Ikra-Cpp: A C++/CUDA DSL for Object-Oriented Programming with Structure-of-Arrays Layout

Matthias Springer
Tokyo Institute of Technology
matthias.springer@acm.org

Hidehiko Masuhara
Tokyo Institute of Technology
masuhara@acm.org

1 Introduction

Object-oriented programming (OOP) is a popular language paradigm in general-purpose computing, but not widely used in high-performance SIMD computing due to insufficient compiler support. Object-oriented code is often several factors slower than tuned, non object-oriented code. In this paper, we present Ikra-Cpp, a DSL for object-oriented high-performance computing embedded in C++/CUDA. Ikra-Cpp allows programmers to write OOP-style code while, behind the curtain, storing data in a Structure of Arrays (SOA) representation; a well-studied best practice for SIMD architectures.

Why Structure of Arrays? Unfortunately, many performance-critical applications are unacceptably slow when expressed in an object-oriented way due to the way virtually any modern compiler structures objects in main memory: Arrays of Structures (AOS). In AOS, every object is stored as a contiguous chunk of data. This is often not ideal for SIMD architectures, which operate on a vector of values.

In Structure of Arrays (SOA; column stores in database systems), all values of a field are grouped and stored contiguously across the entire object space. SOA is a well-studied, established data layout [4, 19, 20, 24, 28, 31, 39] and CUDA best practice which can save on memory access time (memory coalescing [16]), maximize cache usage and allow for vectorization via SIMD instructions [2, 14, 15].
class Body {
  public:
  double pos_x = 0.0;
  double pos_y = 0.0;
  double vel_x = 1.0;
  double vel_y = 1.0;

  Body(double x, double y)
  : pos_x(x), pos_y(y) {}

  void move(double dt) {
    pos_x = pos_x + vel_x * dt;
    pos_y = pos_y + vel_y * dt;
  }

  void create_and_move() {
    Body* b = new Body(1.0, 2.0);
    b->move(0.5);
    assert(b->pos_x == 1.5);
  }
}

class Body : public SoaLayout<Body, 50> {
  public: IKRA_INITIALIZE_CLASS
    double_ pos_x = 0.0;
    double_ pos_y = 0.0;
    double_ vel_x = 1.0;
    double_ vel_y = 1.0;

    Body(double x, double y)
    : pos_x(x), pos_y(y) {}

    void move(double dt) {
      pos_x = pos_x + vel_x * dt;
      pos_y = pos_y + vel_y * dt;
    }

    void create_and_move() {
      Body* b = new Body(1.0, 2.0);
      b->move(0.5);
      assert(b->pos_x == 1.5);
    }
}

double Body_pos_x[50]; double Body_pos_y[50]; double Body_vel_x[50]; double Body_vel_y[50];
int Body_inst = 0;

int new_Body(double x, double y) {
  int id = Body_inst++;
  Body_pos_x[id] = x; Body_pos_y[id] = y;
  Body_vel_x[id] = Body_vel_y[id] = 1.0;
  return id;
}

void Body_move(int id, double dt) {
  Body_pos_x[id] = Body_vel_x[id] * dt;
  Body_pos_y[id] = Body_vel_y[id] * dt;
}

void create_and_move() {
  int b = new_Body(1.0, 2.0);
  Body_move(b, 0.5);
  assert(Body_pos_x[b] == 1.5);
}

(a) C++ Class (AOS Layout) (b) Ikra-Cpp: AOS Syntax, but SOA Layout (c) Hand-written SOA Layout in C++

Figure 1. Comparison of OOP Notation for a simplified 2D N-Body Simulation. Programmers want the notation of (a) but the performance of (c). With Ikra-Cpp, they get the performance of (c) with the notation of (b). A maximum of 50 objects are supported in this example.

Maintaining a SOA layout manually is troublesome. SOA code (Figure 1c) is less readable and expressive than AOS-style code (Figure 1a): Native OOP language constructs such as the new keyword or member access notation for fields cannot be used; instead, programmers must keep track of object allocations by themselves, implement constructor logic in global functions, and access field values through arrays. Methods require an explicit this parameter, and objects are referenced with integer IDs instead of class pointers.

