High-Performance GPU offloading using *XBLang*: an extensible language front-end for MLIR

***Abstract*—This paper presents the design and development of *XBLang*, a language front-end for MLIR, created to allow open access to Multi-Level Intermediate Representation (MLIR) infrastructure, focusing on GPU offloading.**

The *XBLang* compiler (`xbc`) contains an extensible high-level front-end design developed for creating and testing new language constructs and a middle-end based entirely on MLIR. XBLang targets multicore CPUs and GPU parallelism and successfully runs on targets like NVIDIA and AMD GPUs and CPUs like A64FX. Our results demonstrate speedups or comparable performance to vendor compilers for GPUs.

Highlighting one of our results from evaluation, we observe that NAS CG *XBLang*'s GPU version on NVIDIA A100 is $74\times$ faster than Clang-serial, $2.23\times$ faster than Clang's OpenMP offload, and $1.17\times$ faster than NVIDIA's OpenACC compilers; for AMD MI250x, *XBLang* is $36\times$ faster than Clang sequential, $5\times$ faster than Clang's OpenMP offload, and $6\times$ faster than AMD Clang's OpenMP offload.

***Index Terms*—parallel languages, parallel programming, compiler infrastructure**

I. Introduction

High-Performance Computer (HPC) has seen a paradigm shift from homogeneous to heterogeneous systems in the last couple of decades. The heterogeneous system era has seen the introduction of massively parallel architectures such as GPGPUs from vendors such as NVIDIA, AMD, and Intel, vector architectures such as the A64FX from Fujitsu/ARM, and also other types of devices such as FPGAs, ASICs, neural engines, and many-core processors. This shift means plenty of parallelism has to be exposed and expressed at the different hardware and software stack levels.

These new parallel architectures are evolving rapidly and constantly disrupting software development. This disruption leads to significant challenges when adapting legacy algorithms designed for conventional architectures to new ones. These legacy algorithms often require extensive rewrites to take advantage of the advanced features found in new architectures, e.g., adapting the classical GeMM algorithm to use tensor operations in GPU architectures. Therefore, developers are often tasked with creating new parallel algorithms and abstractions designed for these new architectures and adapting these changes to evolving programming models. This nontrivial process requires significant effort.

Unlike CPUs and traditional programming languages like C/C++/Fortran, there is no unified and accessible way to program these new devices, as different hardware vendors use different programming models for their accelerators, exacerbating the software disruption. Parallel programming models like OpenMP [1], OpenACC [2], and Alpaka [3] have been developed over the years to address portability issues. However, these models do not provide access to all hardware features that native languages like CUDA and HIP do, resulting in a gap that can be detrimental to end users.

Compiler developers are also affected by this architecture evolution, facing a never-ending race to implement host language compiler features and a myriad of programming models. For example, while the OpenMP specification continues to grow, they are still developing offloading implementations; see SOLLVE V&V [4] for reference on the status of some of these implementations. An OpenMP 6.0 release is scheduled for 2024, while most compiler developers are still implementing features from OpenMP 5.2, ratified in 2021. To that end, most OpenMP offloading stories [5]–[12] almost only use OpenMP 4.5 features ratified in 2015, elucidating a gap between features in the standard and their usage.

C++ is a prominent language in HPC [13], but its use can be challenging to many due to its complicated syntax and semantics. As a result, Rust [14], Python [15], and, lately, Julia [16] are becoming more popular languages of interest for HPC; however, most new specialized hardware features are still only present in C++.

Given its prevalence, C++ is a popular language for conducting HPC research. However, the constant addition of new features to C++, lengthy ISO specification, and C++'s syntax and semantics have made it difficult for researchers to maintain or develop parallel tools targeting C++. For example, the CETUS [17] and derived projects [18] do not support C++, only C due to C++'s complexity.

With the rise of new AI and quantum architectures in the coming years, further challenges are expected in developing suitable and compelling software for these up-and-coming architectures. Our work addresses the evolving hardware/software landscape by exploring a high-level front-end language in Multi-Level Intermediate Representation (MLIR), a flexible infrastructure for modern optimizing compilers.

II. Motivation

Compiler practitioners often create new HPC programming abstractions for various reasons, such as accommodating new programming paradigms, introducing optimizations, addressing known gaps in existing models, and coping with the evolution of hardware architectures. To add these new abstractions, practitioners often create source-to-source tools, extend

Identify applicable funding agency here. If none, delete this.

existing languages and compilers, or create new programming languages.

Source-to-source (S2S) tools are classic yet popular options for conducting research; see CETUS [17] and derived projects [18], [19], the ROSE compiler [20], and other S2S tools [21]–[26]. However, as the survey article [27] presents, these tools have many perceived shortcomings and detractors. In this survey, the authors seek to answer why and in what contexts HPC practitioners avoided source-to-source transformation tools. Their approach consisted of a survey of papers that intended to use a source-to-source tool but argued against using them. According to the survey, practitioners avoided using these tools because they are difficult to extend to support new programming models, lead to complex and fragile workflows, and may interfere with compiler optimizations. A critical shortcoming not mentioned in the survey is that S2S tools are always limited to being within the compounds of the target language.

Clang-LLVM [28] offers an environment to create new HPC programming model abstractions within C++. However, despite offering several options, such as modifying Clang's source, source-to-source Libtooling tools and LLVM passes, tools developed within the environment also have drawbacks and limitations.

LLVM passes can easily modify the IR produced by Clang, thus providing a robust workflow for introducing optimizations and support for new hardware architectures. However, LLVM-IR has no high-level information about the language and its semantics, limiting the type of possible manipulations.

Modifying Clang's front-end source code allows the introduction of new language constructs; however, this modification process could be more convenient and less error-prone for newcomers. Compiling these source code modifications might take significant time, depending on the system, reducing productivity. Furthermore, language extension modifications on Clang would require patching Clang's source and, therefore, are not easy to combine with other language extensions or with the permanent Clang development, as it requires acceptance from the Clang community.

An MLIR front-end would overcome some of the shortcomings of LLVM passes, thanks to MLIR's capability to represent high-level information; see Section IV for information about MLIR. Two notable projects leveraging MLIR are ClangIR [29], a high-level IR for Clang, and LLVM-Flang [30], a Fortran compiler. However, both are still under active development, with ClangIR in early development and Flang not in production yet. Nonetheless, the struggle to introduce new constructs into these existing front-ends persists, as they are not general MLIR front-ends.

The challenges outlined above inspired the creation of ***XBLang*** - a language front-end for MLIR and compiler designed for extensibility, capable of addressing architecture evolution thanks to the power of the MLIR infrastructure. In this paper, we make the following contributions:

- The ***XBLang*** language, an extensible front-end for MLIR and programming language conceived to target the evolving hardware landscape, constantly demanding newer software techniques for maximizing performance.
- The ***XBLang*** compiler (`xbc`), an MLIR-based extensible compiler. This compiler introduces a new MLIR dialect to express the semantics of a high-level programming language capable of interacting with existing MLIR dialects, thus leveraging existing MLIR features and capabilities.
- The **`par`** MLIR dialect and its corresponding ***XBLang*** front-end extension, which is capable of targeting CPUs and GPUs and it was **created to demonstrate the extensible capabilities of the compiler**. **`par`**'s performance matches or surpasses vendor performance compilers, outperforming other open-source compilers and programming models for our test cases. We envision our model to evolve to meet the demands of the rich hardware features that expose multiple levels of parallelism.

III. Language

This section presents ***XBLang*** and its extensibility properties. Furthermore, we present **`par`**, a language extension for writing portable parallel programs capable of yielding high performance.

The syntax of ***XBLang*** is similar to that of Rust and C, with blocks of statements delimited by curly braces, expressions delimited by semicolons, and the usual control flow statements. Declarations such as structs, functions, and variable definitions follow Rust's syntax style. The semantics of ***XBLang*** are those of C, plus some extensions such as type inference and a module system instead of C includes.

