested parallelism expressed with the common functional array combinators (`map`, `reduce`, `scan`, etc).

Simple Incremental Flattening

At every level of `map`-nesting we have two options:

1. Continue flattening inside the `map`, exploiting the parallelism there.
2. Sequentialise the `map` body; exploiting only the parallelism on top.

- **Full flattening** in the Blelloch style will do the former, maximising utilised parallelism.
- **Incremental flattening** generates *both* versions and uses a predicate to pick at runtime.

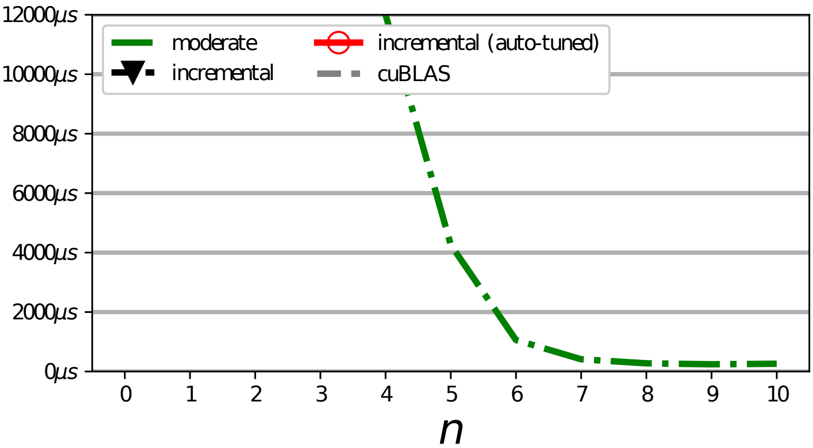
Multi-versioned matrix multiplication

```
xss :  [n][p]i32
yss :  [p][m]i32.
```

```
if n * m > $t_0$ then
  map (\xs ->
         map (\ys ->
                redomap (+) (*) 0 xs ys)
             (transpose yss))
      xss
else
  map (\xs ->
         map (\ys ->
                redomap (+) (*) 0 xs ys)
             (transpose yss))
      xss
```

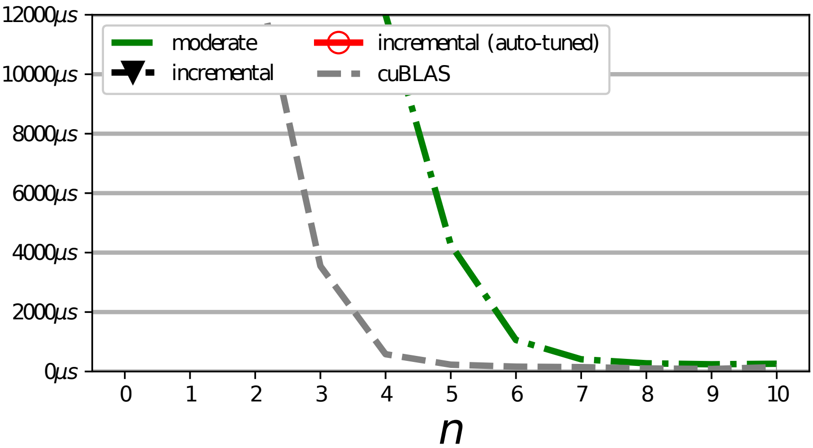
The t_0 *threshold parameter* is used to select between the two versions—and should be auto-tuned on the concrete hardware.