Why a Library/DSL? Ikra-Cpp is a lightweight C++ library/embedded DSL [11] (around 2500 LoC) implemented entirely in C++ with template metaprogramming, operator overloading, helper classes and preprocessor macros. Our goal is to provide a mechanism that lets programmers write object-oriented AOS-style code (Figure 1b) while transparently laying out objects as SOA. Ikra-Cpp works with every modern C++14 compiler and the Nvidia CUDA Toolkit 9.0 or higher (in GPU mode). We originally designed Ikra-Cpp as an intermediate representation for Ikr-Ruby [32], a GPGPU library with object support for Ruby, but it can also be used standalone by C++/CUDA programmers, which is the focus of this paper. Ikra-Cpp could be implemented as a compiler extension. To the best of our knowledge, no such extension exists for a widely used language. We believe that this is due to the high engineering effort of writing a new compiler or such an invasive compiler extension [21].

Contributions and Outline The main contribution of this paper is twofold. First, to the best of our knowledge, Ikra-Cpp is the first C++ tool for SOA data layout that supports OOP features, most notably member function calls and constructors. Second, Ikra-Cpp supports many SOA features supported by projects discussed in the related work section (e.g., referencing objects with class/struct pointers instead of IDs; c.f. ispc), but with standard C++ syntax (c.f. SoAx) and without relying on an external tool or extending the language (c.f. ispc); everything is implemented in C++.

The remainder of this paper is structured as follows. Section 2 gives an overview of the architecture of Ikra-Cpp and describes its basic functionality. Section 3 explains how data is layed out, how addresses are computed and how a seamless notation can be achieved in C++. Sections 4 and 5 present first benchmarks and discuss related work. Finally, Sections 6 and 7 describe future work and conclude the paper.

2 Language Overview

In this section, we describe the basic functionality of Ikra-Cpp, focusing on host (CPU) code.

2.1 Notation

A class that is layed out as SOA is called a SOA class and its instances are called SOA objects. In Ikra-Cpp, every SOA class (Figure 1b) must inherit from SoaLayout, a class that provides useful helper methods and type aliases. The maximum number of instances of a SOA class is a compile-time constant and template parameter of SoaLayout (Line 1).

SOA objects can only be created with the new keyword and must be referenced with pointers. Stack/static allocation

Given a collection of objects of the same type and a method name, execute the method for every object (Figure 2). In host mode, this is done sequentially, but future versions of Ikra-Cpp may support thread pool execution. In device mode, Ikra-Cpp launches a CUDA kernel with (currently) one thread per object. Only methods that are annotated with the \_\_device\_\_ keyword can be executed in device mode.

### Execute and Reduce
Ikra-Cpp provides an API for combined execute and reduce operations. For example, this is useful for termination detection of iterative algorithms where the termination criteria depends on a property of multiple objects. One concrete example is the standard, parallel, frontier-based breadth-first search algorithm [7] terminates if no new vertex is explored in an iteration, i.e., the frontier for the next iteration is empty. This can be written as a conjunction of boolean “Am I part of the frontier?” vertex values, which can be reduced to determine if the algorithm should terminate. In the N-Body example from this paper, execute and reduce can be used to calculate the average distance of all bodies from a given point in space (Figure 3).

The reduction part is not supported in device mode yet and performed on the host. Future versions of Ikra will perform parallel reductions with shared memory [18].

### 3 Implementation
Ikra-Cpp is based on four ideas: (a) Allocate a large storage buffer (char array) in which all data is stored (generated by IKRA\_HOST\_STORAGE). (b) Assign unique integer IDs to objects. (c) Reference objects with “fake pointers” that encode an object ID. (d) Override C++ operators to decode IDs, calculate addresses and access data in the storage buffer.

In this section, we explain those steps in more detail, using a more verbose, but less convoluted notation. The example class in Figure 4 is identical to the one in Figure 1b (without constructor), but with expanded preprocessor macros.

```cpp
float Body::move(float x, float y)
{
float dx = pos.x - x;
float dy = pos.y - y;
return std::sqrt(dx * dx + dy * dy);
}
```
3.1 Overview

SOA object pointers (i.e., the result of a new expression) do not necessarily point to allocated data but are used to encode object IDs ("fake pointers"; similar to tagged pointers where the tag is the entire pointer). All objects of a SOA class C have a unique ID\(^3\) between [1; maxInst(C)]; E.g., calling new C for the first time returns a C* pointer that encodes ID 1.