In ***XBLang***, we use the term "Language Context" to describe a set of semantic and syntactic rules that govern how a particular language construct works. Each construct of ***XBLang*** has a language context, which can have multiple parent and child contexts, with child contexts following the rules set by their parent contexts. For example, typed and named declarations govern function declarations, forcing functions to have a well-defined type and a name.

XBLang uses language dialects to extend its functionality. Each dialect encapsulates a specific set of language constructs and provides the rules for creating the language context for each construct. Dialects can introduce any elements necessary for their semantic and syntactic purposes and determine what is legal inside a context owned by the dialect.

The only overall restriction imposed on dialects is that any side effects produced by a dialect construct must always have a valid semantic meaning that the parent context can understand. Failing this restriction is considered an error, e.g., inside an ***XBLang*** statement, all child contexts must follow statement semantics, i.e., at the top level, they are ordered and structured.

Syntactically, language contexts can be created by specifying the dialect's keyword or by invoking the desired dialect construct directly, provided the dialect registers the construct within the parent context and there are no syntactic conflicts. Listing 1 shows this mechanism; in this listing, the **`par`** keyword establishes the intent of creating a **`par`** construct, like `map` or `region`. In contrast, the `loop` construct does not

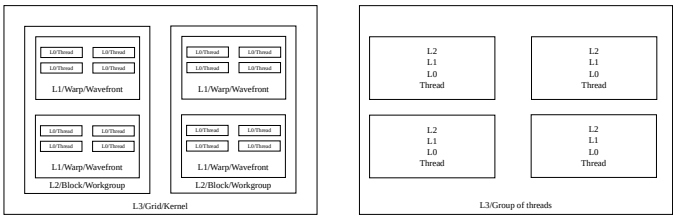

Figure 1: Mapping of the thread hierarchy system in **par** to hardware. The mapping for GPU offloading is on the left, while the diagram on the right shows the mapping for CPU architectures.

need to be surrounded by the **par** keyword, as it is registered globally in ***XBLang***.

```
fn saxpy_xb(x: f64*, y: f64*, a: f64, n: i32) {
  par map(toFrom: x[0 : n]) map(to: y[0 : n])
  par region firstprivate(a, n)
  loop(let i: i32 in 0 : n)
    x[i] = a * y[i] + x[i];
}
```

Listing 1: ***XBLang*** high level version of the SAXPY kernel.

A. *par Programming Model*

We designed the ***XBLang*** **par** dialect to be similar to commonly used programming models, such as OpenMP and OpenACC, so new developers could quickly learn and use it.

However, unlike directive-based programming models such as OpenMP and OpenACC created to convert a sequential program into a parallel program, the **par** dialect adopts the philosophy that users are responsible for constraining parallelism within the model, not the model constraining users.

The complete set of currently available clauses in the **par** dialect is presented in Listing 2. Their definition is close to the definitions of similarly named OpenMP clauses. In many cases, code written using the clauses in Listing 2 will result in portable code across CPUs and GPUs, demonstrated in Section VI.

```
map[[queue]]?([[clause [, clause...]]])
region[[queue]]?[[clause [, clause...]]]
loop[[clause [, clause...]]]
atomic(op: lvalue, rvalue)
reduce(op: [lvalue, [, lvalue...]])
sync
lead
wait
```

Listing 2: A complete list of clauses available in **par**.

The execution model of the **par** dialect is organized into parallelism levels, namely L0, L1, L2, and L3. Higher levels have more parallel resources, while lower levels have less or no parallel resources. Figure 1 shows the mapping of this hierarchy to hardware; it also demonstrates that only the L3 level provides parallelism in the CPU model, with L1 and L2 levels considered *degenerate* providing no parallelism.

The model also defines that constructs executing in *degenerate levels* execute sequentially and always have `id` 0 and `dim` 1. This definition is the key to allowing portability across CPU and GPU and enabling the creation of sequential code.

The `region` construct is the primary method for exposing parallelism in the dialect. These regions always launch an L3 level of threads, with all threads executing the same region. In a CPU, this mechanism is similar to calling `pthread_create`, while for accelerators such as GPUs, opening a parallel region would be identical to calling a kernel in CUDA or HIP. Within parallel regions, there are no distinctions between threads, and the model assumes that potentially all threads could be working simultaneously, meaning all threads are active and ready to distribute work to specific classes of threads.

In **par**, work can be distributed across all parallel hierarchy levels using the `loop` construct. For instance, distributing a `loop` at the top level is equivalent to distributing a `loop` at the grid level, using CUDA terminology —loops default to running on all available levels if no nested loops or other clauses exist. If there is nested parallelism, then the compiler decides how to map the loops onto the available levels; this equates to the outer `loop` scheduled at the top level and the nested `loop` expanded into lower levels. The dialect also allows for specifying how to distribute the work across levels; this is shown in Listing 3, where the top `loop` runs at the L3 (`<L3>`) level. In contrast, the nested `loop` expands over L2, L1 and L0 (`<L2:>`).

```
fn smpv(q: f64*, A: f64*, y: f64*, rowPtr: i32*,
    colIndx: i32*, n: i32) {
 par region firstprivate(n)
 loop<L3>(let i: i32 in 0 : n) {
  let sum: f64 = 0.;
  loop<L2:>[reduce(sum)](let col: i32 in rowPtr[i]
       : rowPtr[i + 1])
    sum += A[colIndx[col]] * y[col];
  lead q[i] = sum;
 }
}
```

Listing 3: Work distribution for an SPMV kernel in ***XBLang***.

In Listings 1, 4 and 5, we compare programming models and languages using the SAXPY kernel as reference code. Listings 1 & 4 represent two possible implementations within **par** for the kernel. The first one represents a high-level version similar to the OpenMP version in listing 5, with the `loop` construct distributing the work automatically into all available parallelism levels. The second version expresses the same idea but uses low-level constructs like `id` and `dim`; this version is akin to a typical CUDA kernel. Both listings show the memory mapping mechanism in **par**; internally, a custom memory manager calling CUDA or HIP handles this mapping.

```
fn saxpy_xb(x: f64*, y: f64*, a: f64, n: i32) {
  par map(toFrom: x[0 : n]) map(to: y[0 : n])
  par region firstprivate(a, n) {
    let i: i32 = id<L2:> + id<L2> * dim<L2:>;
    if (i < n)
      x[i] = a * y[i] + x[i];
  }
}
```

Listing 4: ***XBLang*** low level version of the SAXPY kernel.

```
1 void saxpy_omp(double *x, double *y, double a, int
      n) {
2 #pragma omp target loop map(tofrom: x[0 : n]) \
3                          map(to: y[0 : n])
4   for (int i = 0; i < n; ++i)
5     x[i] = a * y[i] + x[i];
6 }
```

Listing 5: OpenMP version of the SAXPY kernel.

IV. Multi-Level Intermediate Representation (MLIR)

This section presents a general overview of MLIR, an extensible compiler infrastructure [31]. MLIR provides the building blocks for creating compilers; however, it is important to stress that MLIR is *not* a compiler [32]. We leverage these components in Section V-B to construct the `xbc` compiler's middle-end, using the MLIR infrastructure to create intermediate representations for the ***XBLang*** language.

A. Overview

The Multi-Level IR (MLIR) project [33] is an extensible compiler infrastructure capable of representing arbitrary *graph-like* IRs, with *graph-like* meaning that IR operations and IR values form a graph structure. MLIR employs a hierarchical structure to represent IRs, allowing for extensibility and modularity. The key IR concepts in MLIR are:

- *Regions* are ordered list of *Blocks*, allowing modularity, e.g., the body of a function can be represented by a *Region*.
- A *Block* is an ordered list of operations, and in SSACFG *Regions*, they represent compiler basic blocks in an SSA style. One crucial distinction to traditional basic blocks is that in MLIR *Blocks* can have arguments, allowing to perform control flow between *Blocks* without the nuances of *Phi* nodes [34].
- A *Value* is a unique typed result produced by an *Operation*, being the edges communicating between operations.

