Compiling a functional array language with non-semantic memory information

Philip Munksgaard  
philip@munksgaard.me  
University of Copenhagen  
Denmark

Cosmin Oancea  
cosmin.oancea@diku.dk  
University of Copenhagen  
Denmark

Troels Henriksen  
athas@sigkill.dk  
University of Copenhagen  
Denmark

ABSTRACT

We present a technique for introducing a notion of memory in the compiler representation for a parallel functional array language, in a way that allows for transformation and optimization of memory access patterns and uses, while preserving value-based semantics.

Functional languages expose no notion of memory to the user. There is no explicit allocation or deallocation of memory, nor any mapping from arrays to memory locations. Instead, the compiler must infer when and where to perform allocations as well as how values should be represented in memory. Because this affects performance, an optimizing compiler will need freedom to express various memory optimizations. Typically this is done by lowering the functional language to an imperative intermediate representation with a direct notion of memory, losing the ability to use high-level functional reasoning while performing said optimizations. We present a compiler representation where memory information is non-semantic, in the sense that it does not affect the program result, but only the operational details.

We start by defining a simple functional language without memory, Fun, and give it static and dynamic semantics. Next, we define an extended language, Fun\textsuperscript{Mem}, which is Fun with non-semantic memory information in the form of LMADs and an allocation statement. We give the extended language static and dynamic semantics and we provide an algorithm for transforming a Fun program into Fun\textsuperscript{Mem}. We likewise introduce a simple imperative language Imp, which we use to illustrate how one might translate Fun\textsuperscript{Mem} into lower-level code. Finally, we show an example of a useful transformation in Fun\textsuperscript{Mem}, memory expansion, which is used to hoist allocations from parallel GPU kernels.

ACM Reference Format:

1 INTRODUCTION

Functional array languages enable easy manipulation of multi-dimensional arrays, for example by means of (i) second-order array combinators, such as map, reduce and scan, that take arrays as inputs and produce new arrays as output; and (ii) change-of-layout operations, such as slicing, transposition or reshaping, that create new arrays by re-ordering or selecting a subset of existing arrays, but without changing their values. Such languages are promising for high level programming of parallel computers, such as GPUs.

Generating efficient array code depends on the compiler’s ability to make good choices for how arrays should be allocated and represented in memory. For example:

(i) destination-passing style [13] has been used to avoid unnecessary copying of the array result of higher-order functions,

(ii) functional representations of arrays, such as pull arrays [1], model fusion of parallel loops by means of function composition, thus potentially replacing accesses to memory with much cheaper accesses to registers, and

(iii) region inference [14] clusters objects together whose life times are lexically bounded in order to optimize (de)allocations and to guarantee the absence of memory leaks.

Most related approaches, e.g., (1-2), use code transformations that indirectly optimize memory, but do not directly support a notion of memory in the intermediate representation (IR). Instead, memory is typically introduced when switching to an imperative representation. For example, destination-passing style cannot perfectly accommodate arbitrarily nested parallel operations, and operations that cannot be fused may require manifesting results of change-of-layout operations in memory, which is sub-optimal.

Region inference [14] (3) does propose a non-semantic memory extension of the IR, but solely for tracking the lifetime of objects in order to ensure their efficient allocation and deallocation, not for expressing object layout or eliminating unnecessary copying.

This paper presents an extension of a functional IR with a non-semantic notion of memory that (i) enables efficient lowering of the memory-agnostic IR and (ii) is amenable to further optimizations. This IR is a simplified form of the one used in the compiler for Futhark [4], a functional array language.

Our IR supports allocation of memory blocks, and, more importantly, arrays are statically associated with a memory block \( m \) together with an index function \( L \) that dictates the memory layout of its elements: e.g., for a \( k \)-dimensional array, the element at index \( [i_1, \ldots, i_k] \) will be laid out at the flat memory offset \( L(i_1, \ldots, i_k) \) in \( m \). This representation has the advantage that, e.g., transposing or slicing an array just requires computing a new index function at compile time, with no additional allocation/manifestation required.

We use linear-memory access descriptors [5, 10], LMADs, as the representation for the index function. An LMAD consists of a global

\[ m \]

In this context non-semantic means that deleting the memory annotations in the IR results in a valid, semantically-equivalent program.
offset together with the number of elements and a total stride\(^{2}\) for each dimension of the array. The lmad-based representation:

- has negligible runtime overhead, i.e., only requires carrying around and performing computations on a couple of integers per dimension,
- is closed under affine index transformations,\(^{3}\)
- represents sets of quasi-affine indices, and as such is amenable to further index-based analyses.