Matrix multiplication on NVIDIA K40



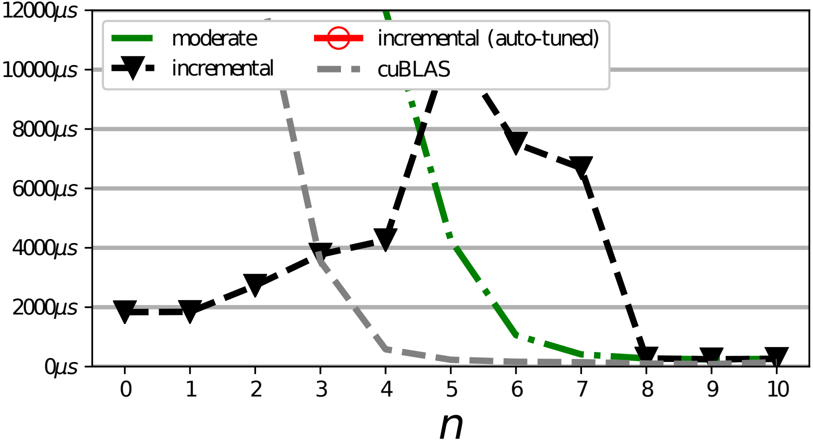
Multiplying matrices of size $2^n \times 2^m$ and $2^m \times 2^n$, where $m = 25 - 2n$, meaning that work is constant as we vary n.

Matrix multiplication on NVIDIA K40



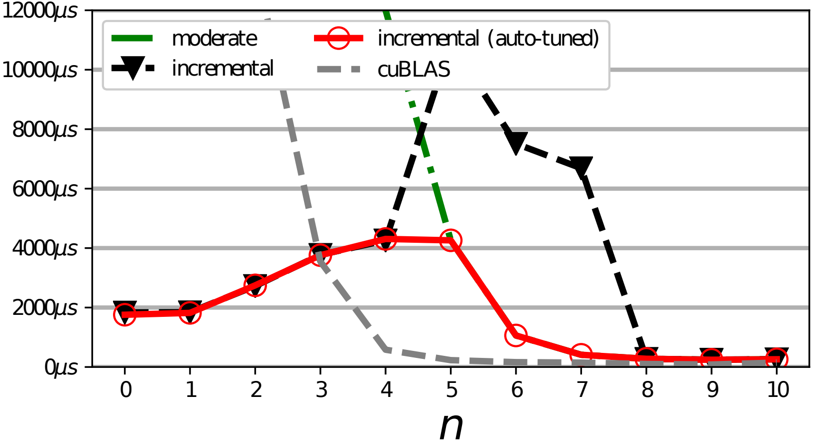
Multiplying matrices of size $2^n \times 2^m$ and $2^m \times 2^n$, where $m = 25 - 2n$, meaning that work is constant as we vary n.

Matrix multiplication on NVIDIA K40



Multiplying matrices of size $2^n \times 2^m$ and $2^m \times 2^n$, where $m = 25 - 2n$, meaning that work is constant as we vary n.

Matrix multiplication on NVIDIA K40



Multiplying matrices of size $2^n \times 2^m$ and $2^m \times 2^n$, where $m = 25 - 2n$, meaning that work is constant as we vary n.

Incremental flattening rule

$$\mathcal{G}(\Sigma, \textbf{map}\ (\lambda \overline{x} \rightarrow e)\ \overline{\mathsf{xs}}) \Rightarrow \begin{array}{l} \texttt{if}\ \mathrm{Parallelism}(\Sigma') \geq t_{\mathrm{top}} \\ \texttt{then segmap}\ \Sigma'\ e \\ \texttt{else}\ \mathcal{G}(\Sigma', e) \end{array}$$

where $\Sigma' = \Sigma, \langle \overline{x} \in \overline{\mathsf{xs}} \rangle$.

Example for

```
map (\xs -> redomap (+) (\x -> x) 0 xs) xss
```

Incremental flattening rule

$$\mathcal{G}(\Sigma, \textbf{map}\ (\lambda\overline{x} \rightarrow e)\ \overline{\texttt{xs}}) \Rightarrow \begin{array}{l} \textbf{if}\ \text{Parallelism}(\Sigma') \geq t_{\text{top}} \\ \textbf{then segmap}\ \Sigma'\ e \\ \textbf{else}\ \mathcal{G}(\Sigma', e) \end{array}$$

where $\Sigma' = \Sigma, \langle \overline{x} \in \overline{\texttt{xs}} \rangle$.