Additionally, MLIR defines an open type system used by *Values*, with the semantics of the types being "application-specific" [34]. MLIR also has the concept of *Attributes* and *Properties* for storing additional information in operations and types. Some examples are integer constants or symbol names. Unlike other intermediate representations like LLVM IR, MLIR allows for symbols, providing a non-SSA procedure to refer to operations, e.g., a function or a global variable.

To allow for extensibility and modularity, MLIR introduces the concept of dialects where *Operations*, *Types*, *Attributes*, and other concepts with a common purpose are grouped to provide semantics and a programming interface. From the implementation standpoint, the creation of dialects also simplifies parsing and printing the IR as it encapsulates the information on how to perform the actions.

B. MLIR dialects

MLIR provides a set of core dialects as a starting point for interacting with the infrastructure. These dialects are maintained and developed by the MLIR community; all are under active development and have ample infrastructure built around them, e.g., optimization passes and IR transformations. The `builtin` dialect defines basic infrastructure, like integer and floating point types, and higher-level types, like `memref`, intended to describe general memory references, like ranked or unranked tensors in memory. Other relevant dialects are: `arith` & `math` for arithmetic and math operations, `func` for defining and interacting with functions, `cf` & `scf` for basic and structured control flow, and `affine` for affine operations like loops.

V. The `xbc` Compiler

This section explores each of the stages of our compiler alongside relevant implementation details. Figure 2 presents a high-level representation of the compilation workflow. This figure shows the compilation process from an input program to LLVM IR, with Clang compiling the IR into an executable. We chose to leave the executable generation to Clang as it natively handles LLVM IR, can call vendor tools for device code compilation, and can handle more complicated compilation tasks such as Link Time Optimization (LTO).

In the following subsections, we explore the main contributions of the paper: the front and middle ends of the compiler. Section V-A elucidates how to introduce new language constructs into the `xbc` front-end, including introducing new syntax effortlessly and generating the AST infrastructure. Then, in Section V-B, we introduce the **`xb`** and **`par`** MLIR dialects, two high-level representations for the language constructs available in XBLang and PAR. Section V-B also presents how these two high-level MLIR dialects are transformed and lowered into LLVM IR.

A. `xbc` front-end

We use LLVM TableGen [35] to specify almost every aspect of XBlang language dialects front-end, with custom TableGen backends generating the necessary code and hooks to interact with the rest of the compiler. This design decision allows the use of all the existing machinery inside the TableGen language, further facilitating the extension of the compiler.

To make the compiler dependent only on the LLVM and MLIR projects and as self-contained as possible, we opted against using existing Lex & Parser generators, such as Flex [36], Bison [37], or ANTLR [38], for specifying the grammar. Instead, we use self-made generators capable of producing classical DFA lexers and Packrat [39] parsers. Any dialect can use the lexer generator to generate a complete DFA or specific routines to lex tokens, simplifying the introduction of intricate tokens, as shown in Listing 6.

Dialect grammars are specified in a custom Parsing Expression Grammar (PEG) format, allowing the specification of complex grammars with little effort. This format permits

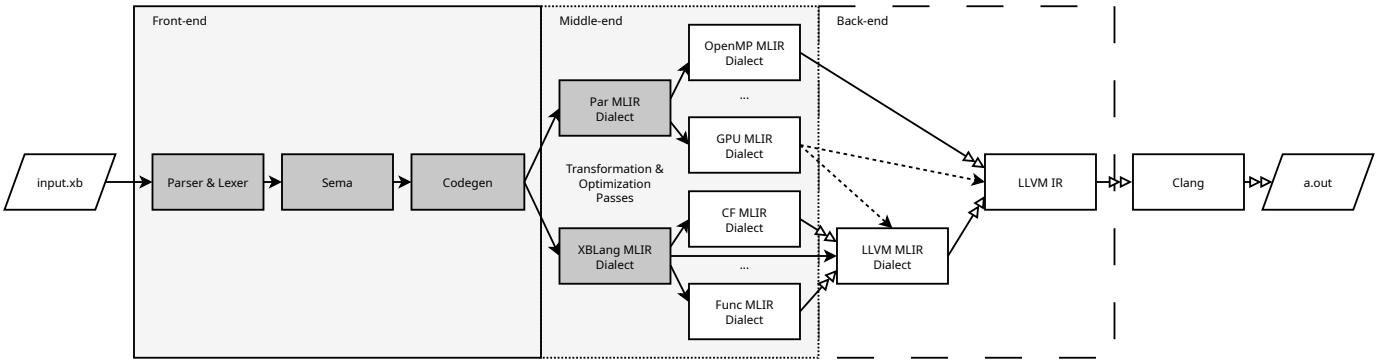

Figure 2: Compilation workflow of the ***XBLang*** compiler. Rectangles colored in shades of gray and black arrowheads represent the contributions of this paper. Dashed arrows indicate contributions to the MLIR community resulting from this work. Double arrowheads indicate parts of the workflow that we did not change for this work. Please refer to Section IV-B for an MLIR overview.

```
class FloatLit<string suffix> {
 list<TokenRule> toks = [
  TokenRule<"FloatLiteral", [], [{(
    (digit_sequence [eE] [+\-]? digit_sequence)
  | (digit_sequence '.' ([eE] [+\-]?
      digit_sequence)?)
  | (digit_sequence? '.' digit+ ([eE] [+\-]?
      digit_sequence)?))}]>
  ];
  list<TokenRule> tokens = !foreach(tok, toks,
    TokenRule<tok.identifier # suffix, [],
      tok.rule # ruleSuffix>);
}
def f32 : FloatLit<"f32">; // 0.f32, .32f32
def f64 : FloatLit<"f64">; // 1.e-10f64, 1.32f64
```

Listing 6: Tablegen class for recognizing float literals with a suffix. A `TokenRule` specifies the creation of a lexing rule associated with a named token.

the introduction of C++ code actions and includes macro expressions. Parse macros enable describing a common concept once and employing it multiple times later, diminishing code bloating, e.g., the `SepList(expr, sep)` macro expands into a list of zero or more `expr` separated by `sep`.

Dialects are also free to provide their lexer, parser, or specific lexing or parsing routines as long as they comply with the compiler interface. For example, the ***XBLang*** dialect provides a custom method for parsing binary expressions, using operator precedence parsing to speed up the parsing of expressions.

Dialects specify AST nodes through TableGen, providing an easy mechanism for introducing new constructs. Listing 7 presents a simplified version of such a node specification for the **`par`** region construct; in this listing, the `RegionStmt` belongs to the **`par`** dialect; it inherits the properties found in the `Stmt` node, defines a symbol table and has multiple children, like a statement body. The syntax for the construct is specified through the format field, allowing a unified front-end specification; for parsing a `RegionStmt`, the construct has an optional `LaunchStmt` and a required body statement.

```
def ParRegionStmt: ParNode<"RegionStmt", Stmt, [
    SymbolTable]> {
 let leaves = (ins "LaunchStmt":$launchParameters,
                   ..., "Stmt":$body);
 let format = [{ 'region' EE
  (LaunchStmt:stmt { result.setLaunchParameters(
      stmt); })? ...
  Stmt:body { result.setBody(body); }
 }];
}
```

Listing 7: Tablegen record for the `par region` construct, showing its children and syntax. Blocks enclosed in curly braces inside the `format` field denote C++ code actions; the remainder of the field represents syntax.

The structure of the Abstract Syntax Tree (AST) used by the front-end mirrors, to some extent, that of Clang's, grouping classes of related semantic objects and forming a class hierarchy, providing a helpful interface for interacting with it. This hierarchy uses type safeness as the first semantic checker mechanism, as inserting a node of type `Decl` into a node that expects a `Stmt` for a particular field is impossible.

The *Sema* stage will perform semantic checks on the AST, name resolution, and type inference for constructs like expressions. It works by traversing the AST recursively and verifying each of the nodes in the AST. In the case of a semantic error, *Sema* aborts further checks and compilation. Finally, the code generation stage handles the creation of the initial **`xb`** & **`par`** MLIR IR from the AST representation; for more details see Section V-B. Dialects can easily accommodate new constructs by providing the necessary semantic or code-generation hooks, currently specified using C++.