Our discussion is through two toy languages—Fun being memory agnostic and FunMem being an extension of Fun with memory information—and we present type rules and big step operational semantics for each, as well as an algorithm that translates a Fun program to a FunMem program. In principle, this algorithm can avoid memory manifestation for an array produced by any sequence of affine layout transformations, even when returned from if expressions. Although we do not prove the correctness of this algorithm, we do state the properties that must hold for the transformation to be correct. We also show a simple imperative language Imp, which we use to illustrate how a compiler might translate FunMem into low-level code, by giving a translation from FunMem to Imp.

We demonstrate that FunMem is a suitable IR for optimizations by presenting an algorithm for memory expansion in section 5. Since dynamic allocation is not efficiently supported from inside GPU kernels, memory expansion hoists inner allocations out of kernels by creating a single large memory block allocated in advance, which is then divided among the threads in non-overlapping (but possibly interleaved) chunks. We show that memory expansion can be expressed for FunMem in a simple and elegant way that ensures spatial locality on GPUs (i.e., coalesced access to global memory).

Contrary to Fun, FunMem does not guarantee that parallelism is correct by construction.\(^{4}\) However, it allows the compiler to perform optimizations that are not expressible in Fun, such as

- re-using memory across arrays whose life spans do not overlap—think register allocation on arrays, or
- allowing parallel algorithms to both read and write its input arrays in-place, as showcased in \([9]\).\(^{5}\)

These transformations are briefly discussed in section 6 and have been demonstrated to result in significant speedups \([9]\), but are otherwise out of scope for this paper.

In summary, the main contributions of this paper are:

1. specifying the static and dynamic semantics of FunMem,
2. presenting the translation from Fun to FunMem, and stating the properties that guarantee the translation correctness,
3. demonstrating the memory expansion transformation on FunMem, which is essential for efficient execution on GPUs.

### 2 The Fun Language

We start by defining a tiny purely functional size-dependently typed core language, corresponding to a subset of the IR used in the Futhark compiler. We concern ourselves solely with expressions; not function definitions as a whole. This is insufficient to express real programs, but sufficient to describe our approach. The language can be thought of as a first-order monomorphic functional language with parallel loops.

The grammar is shown in fig. 1. We assume a denumerably infinite set of program variables, ranged over by \(x, y, z\). We will also use superscripts and subscripts to distinguish distinct variables. We write \(\sigma\) where \(\sigma\) is some syntactical metavariable for a sequence of \(x\) whenever we do not need to address individual terms. We write \(FV(a)\) for the free variables of \(a\).

All variable bindings are of the form \((x : \tau)\); variables can be either integers or multi-dimensional arrays of integers, with explicit sizes for each dimension. A statement \((\sigma)\) is a binding of an expression \((e)\) with some variable bindings (called the pattern). A body \((b)\) is a sequence of statements terminated by a result.

Expressions only occur in statements, meaning that the language is in administrative normal form \([11]\), which for our purposes strongly resembles SSA. Expressions can be either scalar expressions or expressions operating on and creating arrays, and most of them are straightforward. Slices are of the form \(offset : size\). kernel-expressions denote parallel execution of a body, where

\[
\text{kernel } x \leq y \text{ do } b
\]

executes \(b\) \(y\) times in parallel, where in each thread \(x\) is bound to the thread-id, between 1 and \(y\) (both inclusive). Each invocation of \(b\) produces the corresponding element of the final array. A kernel-expression returns only one array, although the language could easily be extended to allow for multiple return values.

We also define the derived forms

\[
\begin{align*}
\text{iota } x & \equiv \text{kernel } x_1 \leq x \text{ do in } (x_i) \\
\text{copy } x & \equiv \text{kernel } y_1 \leq z_1 \cdots \text{kernel } (y_n \leq z_n) \text{ do in } (x[y_1, \ldots, y_n])
\end{align*}
\]

where \(z_1 \ldots z_n\) are the \(n\) dimensions of \(x\).

#### 2.1 Type Rules

The type rules for Fun are shown in fig. 3 and mostly standard. The most interesting part is that whenever an array variable is bound, the sizes of that array must be in scope; possibly by being bound within the same pattern. Note that in \(\Gamma\)-Ir, we use \(FV(\sigma)\) to mean free variables in the types of each pattern, specifically array sizes, not the bound names themselves.

The judgement \(\Gamma \vdash s\) states that \(s\) is a well-typed statement in context \(\Gamma\). We write \(\bullet\) for empty contexts and in general for empty sequences. The judgement \(\Gamma \vdash \sigma \leftarrow b\) states that the body \(b\) can be type-checked in context \(\Gamma\), and its results bound to the bindings \(\sigma\). The base case uses a substitution \(S\) to require that any arrays returned also have their sizes similarly returned.