Example for

```
map (\xs -> redomap (+) (\x -> x) 0 xs) xss
```

$$\Sigma = \bullet \quad \Sigma' = \langle \texttt{xs} \in \texttt{xss} \rangle$$

$$\Sigma' \vdash e \Rightarrow \textbf{segred}\ (\langle \texttt{xs} \in \texttt{xss} \rangle, \langle \texttt{x} \in \texttt{xs} \rangle)\ \texttt{(+)}\ 0\ \texttt{x}$$

$$\Sigma \vdash \ldots \Rightarrow \quad \begin{array}{l} \textbf{if}\ \text{length}(\texttt{xss}) \geq t_{\text{top}} \\ \textbf{then segmap}\ \langle \texttt{xs} \in \texttt{xss} \rangle\ (\texttt{redomap (+) (}\lambda\texttt{x} \rightarrow \texttt{x) 0 xs)} \\ \textbf{else segred}\ (\langle \texttt{xs} \in \texttt{xss} \rangle, \langle \texttt{x} \in \texttt{xs} \rangle)\ \texttt{(+)}\ 0\ \texttt{x} \end{array}$$

Autotuning

- An incrementally flattened program may have dozens of threshold parameters, t_i, used to select versions at runtime.
- As we have seen, the default value (2^{16}) is often not optimal.

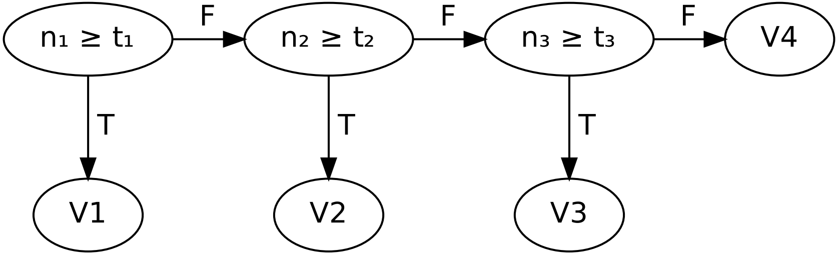
A *configuration* P maps each t_i to an integer $P(t_i)$.

The search problem

Find the P that minimises the cost function $F(P)$, where the the cost function runs the program on a set of user-provided representative datasets and sums the observed runtimes.

- Other cost functions are also possible, e.g. average runtime over datasets.
- **Note:** recompilation is not necessary.

Briefly on the search procedure[4]



- Suppose we are given training data sets $D_j, j < k$, each of which provide a value $v_{i,j}$ for each threshold parameter n_i.
- Starting from the deepest comparison (t_3), for each D_j find an (x_j, y_j) that minimises runtime, take the intersection of the intervals, and use that to determine threshold value.
- Tuning time is linear in the number of comparisons.

[4] https://futhark-lang.org/publications/tfp21.pdf

Using incremental flattening

Compile with a GPU backend (opencl or cuda):

```
$ futhark opencl matmul.fut
```

To autotune:

```
$ futhark autotune -v --backend=opencl matmul.fut
```

Produces matmul.fut.tuning, which is automatically picked up by futhark bench (use --no-tuning to stop this).

Use futhark dev -s --extract-kernels -e matmul.fut to see IR.

Representation and Fusion

Handling nested parallelism

Basic flattening rules

Incremental flattening

Multi-level parallelism

Final words as time permits

Confession

I lied when I claimed that GPU threads were completely isolated.

Confession

I lied when I claimed that GPU threads were completely isolated.

- Most hardware has useful (fixed) levels of parallelism.
- An ideal flattening algorithm maps levels of application parallelism (any number) to hardware parallelism (fixed number) in a way that exploits locality well.

Example of deep nesting: a system consists of multiple *datacenters*, that each contain multiple *computers*, that each contain multiple *GPUs*, that each contain multiple *SMs* (next slide), that each run some number of threads.