SOA field types behave like normal C++ types in most cases, but access data at a location inside the storage buffer. In particular, they must support the following operations.

- **Reading Value**: A double\(--\) value can be converted to a double value without an explicit typecast (implicit conversion operator\(^4\), Figure 5, Line 12).
- **Writing Value**: A double value can be assigned to a double\(--\) field (assignment operator, Line 13).
- **Method Call**: For non-primitive types (does not apply to double), a method call on a SOA field is forwarded to the object at the data location (member of pointer “arrow” operator\(^5\), Line 14).

SOA field types are defined in SoaLayout as template instantiations of Field\_ (Figure 5, Lines 3–6). This class provides the necessary operator implementations and calculates the address inside the storage buffer at which the field value of a certain object can be found (Line 15). In the most basic case, given the address of a field object this, the address of field C::f can be computed as follows, where id is a function that decodes the object ID from a field object address.

\[ \text{addr}(this, C::f) = \text{storage} + \text{maxInst}(C) \cdot \text{offset}(C::f) + (id(this) - 1) \cdot \text{sizeOf}(C::f) \]

The first two lines in the equation compute the beginning of the SOA array storing all values of C::f. The third line computes the offset into that array. How exactly object IDs are encoded in SOA object addresses is determined by the addressing mode. Ikra-Cpp supports three different addressing modes, one of which must be chosen at compile time: Zero Addressing and two variants of Valid Addressing. The former one is more space-efficient but relies on non-standard C++ constructs, so it might not work with some compilers\(^6\).

3.2 Addressing Modes

This section gives an overview of various addressing modes. Zero addressing and storage-relative zero addressing are implemented in Ikra-Cpp. In accordance with the C++ zero overhead principle [34], zero addressing is the default mode.

3.2.1 Zero Addressing

In this addressing mode (Figures 6, 7), an object of class C with ID i is referenced with a C* pointer pointing to memory address i (e.g., obj\textsubscript{10} has address 0xa). Field values are grouped by field and stored from the beginning of the storage buffer. No field values are stored for object 0 (null pointer). Given a C* pointer obj, the memory location of field value C::f, i.e., &obj->f, is calculated as follows. Compile-time constants are in blue.

\(^3\)ID 0 is reserved for null references.
\(^4\)The auto keyword is not supported. E.g., a field value cannot be assigned to a variable declared as auto without an explicit type cast.
\(^5\)Note for experienced C++ programmers: This is similar to how std::unique_ptr is implemented.
\(^6\)We verified that it works with gcc 5.4.0, clang 3.8.0 and CUDA 9.0.
The size of a class or struct should be greater than zero (even if empty) according to the C++ standard \cite{12}, but many compilers can be instructed to use a size of zero. If this is not supported by a compiler, either valid addressing or a different mechanism for instance creation must be used.

### 3.2.2 Valid Addressing Mode

Since zero addressing does not conform to the C++ standard, Ikra-Cpp provides a different addressing mode. In valid addressing, the C++ size of every SOA field is one byte (e.g., `sizeof(double) = 1`), consequently the C++ size of every SOA object is `numFields` bytes. In order to support the new keyword, the address of a SOA object must then point to valid (allocated) memory; thus the name valid addressing.

The challenge of valid addressing is to add an as small as possible amount of padding (wasted memory) such that no data is overwritten by zero initialization. Programmers should use zero addressing if supported by their compiler, since it does not waste any memory. Even though we do not have measurements for valid addressing yet, we expect the same runtime performance as in zero addressing, because address computation is in both cases reduced to strided memory access after constant folding.

#### Storage-relative Zero Addressing

In this addressing mode, an object of class `C` with ID `i` is referenced with a `C*` pointer pointing to the same address as in zero addressing and starts at offset `padding = maxInst + 1 + numFields(C)`, i.e., `padding` many bytes are wasted in this addressing mode. In general, the memory location of a field `C::f` of object with address `obj` is then calculated as follows. Note that the formula is identical to the one in zero addressing, except for the offset of the data segment and the ID computation.
\[ addrValid(obj, C::f) = \text{storage} \]

data segment offset \[ + \text{maxInst}(C) + 1 + \text{numFields}(C) \]
\[ + \text{maxInst}(C) \cdot \text{offset}(C::f) \]
\[ - \text{sizeof}(C::f) \]

ID computation \[ + (obj - \text{storage}) \cdot \text{sizeof}(C::f) \]