The judgement \(\vdash b : (r_1, \ldots, r)\) states that \(b\) is a well-typed program, meaning it is a body with no free variables that returns arrays of the indicated constant sizes. The restriction is merely to keep the rules simple, and would not be present in a real implementation.
2.2 Operational semantics

The operational semantics for Fun are shown in fig. 4, and constitute a standard value-based semantics. Note that evaluation can “get stuck” for out-of-bounds indexing. Arrays are 1-indexed. We rely on some auxiliary functions for indexing, slicing, and transposing arrays, which are defined in fig. 5. The functions “index” and “slice” perform indexing and slicing of arrays, while “transpose” transposes an array. The “nest” function is used in section 3.3.1 to unflatten a one-dimensional array. For instance

\[
\text{nest}([1, 2, 3, 4], [2][2]) = [[1, 2], [3, 4]].
\]

For brevity, we write \(E(\alpha)\) to denote substituting every variable in the syntactic construct \(\alpha\) with its value from \(E\), but also to evaluate any arithmetic operations now applied to constants, following conventional arithmetic rules.

3 THE FUNMEM LANGUAGE

FunMem is an extension of Fun with a new type \(\text{mem}\), a new expression form \(\text{alloc}\), and a change to the syntax of variable bindings such that arrays denote their memory block and index function. The syntax is shown in fig. 6.

In FunMem, \(\text{kernel}\) creates a new array in memory—we call these fresh arrays. We have complete freedom to decide the index function of fresh arrays, by specifying the desired index function in the pattern. Other expressions, such as transpose or slicing, produce derived arrays. These arrays are index space transformations of a prior array, and their index functions are derived from the index function of the prior array. The intuition here is that such transformations are free at runtime, which is only possible if the change-of-indexing can be implemented entirely at compile time.

An example of a FunMem expression can be seen in fig. 7.

Note that this representation allows multiple arrays to concurrently co-exist in the same memory block. This is a crucial feature of the representation, but it is only safe if uses or lifetimes of the two arrays do not overlap. Although our type rules themselves do not verify this, we will return to this issue in section 3.3.1.

3.1 LMADs in FunMem

A \(q\)-dimensional array must be associated with a \(q\)-dimensional LMAD, describing the offset as well as the stride and number of elements of each dimension. Application of an LMAD

\[
L = se + \{(x_1^\text{elems} : x_1^\text{stride}), \ldots, (x_q^\text{elems} : x_q^\text{stride})\}
\]

defines as:

\[
L(y_1, \ldots, y_q) = se + \sum_{1 \leq i \leq q} y_i \cdot x_i^\text{stride}
\]

(1)

Further, the function \(\text{slice}\_f(L, [y_1 : z_1], \ldots, [y_q : z_q])\) defines slicing of a \(q\)-dimensional LMAD \(L\) with \(q\) offset:stride pairs:

\[
\text{slice}_f(L, [y_1 : z_1], \ldots, [y_q : z_q]) = (se + \sum_{1 \leq i \leq q} y_i x_i^\text{stride}) + \{(z_1 : x_1^\text{stride}), \ldots, (z_q : x_q^\text{stride})\}
\]

(2)

Similarly, \(\text{index}\_f(L, k)\) fixes the outermost dimension of a \(q\)-dimensional LMAD, producing a \(q - 1\)-dimensional LMAD:

\[
\text{index}_f(L, k) = (se + k \cdot x_1^\text{stride}) + \{(x_2^\text{elems} : x_2^\text{stride}), \ldots, (x_q^\text{elems} : x_q^\text{stride})\}
\]

(3)

The domain \(\text{dom}(L)\) of an LMAD \(L\) constitutes all valid indexes, while the image \(\text{img}(L)\) constitutes all addresses reachable by applying it to an in-bounds index. An LMAD can be 0-dimensional, representing the offset at which a single scalar value is stored. These do not occur in our type system, as we do not store scalars in memory, but occur in our operational semantics when defining array indexing.
In FunMem, LMADs can use variables in scope. In the operational semantics we will also use LMAD values that are assumed to contain only integer constants, not variables or expressions.

As an abbreviation we define \( R(x_1, \ldots, x_n) \) as an index function for arrays of shape \( [x_1] \cdots [x_n] \) in row-major order with zero offset.

### 3.2 Type rules

The type rules for FunMem are shown in fig. 8. The rules are structured similarly to the ones for Fun. The central judgement \( \Gamma \vdash s \) states the well-typedness of statement \( s \) in a type context \( \Gamma \). Note how T-Mem-Kernel lets us pick any index function for the result, as long as it supports an array of the proper shape.

The judgement \( \Gamma \vdash \overline{p} \leftarrow b \) states that in a context \( \Gamma \), the body \( b \) produces results matching the pattern \( \overline{p} \). As with Fun, we use a substitution \( S \) but this time in addition to array sizes, it requires that any arrays returned also have all their supporting information (sizes, memory blocks, variables used in index functions) returned.