Confession

I lied when I claimed that GPU threads were completely isolated.

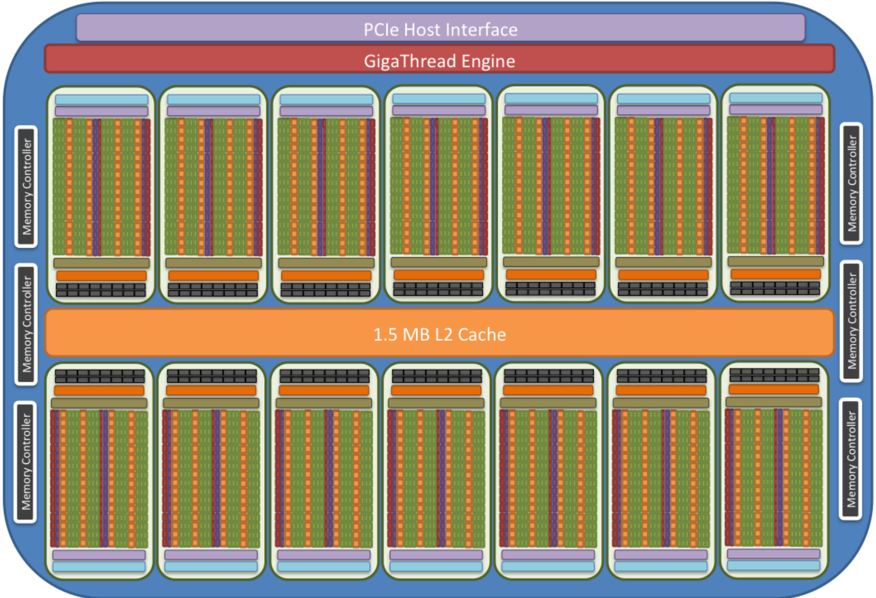
- Most hardware has useful (fixed) levels of parallelism.
- An ideal flattening algorithm maps levels of application parallelism (any number) to hardware parallelism (fixed number) in a way that exploits locality well.

Example of deep nesting: a system consists of multiple *datacenters*, that each contain multiple *computers*, that each contain multiple *GPUs*, that each contain multiple *SMs* (next slide), that each run some number of threads.

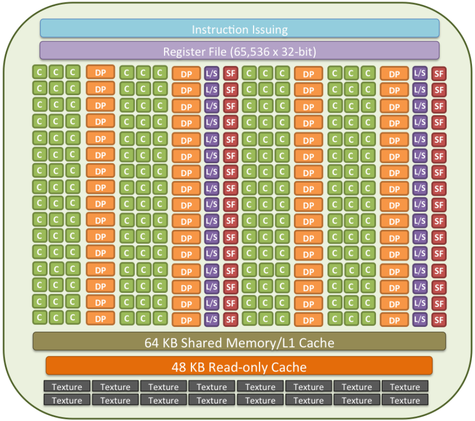
General principle

"Tasks" at the same hardware level cannot communicate, but can "launch" tasks at a lower level.

K20 GPU layout



Streaming Multiprocessor (SM) layout



Instruction Issuing

Register File (65,536 x 32-bit)

64 KB Shared Memory/L1 Cache

48 KB Read-only Cache

C single precision/integer CUDA core

DP double precision FP unit

L/S memory load/store unit

SF special function unit

Level-aware segmented operations

$$l \in \{\mathrm{thread}, \mathrm{group}\}$$

- *Group* is the same as a CUDA *thread block*
- Each segmented operation then tagged with the level at which its *body* executes.

$$\mathbf{segmap}_l\ \Sigma\ e$$
$$\mathbf{segscan}_l\ \Sigma\ \odot\ \overline{d}\ e$$
$$\mathbf{segred}_l\ \Sigma\ \odot\ \overline{d}\ e$$

Restrictions

Both thread and group can occur at top level, but a group construct can contain only thread constructs, and thread cannot contain any segmented constructs.