In this subsection, we demonstrated how to extend the front-end to accommodate new language constructs. In the following subsection, we present the **`xb`** and **`par`** MLIR dialects and all the major transformation stages happening in the middle end, starting from a very high-level IR produced by the front-end

to LLVM IR.

B. `xbc` MLIR middle-end

We built the middle-end of the `xbc` compiler using the MLIR compiler infrastructure; see Figure 2 for a general overview of the workflow and contributions. The middle-end relies on the **xb** and **par** MLIR dialects; these are the direct contributions of this work. It is organized into 6 compilation stages, discussed later in this section.

The **xb** MLIR dialect is one of the main contributions of this work. This dialect provides operations required to represent **xb**'s source code as a high-level IR, particularly suitable for applying high-level transformations. Many of the operations of this dialect represent generalized versions of operations in trunk MLIR; this is particularly true for the control flow operations in **xb**. For example, there are no operations modeling loops with **break** conditions in trunk MLIR; however, they are available in **xb**. Another great example is the **xb.**`bop` operation, as it allows the model of general binary operations, even between operands of different types.

We also created compatibility layers so the **xb** dialect is interoperable with trunk MLIR dialects. For example, the cast operation allows converting **xb** values with pointer or reference types to MLIR `memref`s. All trunk control flow operations can operate inside **xb** control flow operations. However, the opposite is not true; **xb** operations containing return operations might not work inside `scf` or `affine` operations.

The only passes and patterns from trunk MLIR used by the middle-end stages are the *canonicalize* and *cse* passes and patterns used for conversion, like *convert-gpu-to-nvvm* or *convert-cf-to-llvm*.

Another critical contribution of this work is the **par** MLIR dialect, capable of representing both portable and non-portable parallel programs generated by the front-end. Depending on the compilation target, the middle-end converts operations from this dialect to operations in either **xb**, `omp`, or `gpu` dialects.

To further explain the compilation stages used by the middle-end, the code appearing in Listing 8 will be used as a reference. This code represents a simple parallel vector fill function, with subsequent listing showing the parallelization for GPUs; an ellipsis in a listing indicates elided code.

```
1 fn fill(x: f64*, v: f64, n: i32) {
2  par map(toFrom: x[0 : n])
3  par region([1], [1]) firstprivate(v, n)
4  loop (let i: i32 in 0 : n)
5   x[i] = v;
6 }
```

Listing 8: Parallel fill vector function written in ***XBLang***, MLIR representations of this code appear in listings 9 & 10.

Listing 9 shows the MLIR generated by the *CodeGen* stage. This first representation uses mostly operations from the **xb** and **par** dialects. This listing shows that the **xb** MLIR dialect provides a close representation of the source code in Listing 8, with easily identifiable variable declarations (**xb.**`var`) as well as parallel regions (**par.**`region`) and loops (**par.**`loop`).

```
1  xb.func @fill(%arg0: !xb.ptr<f64>, %arg1: f64, %
       arg2: si32) {
2   %0 = xb.constant(1 : si32) : si32
3   %1 = xb.constant(0 : si32) : si32
4   %2 = par.default_queue !xb.address
5   %x = xb.var[param] @x : !xb.ptr<f64> = %arg0
6   %v = xb.var[param] @v : f64 = %arg1
7   %n = xb.var[param] @n : si32 = %arg2
8   %3 = xb.range %1 "<" %n !xb.ref<si32> : !xb.
        range_t<si32>
9   %4 = xb.array_view %x !xb.ref<!xb.ptr<f64>> [%3 !
       xb.range_t<si32>] : tensor<?xf64>
10  par.data_region map(toFrom: %4 : tensor<?xf64> ->
        %x : !xb.ref<!xb.ptr<f64>> [%2:!xb.address])
        {
11   par.region firstprivate(%v : !xb.ref<f64>, %n :
        !xb.ref<si32>) {
12    %5 = xb.cast %n : !xb.ref<si32> -> si32
13    par.loop (%i : si32 in %1 : %0 : %5) {
14     %6 = xb.array %x !xb.ref<!xb.ptr<f64>> [%i :
          si32] : !xb.ref<f64>
15     %7 = xb.bop "=" %6 !xb.ref<f64>, %v !xb.ref<
          f64> : !xb.ref<f64>
16    }
17   }
18  }
19  xb.return
20 }
```

Listing 9: MLIR code generated by the *CodeGen* stage for the code in Listing 8.

We now present the middle-end stages. Each stage comprises multiple passes grouped to accomplish a goal; for example, the lowering stages transform higher-level constructs into lower-level ones. We uploaded the source code for the XSBench kernel used for evaluation in Section VI to Zenodo [40]. We also included some of the outputs produced by the compiler for XSBench to showcase some of its stages for a real-world code.

1) High-level transformations: The output from *CodeGen* is the input for the first middle-end stage, known internally as high-level transformations. In this stage, the **par** dialect applies its first set of transformations, explicitly placing memory mappings in the IR, transforming L3 reduction clauses into an L2 reduction and an atomic operation, and privatizing variables, amongst others.

2) Concretization: The concretization stage is one of the main stages in the pipeline, as it concretizes the IR generated by the high-level stage into an IR suitable for lowering to core MLIR dialects. Listing 10 shows the output of this stage. One of the fundamental roles of this stage is the introduction of implicit casting operations, like loading memory - see line 14 in the listing. Another critical transformation performed at this stage is type promotion, ensuring that binary operations have the same operand type after this stage.

In this stage, the **par** dialect will make explicit any runtime calls, collapse nested `loops`, and distribute `loops` across

the parallel hierarchy. During this stage, the compiler will replace mapped memory values inside parallel regions with the appropriate value.

```
1  xb.func @fill(%arg0: !xb.ptr<f64>, %arg1: f64, %
       arg2: si32) {
2   ...
3   %x = xb.var[param] @x : !xb.ptr<f64> = %arg0
4   ...
5   %10 = xb.call @__xblangMapData(%1, %6, %0, %9,
       %4) : (ui32, !xb.address, index, index, !xb.
       address) -> !xb.address<#gpu.address_space<
       global>>
6   %11 = xb.cast %10 : !xb.address<#gpu.
       address_space<global>> -> !xb.ptr<f64, #gpu.
       address_space<global>>
7   %12 = xb.cast %v : !xb.ref<f64> -> f64
8   %13 = xb.cast %n : !xb.ref<si32> -> si32
9   par.region {
10   %x_0 = xb.var[local] @x : !xb.ptr<f64, #gpu.
        address_space<global>> = %11
11   %v_1 = xb.var[local] @v : f64 = %12
12   %n_2 = xb.var[local] @n : si32 = %13
13   xb.scope {
14    %20 = xb.cast %n_2 : !xb.ref<si32> -> si32
15    %21 = par.id l3_l0 0
16    %22 = par.dim l3_l0 0
17    %23 = xb.bop "+" %21 si32, %3 si32 : si32
18    %i = xb.var[local] @i : si32
19    %24 = xb.bop "=" %i !xb.ref<si32>, %23 si32 : !
         xb.ref<si32>
20    xb.loop condition: {
21     %25 = xb.cast %i : !xb.ref<si32> -> si32
22     %26 = xb.bop "<" %25 si32, %20 si32 : i1
23     xb.yield Fallthrough %26 i1
24    } body : {
25     %25 = xb.cast %x_0 : !xb.ref<!xb.ptr<f64, #gpu
          .address_space<global>>> -> !xb.ptr<f64, #
          gpu.address_space<global>>
26     ...
27     %28 = xb.array %25 !xb.ptr<f64, #gpu.
          address_space<global>> [%27 index] : !xb.
          ref<f64, #gpu.address_space<global>>
28     %29 = xb.cast %v_1 : !xb.ref<f64> -> f64
29     %30 = xb.bop "=" %28 !xb.ref<f64, #gpu.
          address_space<global>>, %29 f64 : !xb.ref<
          f64, #gpu.address_space<global>>
30    } iteration : {
31     %25 = xb.cast %i : !xb.ref<si32> -> si32
32     %26 = xb.bop "+" %25 si32, %22 si32 : si32
33     %27 = xb.bop "=" %i !xb.ref<si32>, %26 si32 :
          !xb.ref<si32>
34    }
35   }
36  }
37  ...
38  %19 = xb.call @__xblangMapData(%2, %15, %0, %18,
       %4) : (ui32, !xb.address, index, index, !xb.
       address) -> !xb.address<#gpu.address_space<
       global>>
39 }
```

Listing 10: IR representation obtained after applying the concretization stage.