The judgement \( \Gamma \vdash L : [x_1] \cdots [x_n] \) states that in a context \( \Gamma \), \( L \) is a well-typed index function for an array of type \( [x_1] \cdots [x_n] \). Similarly, the judgement \( \Gamma \vdash r \) states that the context \( \Gamma \), \( r \) is well-typed, meaning for arrays that the memory block is bound in \( \Gamma \) and that the LMAD is well-typed.

The judgement \( \vdash b : (r, \ldots, r) \) states that \( b \) is a well-typed program returning arrays of the sizes indicated.

The type system is by design unsound, and FunMem programs that are well-typed can still “go wrong”. In particular, we allow programs that perform essentially imperative in-place updates by modifying memory blocks that are in use by other arrays, as well as expressing data races by letting distinct iterations of a kernel expression write to the same memory locations. The reasoning behind this rather unusual design will be discussed in section 3.4.
we denote $\mathcal{H}$ construct the heap the value at offset $i$ are one-dimensional arrays of integers. We write cate. We make use of a fairly standard heap abstraction, which are shown on fig. 9.

<table>
<thead>
<tr>
<th>$E \vdash e \downarrow (v_1, \ldots, v_n)$</th>
<th>$E(x) \neq 0$</th>
<th>$E \vdash b_1 \downarrow (v_1, \ldots, v_n)$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$E \vdash \text{if } x \text{ then } b_1 \text{ else } b_2 \downarrow (v_1, \ldots, v_n)$</td>
<td>[E-If-True]</td>
<td>$E(x) = 0$</td>
</tr>
<tr>
<td>$E \vdash b_2 \downarrow (v_1, \ldots, v_n)$</td>
<td>[E-If-False]</td>
<td>$E \vdash \text{if } x \text{ then } b_1 \text{ else } b_2 \downarrow (v_1, \ldots, v_n)$</td>
</tr>
</tbody>
</table>

$$v = \text{conventional evaluation of se}$$ $E \vdash se \downarrow (v)$ [E-SCALAR]

$$v = \text{transpose } x \downarrow (v)$$ $E \vdash \text{transpose } x \downarrow (v)$ [E-TRANSPOSE]

$$v = \text{slice}(E(x), E(y_1), \ldots, E(y_\ell))$$ $E \vdash x[y_1, \ldots, y_\ell] \downarrow (v)$ [E-SLICE]

$$E, x \mapsto 1 \vdash b \downarrow (v_1)$$ $E, x \mapsto m \vdash b \downarrow (v_m)$ [E-KERNEL]

$$E \vdash b \downarrow (v_1, \ldots, v_n)$$ $E \vdash \text{let } (x_1 : \tau_1) \cdots (x_n : \tau_n) = e \ b \downarrow (v_1', \ldots, v_n')$ [E-LET]

$$E(x_1) = v_1 \text{ for } 0 < i \leq n$$ $E \vdash \text{in } (x_1, \ldots, x_n) \downarrow (v_1, \ldots, v_n)$ [E-IN]

Figure 4: Big-step operational semantics for Fun. We make use of some auxiliary functions from fig. 5.

Figure 5: Auxiliary functions for transforming values.

heap-based semantics that is close to how it would be implemented on a machine. A FunMem program is valid only if its interpretation under these two semantics coincides; a notion we will make precise in section 3.3.1.

The value-based semantics is based on reducing the FunMem program to Fun and then applying the rules from fig. 4. Intuitively, this is done by turning all memory blocks into integers, all allocations into dummy integer literals, and replacing array memory information with the corresponding array type. The pertinent rewrite rules are shown on fig. 9.

The heap-based semantics, shown in fig. 10, is more complicated. We make use of a fairly standard heap abstraction, which we denote $H$, that maps heap labels $\ell$ to memory blocks, which are one-dimensional arrays of integers. We write $H[\ell, i]$ to look up the value at offset $i$ in the block with label $\ell$, and $H[\ell, i] \mapsto \ell'$ to construct the heap $H$ with the value at offset $i$ in block $\ell$ changed to $\ell'$. The alloc statement is used to create new labels and add the corresponding mapping to the heap.