Examples

Each thread transposes part of an array

$$\mathbf{segmap}_{\text{thread}}\ \langle \mathtt{x} \in \mathtt{xs} \rangle\ (\mathtt{transpose\ x})$$

Each workgroup transposes part of an array

$$\mathbf{segmap}_{\text{group}}\ \langle \mathtt{x} \in \mathtt{xs} \rangle\ (\mathtt{transpose\ x})$$

These are both equivalent to `map transpose xs`.

Each workgroup sums the row of an array

$$\mathbf{segmap}_{\text{group}}\ \langle \mathtt{xs} \in \mathtt{xss} \rangle\ (\mathbf{segred}_{\text{thread}}\ \langle \mathtt{x} \in \mathtt{xs} \rangle\ (+)\ 0\ \mathtt{x})$$

Equivalent to `map (reduce (+) 0) xss`.

Tags carry no semantic meaning; used solely for code generation.

Example: LocVolCalib

The following is the essential core of the LocVolCalib benchmark from the FinPar suite.

```
map (\xss ->
      map (\xs ->
            let bs = scan ⊕ $d_{\oplus}$ xs
            let cs = scan ⊗ $d_{\otimes}$ bs
            in  scan ⊙ $d_{\odot}$ cs)
          xss)
    xsss
```

How can we map the application parallelism to hardware parallelism?

Option I: sequentialise the inner scans

$$\begin{array}{l}
\textbf{segmap}_{\text{thread}}\ (\langle \mathtt{xss} \in \mathtt{xssss}\rangle, \langle \mathtt{xs} \in \mathtt{xss}\rangle)\\
\quad \textbf{let}\ \mathtt{bs} = \textbf{scan}\ \oplus\ d_{\oplus}\ \mathtt{xs}\\
\quad \textbf{let}\ \mathtt{cs} = \textbf{scan}\ \otimes\ d_{\otimes}\ \mathtt{bs}\\
\quad \textbf{in}\ \textbf{scan}\ \odot\ d_{\odot}\ cs
\end{array}$$

scan is relatively expensive in parallel, so this is a good option if the outer dimensions provide enough parallelism.

Option II: flatten and parallelise inner scans

Flattening uses *loop distribution* (or *fission*) to create **map** nests:

```
map (\xss ->
      map (\xs ->
            let bs = scan $\oplus$ $d_{\oplus}$ xs
            let cs = scan $\otimes$ $d_{\otimes}$ bs
            in  scan $\odot$ $d_{\odot}$ cs)
          xss)
    xsss
```

Option II: flatten and parallelise inner scans

```
let bsss =
  $\mathbf{segscan}_{\text{thread}}\ (\langle \mathtt{xss} \in \mathtt{xssss} \rangle, \langle \mathtt{xs} \in \mathtt{xss} \rangle, , \langle \mathtt{x} \in \mathtt{xs} \rangle)\ \oplus\ d_{\oplus}\ \mathtt{x}$
let csss =
  $\mathbf{segscan}_{\text{thread}}\ (\langle \mathtt{bss} \in \mathtt{bsss} \rangle, \langle \mathtt{bs} \in \mathtt{bss} \rangle, , \langle \mathtt{b} \in \mathtt{bs} \rangle)\ \oplus\ d_{\oplus}\ \mathtt{b}$
in
  $\mathbf{segscan}_{\text{thread}}\ (\langle \mathtt{css} \in \mathtt{csss} \rangle, \langle \mathtt{cs} \in \mathtt{css} \rangle, , \langle \mathtt{c} \in \mathtt{cs} \rangle)\ \oplus\ d_{\oplus}\ \mathtt{c}$
```

This is what full flattening will do.