For the *XSBench* code uploaded to Zenodo [40], we see that going from the code generation stage to the concretization stage results in an increment of 1.76× lines of MLIR, showcasing how much information gets added during this stage.

3) High lowering: This stage performs the first lowering pass of the compiler, converting many of **xb**'s operations into operations in standard MLIR dialects, such as **xb.**`var` to `memref.alloca`; all these lowering transformations are direct contributions of this work. Listing 11 shows a partial output of this stage. A critical thing to observe from the listing is that not all operations get lowered at this stage; for example, casting, scope, loop operations, and types from the **xb** dialect remain.

There are multiple reasons for performing a partial instead of a complete lowering; the first reason is that not all operations in **xb** have an adequate equivalent operation in trunk high-level dialects, like **xb.**`loop`. The second reason is that the `memref` dialect models well-defined references to memory regions, not raw pointers. Thus, we overcome this restriction by providing casts between **xb** pointers and `memref`s (see line 15 in the listing) and model raw pointers inside the `memref` dialect as nearly infinite `memref`s (see line 16).

At this stage, the **par** dialect lowers into the `gpu` for GPU parallelism or the `omp` dialect for CPU parallelism.

```
1  func.func @fill(%arg0: !xb.ptr<f64>, %arg1: f64, %
       arg2: i32) {
2   ...
3   %alloca = memref.alloca() : memref<!xb.ptr<f64>>
4   memref.store %arg0, %alloca[] : memref<!xb.ptr<
       f64>>
5   ...
6   %11 = gpu.launch async blocks ... threads ... {
7    ...
8    xb.scope {
9     %17 = memref.load %alloca_4[] : memref<i32>
10    %18 = gpu.global_id  x
11    %19 = index.casts %18 : index to i32
12    %20 = gpu.grid_dim  x
13    %21 = gpu.block_dim  x
14    %22 = index.mul %20, %21
15    %23 = index.casts %22 : index to i32
16    %alloca_5 = memref.alloca() : memref<i32>
17    memref.store %19, %alloca_5[] : memref<i32>
18    xb.loop condition : {
19     %24 = memref.load %alloca_5[] : memref<i32>
20     %25 = arith.cmpi slt, %24, %17 : i32
21     xb.yield Fallthrough %25 i1
22    } body ...
23   }
24   gpu.terminator
25  }
26  ...
27  %16 = func.call @__xblangMapData(...)
28  xb.return
29 }
```

Listing 11: Lowered version of the IR from the Listing 10.

4) Low transforms: Low transforms involve transformations in lower-level dialects and translating the `gpu` dialect into LLVM IR. One example of such transformation is transforming `gpu.allreduce` operations into a combination of shuffle operations and `gpu.barrier` operations, MLIR already has an existing transformation for this operation. However, it does not apply to AMDGPU targets, and it is

computationally expensive as it has little information on the semantic nature of the reduction; for example, a reduction clause attached to the `loop` might not need broadcast its result to the neighboring threads.

5) LLVM lowering: This stage is the final MLIR stage and the second lowering stage. It is in this stage that all **`xb`** operations are fully lowered into the `llvm` MLIR dialect, making explicit all the casts, flattening scopes, and fully expanding **`xb.`**`loop`s into explicit control flow operations.

6) LLVM Translation: After the code gets lowered to the `llvm` MLIR dialect, the final stage is the invocation of the translation infrastructure in MLIR, translating the `llvm` MLIR dialect into LLVM IR, with LLVM IR being the final output of the compiler.

VI. Results

To evaluate the efficacy of ***XBLang***, we have chosen applications representing some of Berkeley's seven dwarfs [41]. These seven dwarfs represent common HPC application patterns. As part of the selection criteria, we choose applications that utilize OpenMP offloading capabilities, can run on GPU-based systems, and are maintained regularly. The benchmarks used for our evaluation are CG, representing Sparse Linear Algebra; Miniweather, representing Structured Grids; EP, XSBench, and RSBench, representing Monte Carlo simulations. These applications are part of state-of-the-art benchmark suites like SPEC-ACCEL [42], SPEC-HPC [43], and ECP proxy applications [44]. Furthermore, we noticed that these applications are repeatedly used in research papers for different evaluation types.

The raw files of the results presented have been uploaded in Zenodo [40]. The specific test cases used to demonstrate the efficacy of ***XBLang*** are the NAS parallel benchmark codes EP (Embarrassingly Parallel) & CG (Conjugate Gradient), both using Class=C and three mini-apps: MiniWeather(200x100 grid size, 5000s being the length of the simulation), RSBench (large simulation) and XSBench (large simulation). We use Perlmutter at LBNL and Frontier at ORNL for evaluation purposes. All tests were run ten times and averaged; time measurements were made using C++ 'high_resolution_clock' timers. In the case of GPU runs, we also included device-wide synchronization calls before making the measurement.

We use a single GPU for all our runs, NVIDIA A100 and AMD MI250x on Perlmutter and Frontier. For all benchmarks and platforms, we verified the output correctness of the programs generated by ***XBLang*** using the built-in validation mechanism for validating their output. All benchmarks were compiled using only *-O3* optimization flags.

Fig. 3 shows results using Perlmutter. The compilers used on Perlmutter include Clang 18.0.0 (8823e961), GCC OpenMP offloading 12.1.1, Clang OpenMP offloading 18.0.0 (8823e961), nvc OpenMP offloading 23.5.0, Cray OpenMP offloading 15.0.1, nvc OpenACC 23.5.0 and our ***XBLang*** compiler. We observe that ***XBLang*** performs the best for EP and CG even compared to vendor compilers; for XSBench ***XBLang*** performs close to the other compilers targeting GPU with *clang-omp* performing the best; for mini-weather ***XBLang*** performs almost close to the rest besides *clang-omp* that performs the worst; for RSBench & XSBench ***XBLang*** shows there is room for improvement.

Fig. 4 presents results using Frontier. The compilers used on Frontier include Clang 18.0.0, (e816c89c) Clang OpenMP offloading 18.0.0 (e816c89c), ROCM's 5.7.0 AMD Clang, Cray OpenMP offloading 16.0.1 and our ***XBLang*** compiler. We observe that ***XBLang*** performs better than all other compilers in 3 benchmarks: EP, CG, and Mini-Weather; for RSBench ***XBLang*** performs close to the best-performing compiler AMD-Clang; for XSBench, the figure shows there is room for improvement in ***XBLang***. After profiling RSBench and XSBench using *rocprof*, we became aware that ***XBLang***'s kernels appear to be faster than the other compilers, with the performance downgrade coming from ***XBLang***'s runtime for mapping memory to and back from the device.

We investigated why ***XBLang*** performs so much better in specific programs like CG. For this, we profiled both Clang and ***XBLang*** versions of the benchmark using the NVIDIA profiling tool Nsys on Perlmutter's A100. From these profiles, we noticed that we overperform in kernels that perform reductions. For example, in kernels with reduction having a reduction, ***XBLang***'s version of the kernel is 2.9 times faster than Clang's version. After looking at the LLVM IR generated by Clang and speaking to LLVM OpenMP developers, we can confirm that ***XBLang***'s implementation of reductions is the primary source for overperformance.