Since address computation is done inside SOA field classes (\texttt{Field\_}, not \texttt{SoaLayout}), we have to express the above formula in terms of the address (this pointer) of a SOA field instead of the object address \texttt{obj}. The address of a SOA object \texttt{obj} inside of field \texttt{C::f} is defined as \texttt{obj = this-\text{index}(C::f)+1}, where \texttt{index(C::f)} is the field index of \texttt{C::f}. E.g., the address of the third field \texttt{vel\_x} of \texttt{Body\_1} is \texttt{0x4003} in Figure 8 (striped box). Consequently, \texttt{obj = 0x4003 - 3 + 1 = 0x4001}. This object address can be used in the above formula. Putting both definitions together, the memory location of a field \texttt{C::f} with respect to its \texttt{this} pointer is then calculated as follows.

\[ addrValid(this, C::f) = \text{storage} \]
\[ + \text{maxInst}(C) + 1 + \text{numFields}(C) \]
\[ + \text{maxInst}(C) \cdot \text{offset}(C::f) \]
\[ - \text{sizeof}(C::f) \cdot (\text{index}(C::f) + \text{storage}) \]
\[ + \text{this} \cdot \text{sizeof}(C::f) \]

The formula above was rearranged to keep the number of terms small. After constant folding, the address of a field value can be calculated with the same instructions as in zero addressing mode.

4 Preliminary Performance Evaluation

We evaluated Ikra-Cpp1 on a computer with an Intel Core \texttt{i7-5960X} CPU (4x 3.00 GHz), 32 GB RAM and an Nvidia GeForce GTX 980 GPU, a 64-bit Ubuntu 16.04.1, gcc 5.4.0 and the Nvidia CUDA Toolkit 9.0.176 in zero addressing mode.

We benchmarked an iterative application of \texttt{Body::move} for all bodies (Figure 2a in a loop)\(^8\). The number of iterations was chosen such that every program ran for at least 5 sec. We calculated the average running time per iteration and report the minimum time out of 12 program runs.

Running Time Figures 9 and 10 show the running time on CPU and GPU. The upper subfigure shows the average running time of one entire iteration and the lower subfigure shows the average running time for a single body instance.

In host mode, Ikra-Cpp’s performance is almost identical to hand-written SOA code. AOS-32 is variant of AOS where 16 supplemental double fields were added to the Body class, similarly to the SoA\_x benchmark section [10]. We can think

\(^8\)This benchmark is quite simple, but it clearly isolates the overheads of Ikra-Cpp, specifically address computation.
of such fields as additional properties of a body (e.g., mass or radius) that are not utilized in this particular computation. The AOS-32 graph of the lower subfigure clearly shows the effect of the L1, L2, L3 caches (32KB, 256KB, 20MB).

The performance difference in device mode is due to a higher kernel invocation overhead in Ikra-Cpp. More than 10000 iterations (kernel invocations) are performed for small problem sizes. With a larger number of bodies, we get closer to hand-written SOA code, because of fewer iterations.

**Code Generation** We verified that the generated binary code for reading/writing a single field from a SOA object is identical in Ikra-Cpp and hand-written SOA (gcc and clang). This shows that modern compilers can constant-fold the complex address computations in Ikra-Cpp.

Unfortunately, this is not always the case for other compiler optimizations. One example is automatic loop vectorization. In hand-written SOA code, gcc and clang vectorize the loop that calls Body::move for every Body instance. However, only gcc performs the equivalent loop vectorization with Ikra-Cpp. Clang is able to apply optimizations like loop unrolling but considers the operations involved in address computation as potentially “dependent” memory operations and thus unsafe for vectorization.

There are three approaches to solve this problem. First, we can try rewriting the address computation part of Ikra-Cpp, in an attempt to give the compiler additional hints that trigger optimizations. This approach is fragile and could break at any time. Second, code can be vectorized manually, either with C++ SSE intrinsics or with a vectorization framework like Sierra [14, 15]. Considering that real applications, which exhibit code that is more complex than our example here, cannot be automatically vectorized (yet) with today’s compilers, even if written in SOA style, this approach seems feasible to us. Third, Ikra-Cpp could be implemented as a compiler extension, which is the cleanest and most stable solution. We describe this approach in the context of the *Intel ispc Compiler* in more detail in Section 5.