Option III: Mapping innermost parallelism to the workgroup level

```
map (\xss ->
      map (\xs ->
            let bs = scan $\oplus$ $d_{\oplus}$ xs
            let cs = scan $\otimes$ $d_{\otimes}$ bs
            in   scan $\odot$ $d_{\odot}$ cs)
          xss)
    xsss
```

Option III: Mapping innermost parallelism to the workgroup level

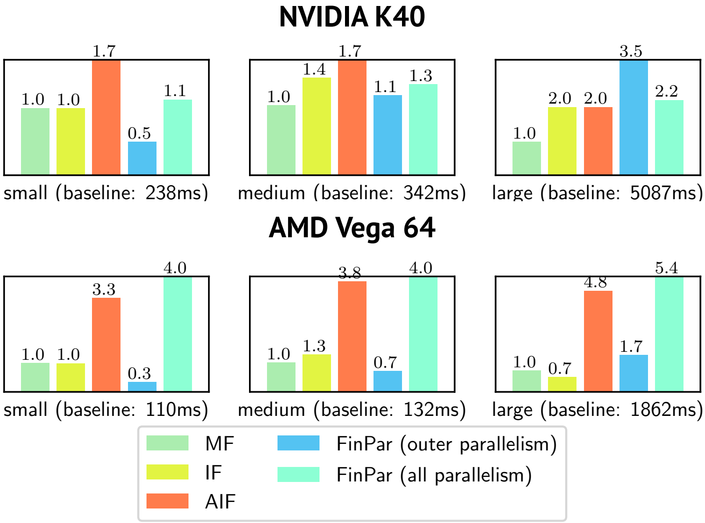
$$\begin{array}{l}
\mathbf{segmap}_{\text{group}}\ (\langle \mathtt{xss} \in \mathtt{xssss} \rangle, \langle \mathtt{xs} \in \mathtt{xss} \rangle) \\
\quad \mathbf{let}\ \mathtt{bs} = \mathbf{segscan}_{\text{thread}}\ \langle \mathtt{x} \in \mathtt{xs} \rangle\ \oplus\ d_{\oplus}\ \mathtt{x} \\
\quad \mathbf{let}\ \mathtt{cs} = \mathbf{segscan}_{\text{thread}}\ \langle \mathtt{b} \in \mathtt{bs} \rangle\ \otimes\ d_{\otimes}\ \mathtt{b} \\
\quad \mathbf{in}\ \ \mathbf{segscan}_{\text{thread}}\ \langle \mathtt{c} \in \mathtt{cs} \rangle\ \otimes\ d_{\otimes}\ \mathtt{c}
\end{array}$$

- Iterations of outer **segmap**s assigned to GPU workgroups[5].
- Each $\mathbf{segscan}_{\text{thread}}$ is executed collaboratively by a workgroup and in local memory[6].
- Only works if the innermost parallelism fits in a workgroup.

[5] *Thread block* in CUDA

[6] *Shared memory* in CUDA

LocVolCalib speedup (higher is better)



Sequential scans (MF) is the baseline.

Level-aware incremental flattening

$\boxed{\mathcal{G}_l(\Sigma, e) \Rightarrow e'}$ In a map-nest context Σ, the source expression e can be translated at machine level l into the target expression e'.

$$
\begin{array}{rl}
\mathcal{G}_{l+1}(\Sigma, \mathbf{map}\ (\lambda \overline{x} \rightarrow e)\ \overline{\mathtt{xs}}) \Rightarrow & \mathbf{if}\ \mathrm{Par}(\Sigma') \geq t_{\mathrm{top}} \\
& \mathbf{then}\ \mathbf{segmap}_{l+1}\ \Sigma'\ e \\
& \mathbf{else}\ \mathbf{if}\ \mathrm{Par}(e_{\mathrm{middle}}) \geq t_{\mathrm{intra}} \\
& \mathbf{then}\ \mathbf{segmap}_{l+1}\ \Sigma'\ \mathcal{G}_l(\bullet, e) \\
& \mathbf{else}\ \mathcal{G}_{l+1}(\Sigma', e)
\end{array}
$$

where $\Sigma' = \Sigma, \langle \overline{x} \in \overline{\mathtt{xs}} \rangle$ and $t_{\mathrm{top}}, t_{\mathrm{intra}}$ fresh.
In the Futhark compiler, only two levels are handled (thread, group), but we believe the idea generalises well.