Fig. 5 presents results using Perlmutter CPU-only nodes. The compilers used on Perlmutter include Clang 18.0.0 (8823e961), GCC OpenMP offloading 12.2.0, Clang OpenMP offloading 18.0.0 (8823e961), nvc OpenMP offloading 23.5.0, Cray OpenMP offloading 15.0.1, nvc OpenACC 23.5.0 and our ***XBLang*** compiler. We observe that ***XBLang*** performs similarly to Clang in 3 of the 5 benchmarks, while ***XBLang***'s CG and MiniWeather performance is not as good. After further analysis, we concluded that the reason for ***XBLang***'s performance degradation is a bug where certain variables are not being privatized correctly; if said variables are manually privatized, then performance is again comparable to Clang's.

XBLang **on 64-bit ARM architecture:** We evaluated ***XBLang*** on a Fujitsu A64fx processor housed in the system (name and location omitted for double-blind reviews); this processor is the same as the one found in the Fugaku Supercomputer in Japan. For full specs of the test machine, please refer to [45]. Since the difference in results between a traditional X86 and an A64fx was not significant enough, we have not discussed them in detail here, but for further reference on Wombat results, please refer to [40].

In summary, results show that ***XBLang*** for GPUs, with its simplified syntax and semantics, has the potential to be either at par or better than directive-based compilers for GPUs and comparable performance for CPUs. While we have more experiments to perform, we hypothesize that ***XBLang*** offers the programmer and compiler researchers a lightweight language and compiler. It is quite appealing that our language

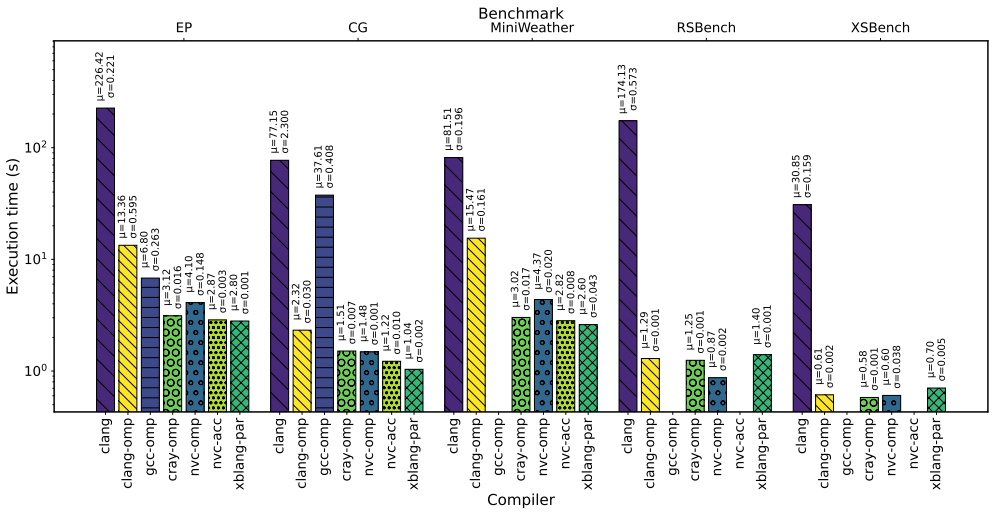

Figure 3: **Lower the better:** Execution time for five benchmarks on Perlmutter NVIDIA A100 GPUs comparing **XBLang** with other compilers; clang (first bar) represents the sequential run while the rest represent GPU runs; *gcc-omp* offloading for RSBench failed to compile; *gcc-omp* offloading for MiniWeather and XSBench failed to execute; RSBench and XSBench do not have an OpenACC equivalent hence *nvc-acc* bar does not exist. μ represents the average execution time, while σ is the standard deviation. All benchmarks were compiled using only *-O3* flags

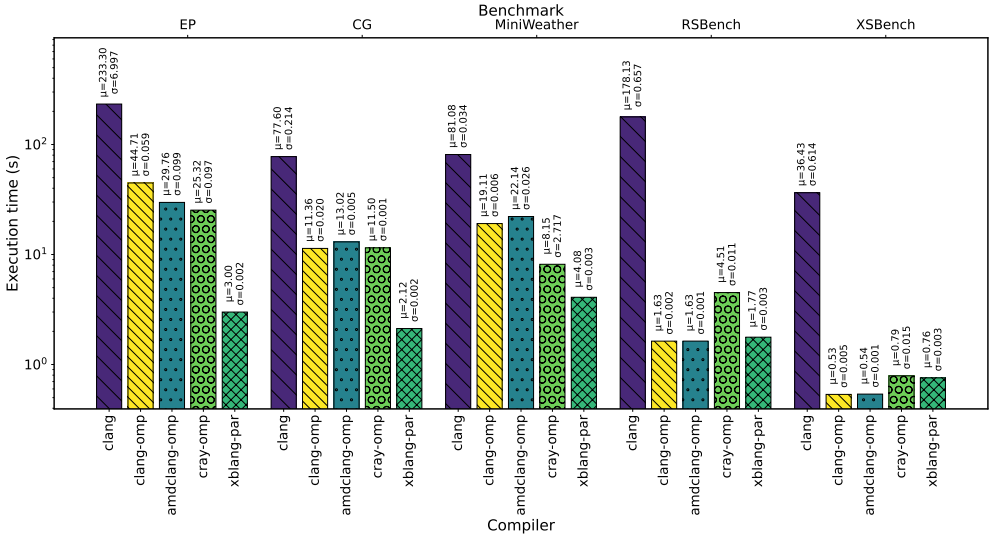

Figure 4: **Lower the better:** Execution time for five benchmarks on Frontier AMD MI250x GPUs comparing **XBLang** with other compilers; clang (first bar) represents the sequential run while the rest represent GPU runs. μ represents the average execution time, while σ is the standard deviation. All benchmarks were compiled using only *-O3* flags

is at par with vendor compilers and other compilers such as GNU and Clang under multiple situations.

VII. Related work

This section presents the related work, divided into high-level MLIR front-ends and new programming languages created to tackle the challenges of an evolving hardware landscape.

The COMET compiler [46] is a "Domain Specific Language" for sparse and dense tensor algebra computations based on MLIR. It leverages MLIR at multiple stages to perform high and low-level transformations, such as transforming Tensor contractions into Transpose-Transpose-GEMM-Transpose operations. It also supports execution in multiple targets, including CPUs and GPUs. However, despite its impressive performance success story, as a DSL, it is not a general front-end for MLIR.

Polygeist [47] is an MLIR front-end for a subset of C/C++ designed to leverage MLIR's representation power and existing dialects. It uses the Clang Libtooling API to process the C++ input, traversing the AST emitting MLIR. Polygeist uses the

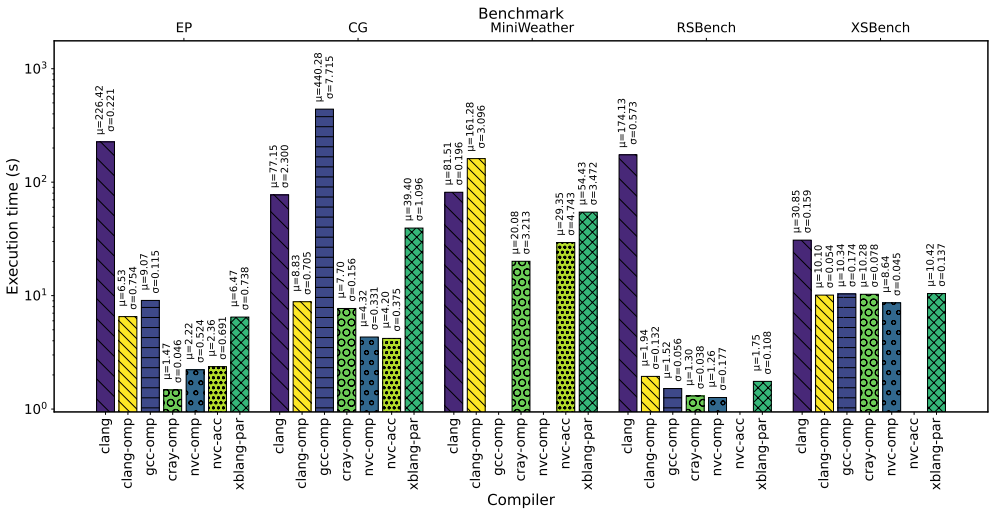

Figure 5: **Lower the better:** Execution time for five benchmarks on Perlmutter's AMD EPYC 7763 CPUs comparing **XBLang** with other compilers; clang (first bar) represents the sequential run while the rest represent multi-core runs; *gcc-omp* and *nvc-omp* for MiniWeather failed to execute; RSBench and XSBench do not have an OpenACC equivalent hence *nvc-acc* bar does not exist. μ represents the average execution time, while σ is the standard deviation. All benchmarks were compiled using only *-O3* flags

emitted MLIR to perform polyhedral optimization and parallel code transpilation, achieving excellent performance in both cases. Polygeist is a front-end based in Clang for C/C++; hence, by design, it is not an extensible front-end for MLIR. Instead, it should be regarded as a high-level C/C++ code optimizer.