The central judgement is

$$E; H \vdash s \downarrow E'; H' ; (\mathcal{R}, \mathcal{W})$$

which denotes the evaluation of a statement $s$ in a value environment $E$, and heap $H$, yielding an extended value environment $E'$ and modified heap $H'$, as well as a trace of memory accesses. The trace consists of two sets of pairs of heap labels and offsets: a read set $\mathcal{R}$ and a write set $\mathcal{W}$. The trace is not semantically significant, but is used as a side condition in the rule $E; \text{Mem-Kernel}$ to avoid data races, by saying that a location written in one iteration may not be used in any way by another. This is intended to allow an implementation to execute the iterations of a kernel construct concurrently. Given the read and write sets of $k$ iterations, we define

$$\mathcal{R}_k = \{ (\ell_1, m_1), \ldots, (\ell_k, m_k) : \ell_1, \ldots, \ell_k \in \mathcal{R}, m_1, \ldots, m_k \in \{ 0, \ldots, \ell - 1 \} \}$$

and

$$\mathcal{W}_k = \{ (\ell_1, m_1), \ldots, (\ell_k, m_k) : \ell_1, \ldots, \ell_k \in \mathcal{W}, m_1, \ldots, m_k \in \{ 0, \ldots, \ell - 1 \} \}.$$
heap label and index function, yielding a new heap. If the value to
that defines how to copy a value to a specified destination, given as
and an index function containing only constants.

The reasoning behind FunMem’s design is that it is intended as
an internal representation in a research compiler, not as a formalism
or calculus. While fully verified compilers such as CompCert [7]
or CakeML [6] exist, implementing fully verified compiler passes
remains very time-consuming. Our design prioritizes ease of trans-
formation over ease of verification. Therefore, while we have tried
to make it precise which properties must hold for a program (valid-
ity and absence of data races), so it can be argued whether a specific
$\Gamma \vdash s$

$$\begin{array}{c}
\Gamma \vdash se : \text{int} \quad [\text{T-MEM-SCALAR}]
\Gamma \vdash se : \text{int} \quad [\text{T-MEM-ALLOC}]
\end{array}$$

$\Gamma \vdash \text{let} (x : \text{int}) = se$

$\Gamma \vdash (y : [z_1, \ldots, z_n]@y_{\text{mem}} \mapsto L) \in \Gamma \quad \forall i. (y_i : \text{int}) \in \Gamma$

$$\begin{array}{c}
\Gamma \vdash \text{let} (x : \text{int}) = y[y_1, \ldots, y_n] \quad [\text{T-MEM-INDEX}]
\end{array}$$

$\frac{(x : \text{int}) \in \Gamma \quad \Gamma \vdash \bar{p} \leftarrow b_1 \quad \Gamma \vdash \bar{p} \leftarrow b_2 \quad \forall(y : \tau) \in \bar{p}, \Gamma \vdash \bar{r}}{\Gamma \vdash \text{let} \bar{p} = \text{if} x \text{ then } b_1 \text{ else } b_2} \quad [\text{T-MEM-IF}]

$\Gamma \vdash (y : \text{int}) \in \Gamma \quad \Gamma, (x : \text{int}) \vdash (z : \tau) \leftarrow b \quad \Gamma \vdash [y]r$

$$\begin{array}{c}
\Gamma \vdash \text{let} \bar{r} = \text{kernel } x \leq y \text{ do } b \quad [\text{T-MEM-KERNEL}]
\end{array}$$

$\Gamma \vdash \bar{p} \leftarrow b$

$$\begin{array}{c}
S = \{y_1 \mapsto x_1, \ldots, y_n \mapsto x_n\}
(x_1 : S(\tau_1)) \in \Gamma \quad \ldots \quad (x_n : S(\tau_n)) \in \Gamma
\end{array}$$

$$\Gamma \vdash (y_1 : \tau_1) \cdots (y_n : \tau_n) \leftarrow \text{in} (x_1, \ldots, x_n) \quad [\text{T-MEM-RESULT}]

$$\begin{array}{c}
\Gamma \vdash \text{let} \bar{r}' = e \quad \Gamma, \bar{p} \vdash \bar{p} \leftarrow b \quad [\text{T-MEM-STM}]
\end{array}$$

$$\begin{array}{c}
\Gamma \vdash \text{let} \bar{p}' \leftarrow e \quad \Gamma \vdash \bar{p} \leftarrow \text{let} \bar{p}' = e \ b \quad [\text{T-MEM-STM}]
\end{array}$$

$\Gamma \vdash L : [x_1] \cdots [x_n]$

$$\begin{array}{c}
L = se_0 + \{(x_1 : y_1) \cdots (x_n : y_n)\} \quad \forall z \in \text{FV}(L). (z : \text{int}) \in \Gamma
\end{array}$$

$$\begin{array}{c}
\Gamma \vdash L : [x_1] \cdots [x_n] \quad [\text{T-MEM-LMAD}]
\end{array}$$