Representation and Fusion

Handling nested parallelism

Basic flattening rules

Incremental flattening

Multi-level parallelism

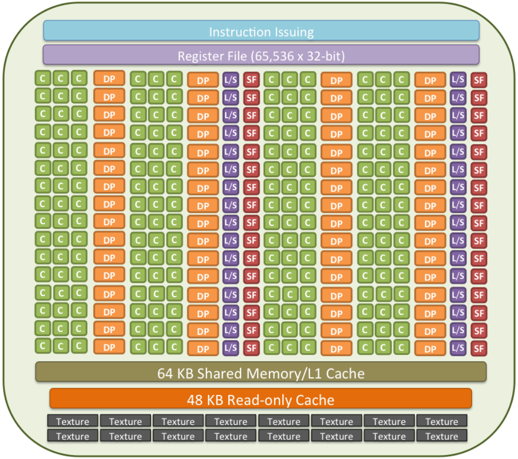
Final words as time permits

Block tiling

Level-aware constructs can also be used for expressing other powerful optimisations.

Block tiling

Level-aware constructs can also be used for expressing other powerful optimisations.



Threads accessing same memory can cooperatively cache it in on-chip memory.

Motivation for block tiling

```
map (\x -> redomap (+) (\y -> y + x) 0 xs) xs
```

After flattening we get this inner-sequential version:

$$\mathbf{segmap}_{\text{thread}}\ \langle \mathtt{x} \in \mathtt{xs} \rangle\ (\mathbf{redomap}\ (+)\ (\lambda \mathtt{y} \rightarrow \mathtt{y} + \mathtt{x})\ 0\ \mathtt{xs})$$

Operation One thread for each element of xs, and each sequentially traverses xs.

Problem ?

Motivation for block tiling

```
map (\x -> redomap (+) (\y -> y + x) 0 xs) xs
```

After flattening we get this inner-sequential version:

$$\mathbf{segmap}_{\text{thread}}\ \langle \mathtt{x} \in \mathtt{xs} \rangle\ (\mathbf{redomap}\ (+)\ (\lambda \mathtt{y} \to \mathtt{y} + \mathtt{x})\ 0\ \mathtt{xs})$$

Operation One thread for each element of `xs`, and each sequentially traverses `xs`.

Problem Poor utilisation of memory bus.

- Many threads simultaneously read same address, which is redundant.
- **Better:** *cooperatively* copy *block* into on-chip memory and iterate from there.

Strip mining/chunking the outer segmap

$$\texttt{\textbf{segmap}}_{\text{thread}}\ \langle \texttt{x} \in \texttt{xs}\rangle\ (\texttt{\textbf{redomap}}\ (+)\ (\lambda \texttt{y} \rightarrow \texttt{y} + \texttt{x})\ 0\ \texttt{xs})$$

Assuming we can split `xs` into `m` equally sized *tiles* each of size `t`, giving `xss : [m][t]f32`, then we can rewrite to

$$\begin{array}{l}\texttt{\textbf{segmap}}_{\text{group}}\ \langle \texttt{xs'} \in \texttt{xss}\rangle \\ \quad \texttt{\textbf{segmap}}_{\text{thread}}\ \langle \texttt{x} \in \texttt{xs'}\rangle \\ \quad\quad \texttt{\textbf{redomap}}\ (+)\ (\lambda \texttt{y} \rightarrow \texttt{y} + \texttt{x})\ 0\ \texttt{xs}\end{array}$$

Question: does this compute the same value as the original?