Mojo [48] is a programming language introduced in 2023 by Modular. Mojo is a Python superset capable of delivering up to 68,000x speedup over Python [49]. It uses MLIR for its internal representation, enabling it to perform optimizations at a higher level. However, GPU support is still under development, with no performance results. It is also important to mention that porting HPC codes to a Pythonic language might pose additional challenges compared to other options due to the high-level essence of Pythonic languages and the traditional low-level nature of HPC codes.

Halide [50] is a programming language used for high-performance image and array processing. Programs are written in C++ using Halide's API rather than being a standalone language. The API then creates an internal representation, with the compiler transforming this IR to increase performance and then lowering it to LLVM IR. It supports multiple CPU and GPU architectures.

Exo [51] is a domain-specific language that uses the principle of exocompilation: externalizing target-specific code generation support and optimization policies to user-level code, i.e., the developer decides which optimizations to perform and when. An LLVM/MLIR is currently under construction [52].

There are other up-and-coming models such as Carbon [53], an experimental programming language and an experimental replacement to C++, introduced by Google in 2022, and Triton [54], a language for facilitating the writing of efficient Deep-Learning primitives, being developed by OpenAI and using MLIR for its IR.

The main difference between ***XBLang*** and previous works is that it is designed as an extensible front-end for MLIR, thus creating the necessary tools for HPC and compiler practitioners to face the challenges of hardware evolution using a high-level language. We demonstrate the power of extensibility with the **`par`** extension, a general-purpose portable parallel programming model.

VIII. Conclusions and Future Work

We introduced ***XBLang***, an extensible language front-end for MLIR, and **`par`**, an MLIR dialect for expressing parallelism, with GPU and CPU lowerings to LLVM. Both were created for compiler developers wanting to develop new programming language abstractions. Promising results show potential for this novel direction of compiler research. As immediate future steps, we will create a translator to convert a subset of conventional C codes to ***XBLang*** so that any programmer can seamlessly port their codes to our language without needing to go through the learning curve of understanding a new language. We are also in the process of adding support for Intel GPUs by using the SPIR-V MLIR dialect. We will also introduce meta-programming capabilities, templates, and other high-level constructs to ***XBLang*** programming language and increase the number of constructs available in the **`par`** dialect. We will evaluate the updated language and compiler with more tests and mini-apps.

References

[1] OpenMP ARB, "OpenMP 5.2," 2020. [Online]. Available: https://www.openmp.org/wp-content/uploads/OpenMP-API-Specification-5-2.pdf

[2] OpenACC, "OpenACC 3.2," 2020. [Online]. Available: https://www.openacc.org/sites/default/files/inline-images/Specification/OpenACC-3.1-final.pdf

[3] E. Zenker, B. Worpitz, R. Widera, A. Huebl, G. Juckeland, A. Knüpfer, W. E. Nagel, and M. Bussmann, "Alpaka - an abstraction library for parallel kernel acceleration." IEEE Computer Society, May 2016. [Online]. Available: http://arxiv.org/abs/1602.08477

[4] T. Huber, S. Pophale, N. Baker, M. Carr, N. Rao, J. Reap, K. Holsapple, J. H. Davis, T. Burnus, S. Lee, D. E. Bernholdt, and S. Chandrasekaran, "Ecp sollve: Validation and verification testsuite status update and compiler insight for openmp," in *2022 IEEE/ACM International Workshop on Performance, Portability and Productivity in HPC (P3HPC)*, 2022, pp. 123–135.

[5] I. Karlin, T. Scogland, A. C. Jacob, S. F. Antao, G.-T. Bercea, C. Bertolli, B. R. d. Supinski, E. W. Draeger, A. E. Eichenberger, J. Glosli *et al.*, "Early experiences porting three applications to openmp 4.5," in *International Workshop on OpenMP*. Springer, 2016, pp. 281–292.

[6] G. Juckeland, O. Hernandez, A. C. Jacob, D. Neilson, V. G. V. Larrea, S. Wienke, A. Bobyr, W. C. Brantley, S. Chandrasekaran, M. Colgrove *et al.*, "From describing to prescribing parallelism: Translating the spec accel openacc suite to openmp target directives," in *International Conference on High Performance Computing*. Springer, 2016, pp. 470–488.

[7] R. Gayatri, C. Yang, T. Kurth, and J. Deslippe, "A case study for performance portability using openmp 4.5," in *International Workshop on Accelerator Programming Using Directives*. Springer, 2018, pp. 75–95.

[8] J. M. Diaz, S. Pophale, K. Friedline, O. Hernandez, D. E. Bernholdt, and S. Chandrasekaran, "Evaluating support for openmp offload features," in *Proceedings of the 47th International Conference on Parallel Processing Companion*, 2018, pp. 1–10.

[9] J. H. Davis, C. Daley, S. Pophale, T. Huber, S. Chandrasekaran, and N. J. Wright, "Performance assessment of openmp compilers targeting nvidia v100 gpus," in *International Workshop on Accelerator Programming Using Directives*. Springer, 2020, pp. 25–44.

[10] C. Daley, H. Ahmed, S. Williams, and N. Wright, "A case study of porting hpgmg from cuda to openmp target offload," in *International Workshop on OpenMP*. Springer, 2020, pp. 37–51.

[11] B. Chapman, B. Pham, C. Yang, C. Daley, C. Bertoni, D. Kulkarni, D. Oryspayev, E. D'Azevedo, J. Doerfert, K. Zhou *et al.*, "Outcomes of openmp hackathon: Openmp application experiences with the offloading model (part ii)," in *OpenMP: Enabling Massive Node-Level Parallelism: 17th International Workshop on OpenMP, IWOMP 2021, Bristol, UK, September 14–16, 2021, Proceedings 17*. Springer, 2021, pp. 81–95.

[12] S. Bak, C. Bertoni, S. Boehm, R. Budiardja, B. M. Chapman, J. Doerfert, M. Eisenbach, H. Finkel, O. Hernandez, J. Huber *et al.*, "Openmp application experiences: porting to accelerated nodes," *Parallel Computing*, vol. 109, p. 102856, 2022.

[13] T. M. Evans, A. Siegel, E. W. Draeger, J. Deslippe, M. M. Francois, T. C. Germann, W. E. Hart, and D. F. Martin, "A survey of software implementations used by application codes in the exascale computing project," *The International Journal of High Performance Computing Applications*, vol. 36, no. 1, pp. 5–12, 2022. [Online]. Available: https://doi.org/10.1177/10943420211028940

[14] N. D. Matsakis and F. S. Klock, "The rust language," *ACM SIGAda Ada Letters*, vol. 34, no. 3, pp. 103–104, 2014.

[15] G. VanRossum and F. L. Drake, *The python language reference*. Python Software Foundation Amsterdam, Netherlands, 2010.

[16] W.-C. Lin and S. McIntosh-Smith, "Comparing julia to performance portable parallel programming models for hpc," in *2021 International Workshop on Performance Modeling, Benchmarking and Simulation of High Performance Computer Systems (PMBS)*. IEEE, 2021, pp. 94–105.