**5 Related Work**

The AOS-SOA tradeoff is a well-known problem and has been studied in previous work in the context of C structs. To the best of our knowledge, there is no system that provides an AOS-like programming style for object-oriented programming with an implicit SOA data layout.

Homann and Laenen developed SoAx [10], a C++ library for AOS-style C/C++ programming with implicit SOA layout. Their implementation is based on preprocessor macros and template metaprogramming. SoAx does not support OOP concepts like classes or methods. SOA struct types are defined using std::tuple instantiations and a helper macro

The problem is pointer casting. In the simplest case, an expression like array[reinterpret_cast<uintptr_t>(id)], where id is a pointer encoding an integer array offset is already considered unsafe.

Ikra-Cpp: A C++/CUDA DSL for OOP with SOA

---

10The problem is pointer casting. In the simplest case, an expression like array[reinterpret_cast<uintptr_t>(id)], where id is a pointer encoding an integer array offset is already considered unsafe.

11ispc does not support OOP features like classes and methods, yet (01/2018).
7 Summary
We presented a first implementation of Ikra-CPP, a C++/CUDA DSL for object-oriented programming. Ikra-CPP allows programmers to write object-oriented code in AOS notation, while data is layed out as SOA for better performance. SOA object members are always accessed through pointers. How exactly an object ID is encoded in a pointer is determined by the addressing mode. Our main insights are that (a) object ID decoding and field address computation can be done efficiently after constant folding and (b) an AOS-style notation can be achieved transparently in C++ with operator overloading, template metaprogramming, and preprocessor macros. Preliminary benchmarks show that simple examples written with Ikra-CPP and compiled with gcc are on par with hand-written SOA code.

References
A First Field Addressing

This addressing mode is a variant of valid addressing. Its purpose is to reduce the amount of waste due to the padding area. It is also required for virtual function support in the future. An object of class \( C \) with ID \( i \) is referenced with a \( C* \) pointer pointing to the memory location of the value of the first field of the object \( i \) (Figure 11). If the SOA class has at least one virtual function, the first field is the vtable pointer. If the number of fields of the SOA class is larger than the size of the first field, then the memory of the first field must be padded with \( \text{sizeof}(C::f) - \text{numFields}(C) \) bytes to avoid overwriting first field values of following objects due to zero initialization.\(^{12}\) Given a \( C* \) pointer \( obj \), the memory location of a field \( C::f \) is calculated as follows.

\[
addr_{first}(obj, C::f) = \text{storage} + \maxInst(C) \cdot \text{offset}^*(C::f) + \left(\frac{\text{obj} - \text{storage}}{\text{sizeof}^*(C::f)} - 1\right) \cdot \text{sizeof}^*(C::f)
\]

\( \text{sizeof}^* \) and \( \text{offset}^* \) take into account padding that might be added to the first field. The memory location of a field \( C::f \) with respect to its \( this \) pointer is calculated as follows.

\[
addr_{first}(this, C::f) = \text{storage} - \text{sizeof}(C::f) + \maxInst(C) \cdot \text{offset}^*(C::f) - (\text{index}(C::f) + \text{storage}) \cdot R + this \cdot R
\]

where \( R = \frac{\text{sizeof}^*(C::f)}{\text{sizeof}^*(C::first)} \)

Even though the definition of \( addr_{first} \) contains a fraction, its value is always an integer. However, its calculation is not straightforward. On the one hand, address calculation can be optimized by isolating the variable parts. In the formula above, the only variable part \( this \) is multiplied by a compile-time constant and added to a compile-time constant (strided memory access). On the other hand, \( R \) might not be an integer and neither might be the constant-folded parts of the strided memory access. Floating point operations as part of the address computation should be avoided by all means. \text{Ikra-Cpp} supports first field addressing only for SOA classes where \( R \) is an integer, i.e., the size of every field is a multiple of the size of the first field. Furthermore, this addressing mode is superior to storage-relative zero addressing only if the field padding size is zero or one byte. Note that field padding is incurred for every instance, i.e., \( \maxInst \) many times. First field addressing is optimal for the minimal N-Body example, where all fields have the same size and the number of fields equals the size of the first field.

B Additional Evaluation

We repeated the experiments from Section 4 on another machine with an Intel Core i7-6820HQ CPU (4x 2.70 GHz), 32 GB RAM and a GeForce 940MX GPU.