Strip mining/chunking the outer segmap

$$\mathbf{segmap}_{\text{thread}}\ \langle \mathtt{x} \in \mathtt{xs}\rangle\ (\mathbf{redomap}\ (+)\ (\lambda \mathtt{y} \rightarrow \mathtt{y} + \mathtt{x})\ 0\ \mathtt{xs})$$

Assuming we can split `xs` into `m` equally sized *tiles* each of size `t`, giving `xss : [m][t]f32`, then we can rewrite to

$$\begin{array}{l} \mathbf{segmap}_{\text{group}}\ \langle \mathtt{xs'} \in \mathtt{xss}\rangle \\ \quad \mathbf{segmap}_{\text{thread}}\ \langle \mathtt{x} \in \mathtt{xs'}\rangle \\ \quad\quad \mathbf{redomap}\ (+)\ (\lambda \mathtt{y} \rightarrow \mathtt{y} + \mathtt{x})\ 0\ \mathtt{xs} \end{array}$$

Question: does this compute the same value as the original?

- *No*—the original expression had type `[n]f32`, while this has type `[m][t]f32`
- This can be `flattened` away.

```
segmap_group ⟨xs' ∈ xss⟩
  segmap_thread ⟨x ∈ xs'⟩
    redomap (+) ($\lambda$y → y + x) 0 xs
```

Chunking/strip-mining the **redomap**, we get

```
segmap_group ⟨xs' ∈ xss⟩
  segmap_thread ⟨x ∈ xs'⟩
    loop acc = 0 for ys in xss do
      redomap (+) ($\lambda$y → y + x) acc ys
```

```
segmap_group ⟨xs' ∈ xss⟩
  segmap_thread ⟨x ∈ xs'⟩
    redomap (+) (λy → y + x) 0 xs
```

Chunking/strip-mining the **redomap**, we get

```
segmap_group ⟨xs' ∈ xss⟩
  segmap_thread ⟨x ∈ xs'⟩
    loop acc = 0 for ys in xss do
      redomap (+) (λy → y + x) acc ys
```

Distributing and interchanging **segmap**$_{\text{thread}}$ gives

```
segmap_group ⟨xs' ∈ xss⟩
  loop accs = replicate t 0
  for ys in xss do
   segmap_thread ⟨x, acc ∈ xs', accs⟩
     redomap (+) (λy → y + x) acc ys
```

```
$\mathbf{segmap}_{\text{group}}$ ⟨xs' ∈ xss⟩
  loop accs = replicate t 0
  for ys in xss do
    $\mathbf{segmap}_{\text{thread}}$ ⟨x, acc ∈ xs', accs⟩
      redomap (+) (λy → y + x) acc ys
```

Collectively copy ys to shared/local memory

```
$\mathbf{segmap}_{\text{group}}$ ⟨xs' ∈ xss⟩
  loop accs = replicate t 0
  for ys in xss do
    let ys' = copy ys in
      $\mathbf{segmap}_{\text{thread}}$ ⟨x, acc ∈ xs', accs⟩
        redomap (+) (λy → y + x) acc ys'
```

- Now the many iterations of the **redomap** read from fast on-chip memory rather than slower global memory!
- **copy** done collectively by all threads in group

The fine print

```
map (\x -> redomap (+) (\y -> y + x) 0 xs) xs
```

to

```
segmap$_{\text{group}}$ $\langle$xs' $\in$ xss$\rangle$
  loop accs = replicate t 0
  for ys in xss do
    let ys' = copy ys in
      segmap$_{\text{thread}}$ $\langle$x, acc $\in$ xs', accs$\rangle$
        redomap $(+)$ $(\lambda$y $\rightarrow$ y $+$ x$)$ acc ys'
```

- Very simple case (e.g. `xss` traversed in both loops)
- 2D tiling much more complex
- The *tile size* `t` is a sensitive tuning parameter; in this case it should coincide with workgroup size
- Appreciate what a compiler can do for you

Summary

- There is no *one size fits all:* for optimal performance, we need different amounts of parallelisation for different workloads.
- Incremental flattening generates a *single program* that for varying datasets exploits only as much parallelism as profitable.
- Autotuning for specific hardware and program is needed to select the optimal version at runtime.
- A good IR is as crucial to a compiler as a good language is to a human.