$\Gamma \vdash \tau$

$$\begin{array}{c}
\Gamma \vdash \text{int} \quad [\text{T-MEM-INT}]
\end{array}$$

$$\begin{array}{c}
\frac{(x_{\text{mem}} : \text{mem}) \in \Gamma \quad \Gamma \vdash L_x : [x_1] \cdots [x_n]@x_{\text{mem}} \mapsto L}{\Gamma \vdash [x_1] \cdots [x_n]@x_{\text{mem}} \mapsto L_x} \quad [\text{T-MEM-ARRAY}]
\end{array}$$

$\vdash b : (r, \ldots, r)$

$$\frac{\forall l. p_{l_{\text{mem}}} \vdash p_{l_{\text{cal}}} \vdash b}{\vdash b : (k_{i,1}, \ldots, k_{i,m_i}@x_{j_{\text{mem}}} \mapsto R(k_{i,1}, \ldots, k_{i,m_i}))} \quad [\text{T-MEM-PROG}]
$$
location and index function of the array being sliced or transposed so we can add the necessary information in the result pattern.

For kernel expressions, we first transform the inner bodies. Then we create a fresh variable for the memory block of the result and create the allocation statement. We then construct a row-major index function and finally combine it all in the pattern $p'$.

The case for if is the most complicated, as we need to return the supporting information of any arrays in the return values. We start by transforming the statements inside each body. Then, for each pair of values returned from the branches and the pattern being matched with, we do the following: First we lookup the result types, then extract supporting information using the SUPPORT procedure, as defined in procedure Support. The SUPPORT function also creates a fresh variable for the scalar expression denoting the LMAD offset, and a statement binding it. We append these statements to the transformed list of statements for each body. Then, for each supporting variable, we create a corresponding variable for the outer pattern and a substitution, which we apply to get the transformed type of the pattern.

Note that we return all supporting information, even though it might not be necessary. For instance, the size of the arrays returned in two branches of an if-statement could be invariant to the branch, meaning that the branches always return arrays of the same size. However, we defer the removal of such redundant return values to a fairly straightforward simplifier in a later pass.

### Procedure TransformProgram($\mathcal{P}$)

<table>
<thead>
<tr>
<th>input</th>
<th>A well-typed Fun program $\mathcal{P}$ is $\mathcal{X}$ in $\Gamma$.</th>
</tr>
</thead>
<tbody>
<tr>
<td>output</td>
<td>A FunMem program corresponding to $\mathcal{P}$.</td>
</tr>
</tbody>
</table>

$\mathcal{X}, \Gamma \leftarrow \text{TransformStms}(\mathcal{X}, \bullet)$,

$\overline{\mathcal{F}} \leftarrow \bullet$,

foreach $x_i$ in $\mathcal{X}$ do

$[x_i]@\text{mem} \leftarrow L \leftarrow \text{lookup}(x_i, \Gamma)$;

$\text{mem}', y \leftarrow \text{fresh}$;

$s_{\text{alloc}} \leftarrow \text{let (mem' : mem) = alloc } ([x] z)$;

$s_{\text{lin}} \leftarrow \text{let (y : } [x]@\text{mem}' \rightarrow \mathcal{R}(\mathcal{F}) \text{) = copy } x_i)$;

$\text{Append } s_{\text{alloc}, s_{\text{lin}}} \text{ to } \overline{\mathcal{F}}$;

$\text{Append } y \text{ to } \overline{\mathcal{F}}$;

return $s'$ in $\overline{\mathcal{F}}$.

### 4 THE IMP LANGUAGE

Imp is a simple imperative language with a parallel loop construct, and is used to show that translating FunMem to imperative code is straightforward.

Figure 11 shows the grammar for Imp, wherein we reuse the scalar expressions from Fun. It supports two types: int and mem,

the latter of which is a single-dimensional integer array that must be allocated before use. Variables can be declared and assigned. Finally, conditionals are supported in the form of if, while kernel constitutes a parallel loop where each thread gets a thread identifier $x$. The type rules for Imp are trivial and uninteresting, so we will elide them here. An example of a program is shown in fig. 13.

Figure 12 shows the dynamic semantics for Imp. The rules are fairly conventional. We have a mutable environment $E$ that maps variable names to values, and a heap $H$ that maps labels to memory blocks. The evaluation judgement

$H : E \cdot s \rightarrow H', E'$

is read "evaluation of statement $s$ with heap $H$ and environment $E$ produces a new heap $H'$ and environment $E'$." The actual rules are straightforward. The alloc statement creates a new label and adds a corresponding fresh block to the heap.

Procedure FunMemToImp translates a FunMem program to Imp, which is somewhat tedious but fairly straightforward. The most significant detail is that we must apply index functions symbolically to translate accesses of the multidimensional arrays of FunMem into the single-dimensional arrays of Imp, which can be seen in e.g. the case for array indexing.

### 5 MEMORY EXPANSION

Memory expansion (MEMORYEXPAND) is a technique for hoisting allocations out of kernels. This is critical for compilation targeted at GPUs, since GPU kernels cannot efficiently allocate memory. Although neither Fun nor FunMem prohibit allocations inside kernel expressions, generation of working GPU code must use memory expansion to hoist out such allocations. This is accomplished by:

1. allocating shared memory blocks big enough to accommodate the sum of memory requirements of all threads—i.e., of all (parallel) iterations of the kernel body, and

2. modifying the index functions of the arrays created inside the kernel body to refer to the shared blocks, in a way that satisfies the property that any two threads use non-overlapping partitions of the shared block.

For simplicity of exposition we treat here only the case when the size of the inner allocation is the same for all threads, i.e., invariant
\[ E;H \vdash s \Downarrow E;H; (\mathcal{R},W) \]

\[
E(s) = m \quad \text{fresh} \quad E' = E, y \mapsto t \quad H' = H, t \mapsto [0, \ldots, 0] \quad \text{[E-Mem-Alloc]} 
\]

\[
E(x) = (t_x, L_x) \quad E' = E, z \mapsto (t_x, \text{slice}_f(L_x, \{E(y_l) : E(y_{l+1}), \ldots, E(y_{2n-1}) : E(y_{2n})\})) \quad \text{[E-Mem-Slice]} 
\]

\[
E(x) = (t_x, L_x) \quad k = L_x(E(y_1), \ldots, E(y_n)) \quad E' = E, z \mapsto H[t_x,k] \quad \mathcal{R} = \{t_x,k\} \quad \text{[E-Mem-Index]} 
\]

\[
E(x) = (t_x, k_0 + \{(k_1 : k_2) : (k_3 : k_4)\}) \quad E' = E, z \mapsto (t_x, k_0 + \{(k_3 : k_4) : (k_1 : k_2)\}) \quad \text{[E-Mem-Transpose]} 
\]

\[
E, x \mapsto 1; H_0 \vdash b \Downarrow E; H; \langle \mathcal{R}_1, W_1 \rangle \quad (H'_1, \mathcal{R}'_1, W'_1) = \text{memcpy}(H_1, E(z_{res}), \text{zmem}, \text{index}(L_z, 1)) 
\]

\[
E, x \mapsto k; H_{k-1} \vdash b \Downarrow E; H_k; \langle \mathcal{R}_k, W_k \rangle \quad (H'_{k}, \mathcal{R}'_{k}, W'_{k}) = \text{memcpy}(H_k, E(z_{res}), \text{zmem}, \text{index}(L_z, k)) 
\]

\[
E(y) = k \quad \text{racefree}(\mathcal{R}_1 \cup \mathcal{R}_1', \ldots, \mathcal{R}_k \cup \mathcal{R}_k', W_1 \cup W_1', \ldots, W_k \cup W_k') \quad E' = E, z \mapsto (E(z_{mem}), E(L_z)) \quad \text{[E-Mem-Kernel]} 
\]

\[
E(x) \neq 0 \quad E; H \Downarrow \bar{x} \Downarrow E; H; \langle \mathcal{R}_x, W_x \rangle \quad E' = E, x_1 \mapsto E(x_1), \ldots, x_n \mapsto E(x_n) \quad \text{[E-Mem-IF-T]} 
\]

\[
E(x) = 0 \quad E; H \Downarrow \bar{x} \Downarrow E; H; \langle \mathcal{R}_x, W_x \rangle \quad E' = E, x_1 \mapsto E(x_1), \ldots, x_n \mapsto E(x_n) \quad \text{[E-Mem-IF-F]} 
\]

\[
E; H \Downarrow \bar{x} \Downarrow E; H; \langle \mathcal{R}, W \rangle 
\]

\[
E; H \Downarrow \text{let } \bar{c}_1 = c_1 \Downarrow E; H; \langle \mathcal{R}_1, W_1 \rangle \quad \ldots \quad E; H \Downarrow \text{let } \bar{c}_n = c_n \Downarrow E; H_n; \langle \mathcal{R}_n, W_n \rangle \quad E; H \Downarrow \text{let } \bar{c}_1 \quad \ldots \quad \text{let } \bar{c}_n = c_n \Downarrow E; H_n; \langle \bigcup_{i=1}^{n} \mathcal{R}_i, \bigcup_{i=1}^{n} W_i \rangle \quad \text{[E-Stms]} 
\]

\[
\vdash b \Downarrow (v_1, \ldots, v_n); H 
\]

\[
\Downarrow \bar{x} \Downarrow (E(x_1), \ldots, E(x_n)); H \quad \text{[E-Mem-Prog]} 
\]

Figure 10: Heap-based big-step operational semantics rules for FUNMEM. The boxed parts serve to detect data races, but are not otherwise significant for the evaluation result. We use the definition of \text{slice}_f from eq. (3) and racefree from eq. (4). For space reasons we elide the rules for scalar expressions, as they are conventional.
The procedure MemoryExpand performs memory expansion on a FUNMem program. It assumes that allocations and kernel-invariant computations have been hoisted as much as possible. The procedure repeatedly pattern matches the first statement \( s \) of a kernel body with an allocation of a memory block \( z \), which is not used by the result of the kernel body. If the match succeeds, the procedure first creates a fresh variable \( z' \) to hold the allocation of the hoisted variable. Then, all array types referencing \( z \) inside the kernel-body are replaced with array types referencing \( z' \), by means of a (new) index function that exhibits a kernel-variant offset, i.e.,
with an outer dimension whose length is equal to the number of
which depends on the thread number. Finally, s is replaced with
two new statements: the allocation of z' and the kernel-statement with the updated body.

The modification of the index function shown in procedure MemoryExpand essentially corresponds to expanding the original array
with an outer dimension whose length is equal to the number of
kernel threads y, and to position the local array of thread x to the
index x of the expanded array. This leads to good spatial locality
for CPU-based systems—consecutive accesses of the same thread
are localized—but is very inefficient for GPU execution because
it would result in fully-uncoalesced accesses to global memory,
and consecutive threads would access the memory with a stride
equal to the array size. For GPU execution, the strategy is to insert
the expanded dimension innermost (i.e., maintaining the array in
transposed form), which is accomplished with the index function:

\[
se' + x + \{(x_1 : x_{n+1} \cdot y) \cdots (x_n : x_{2n} \cdot y)\}
\]

which is then replaced with the following expression:

\[
se' + x + \{(x_1 : x_{n+1} \cdot y) \cdots (x_n : x_{2n} \cdot y)\}
\]

The buffers for each iteration are now interleaved, but the form of
the lmad s ensure that the reads and writes from each thread do not
overlap. This procedure shows how an lmad-based representation
lends itself well to high-level analysis and optimization.

6 REMARKS ON IMPLEMENTATION
The Fun and FunMem languages presented above are simplified
forms of the IR used in the Futhark compiler. The real IRs contain
more constructs—in particular more index transformations than just
transpose, and more ways to construct fresh arrays than kernel. However, the core concepts, particularly the use of lmad s and the
way the necessary supporting information is passed around, is identical. The Futhark implementation also supports functions, which
are implemented by adding memory blocks and lmad information
as parameters to functions.

The compiler uses type rules similar to those in fig. 8 to sanity-
check the result of compiler passes. This has proven very useful
for catching compiler bugs that might otherwise result in memory
corruption at run time. As discussed in section 3.4, these rules do
not rule out all memory errors, and there have certainly been cases
where a faulty optimization silently caused miscompilation.

We have also used lmad s for two other optimizations:

Memory block merging, which uses the same memory block
for multiple arrays, as long as their lifetimes do not overlap. This
can reduce memory footprint in cases where dynamic memory
allocation is not practical, such as within GPU kernels.

Memory short circuiting, discussed in detail in [9], which
identifies arrays that will eventually be copied to some specific
location (e.g. due to being concatenated with some other array),
and modifies the memory block and index function of the original
array such that it is constructed in-place. The “copy” can then be
with the key difference that we support non-lexical lifetimes of memory blocks, while we use index functions to describe object layout. See also the work discussed in section 1.

7 RELATED WORK

Most compilers for functional languages will eventually lower the program being compiled to a representation that makes memory allocation, typically coinciding with a general lowering of the abstraction level, such as by translating the program to an imperative form. In our approach, the memory information is an "extension" of an underlying functional language, which maintains purely functional semantics. A similar idea can be found in Destination-Passing Style [8, 13] or region-based memory management [15], although with the key difference that we support non-lexical lifetimes of memory blocks, as well as using index functions to describe object layout. See also the work discussed in section 1.

8 CONCLUSIONS

We have shown how to extend a functional language with a notion of memory that is rich enough to express layout and allocation optimizations. The memory annotations are non-semantic, although we have not provided an algorithm for verifying memory safety.

Previous work on dope vectors in ALGOL 60 [12] and APL [3] used dynamic structures similar to LMADs, but as actual metadata carried around at runtime, and with less expressive power. For instance, they cannot express the difference between row-major and column-major layouts.

The functional array language Single Assignment C (SaC) is similar to Futhark, and also uses a memory management strategy similar to ours [2], in particular distinguishing conceptual array values from the memory blocks used to store them, and using reference counting for the latter. One significant difference is that SaC's scheme uses a notion of "sub-allocations" to reference parts of memory blocks, while we use index functions, which also provides a mechanism for describing array layouts.
REFERENCES