[17] C. Dave, H. Bae, S.-J. Min, S. Lee, R. Eigenmann, and S. Midkiff, "Cetus: A source-to-source compiler infrastructure for multicores," *Computer*, vol. 42, no. 12, pp. 36–42, 2009.

[18] S. Lee and J. S. Vetter, "OpenARC: open accelerator research compiler for directive-based, efficient heterogeneous computing," in *Proceedings of the 23rd international symposium on High-performance parallel and distributed computing*, ser. HPDC '14. New York, NY, USA: Association for Computing Machinery, Jun. 2014, pp. 115–120. [Online]. Available: https://doi.org/10.1145/2600212.2600704

[19] ——, "OpenARC: Extensible OpenACC Compiler Framework for Directive-Based Accelerator Programming Study," in *2014 First Workshop on Accelerator Programming using Directives*, Nov. 2014, pp. 1–11.

[20] D. Quinlan and C. Liao, "The ROSE source-to-source compiler infrastructure," in *Cetus users and compiler infrastructure workshop, in conjunction with PACT*, vol. 2011. Citeseer, 2011, p. 1.

[21] J. E. Denny, S. Lee, and J. S. Vetter, "Clacc: Translating openacc to openmp in clang," in *2018 IEEE/ACM 5th Workshop on the LLVM Compiler Infrastructure in HPC (LLVM-HPC)*. IEEE, 2018, pp. 18–29.

[22] X. Wu, M. Kruse, P. Balaprakash, H. Finkel, P. Hovland, V. Taylor, and M. Hall, "Autotuning polybench benchmarks with llvm clang/polly loop optimization pragmas using bayesian optimization (extended version)," *arXiv preprint arXiv:2104.13242*, 2021.

[23] G. D. Balogh, G. R. Mudalige, I. Z. Reguly, S. Antao, and C. Bertolli, "Op2-clang: A source-to-source translator using clang/llvm libtooling," in *2018 IEEE/ACM 5th Workshop on the LLVM Compiler Infrastructure in HPC (LLVM-HPC)*. IEEE, 2018, pp. 59–70.

[24] M. R. Gadelha, J. Morse, L. Cordeiro, and D. Nicole, "Using clang as a frontend on a formal verification tool," 2017.

[25] J. Balart, A. Duran, M. Gonzàlez, X. Martorell, E. Ayguadé, and J. Labarta, "Nanos mercurium: a research compiler for openmp," in *Proceedings of the European Workshop on OpenMP*, vol. 8, 2004, p. 56.

[26] P. Gschwandtner, J. J. Durillo, and T. Fahringer, "Multi-objective auto-tuning with insieme: Optimization and trade-off analysis for time, energy and resource usage," in *European Conference on Parallel Processing*. Springer, 2014, pp. 87–98.

[27] R. Milewicz, P. Pirkelbauer, P. Soundararajan, H. Ahmed, and T. Skjellum, "Negative perceptions about the applicability of source-to-source compilers in hpc: A literature review," in *International Conference on High Performance Computing*. Springer, 2021, pp. 233–246.

[28] C. Lattner and V. Adve, "LLVM: a compilation framework for lifelong program analysis amp; transformation," in *International Symposium on Code Generation and Optimization, 2004. CGO 2004.*, Mar. 2004, pp. 75–86.

[29] ClangIR · A new high-level IR for clang. LLVM. [Online]. Available: https://llvm.github.io/clangir/

[30] The Flang Compiler. LLVM. [Online]. Available: https://flang.llvm.org/docs/

[31] MLIR, "Mlir," 2023. [Online]. Available: https://mlir.llvm.org/

[32] A. Zinenko. 2023 LLVM Dev Mtg - MLIR Is Not an ML Compiler, and Other Common Misconceptions. LLVM. [Online]. Available: https://youtu.be/lXAp6ZAWyBY

[33] C. Lattner, M. Amini, U. Bondhugula, A. Cohen, A. Davis, J. Pienaar, R. Riddle, T. Shpeisman, N. Vasilache, and O. Zinenko, "Mlir: Scaling compiler infrastructure for domain specific computation," in *2021 IEEE/ACM International Symposium on Code Generation and Optimization (CGO)*, 2021, pp. 2–14.

[34] "Mlir language reference," 2023. [Online]. Available: https://mlir.llvm.org/docs/LangRef

[35] LLVM, "TableGen Programmer's Reference - LLVM," 2023. [Online]. Available: https://llvm.org/docs/TableGen/ProgRef.html

[36] W. Estes, "Lexical analysis with flex, for flex 2.6.2: Top," 2017. [Online]. Available: https://westes.github.io/flex/manual/

[37] F. S. Foundation, "Bison - gnu project - free software foundation," 2021. [Online]. Available: https://www.gnu.org/software/bison/

[38] T. J. Parr and R. W. Quong, "Antlr: A predicated-ll(k) parser generator," *Software: Practice and Experience*, vol. 25, no. 7, pp. 789–810, 1995. [Online]. Available: https://onlinelibrary.wiley.com/doi/abs/10.1002/spe.4380250705

[39] B. Ford, "Packet parsing: a practical linear-time algorithm with backtracking," Ph.D. dissertation, Massachusetts Institute of Technology, 2002.

[40] Anonymous, "High-Performance GPU offloading using XBLang: an extensible language front-end for MLIR," Dec. 2023. [Online]. Available: https://doi.org/10.5281/zenodo.10393464

[41] K. Asanovic, R. Bodik, J. Demmel, T. Keaveny, K. Keutzer, J. Kubiatowicz, N. Morgan, D. Patterson, K. Sen, J. Wawrzynek, D. Wessel, and K. Yelick, "A view of the parallel computing landscape," *Commun. ACM*, vol. 52, no. 10, p. 56–67, oct 2009. [Online]. Available: https://doi.org/10.1145/1562764.1562783

[42] G. Juckeland, W. Brantley, S. Chandrasekaran, B. Chapman, S. Che, M. Colgrove, H. Feng, A. Grund, R. Henschel, W.-M. W. Hwu, H. Li, M. S. Müller, W. E. Nagel, M. Perminov, P. Shelepugin, K. Skadron, J. Stratton, A. Titov, K. Wang, M. van Waveren, B. Whitney, S. Wienke, R. Xu, and K. Kumaran, "Spec accel: A standard application suite

for measuring hardware accelerator performance," in *High Performance Computing Systems. Performance Modeling, Benchmarking, and Simulation*, S. A. Jarvis, S. A. Wright, and S. D. Hammond, Eds. Cham: Springer International Publishing, 2015, pp. 46–67.

[43] J. Li, A. Bobyr, S. Boehm, W. Brantley, H. Brunst, A. Cavelan, S. Chandrasekaran, J. Cheng, F. M. Ciorba, M. Colgrove, T. Curtis, C. Daley, M. Ferrato, M. G. de Souza, N. Hagerty, R. Henschel, G. Juckeland, J. Kelling, K. Li, R. Lieberman, K. McMahon, E. Melnichenko, M. A. Neggaz, H. Ono, C. Ponder, D. Raddatz, S. Schueller, R. Searles, F. Vasilev, V. M. Vergara, B. Wang, B. Wesarg, S. Wienke, and M. Zavala, "Spechpc 2021 benchmark suites for modern hpc systems," in *Companion of the 2022 ACM/SPEC International Conference on Performance Engineering*, ser. ICPE '22. New York, NY, USA: Association for Computing Machinery, 2022, p. 15–16. [Online]. Available: https://doi.org/10.1145/3491204.3527498

[44] ECP Proxy Applications. Exascale Computing Project. [Online]. Available: https://proxyapps.exascaleproject.org/app/

[45] Oak Ridge National Lab, "Wombat." [Online]. Available: https://www.olcf.ornl.gov/olcf-resources/compute-systems/wombat/

[46] R. Tian, L. Guo, J. Li, B. Ren, and G. Kestor, "A high performance sparse tensor algebra compiler in mlir," in *2021 IEEE/ACM 7th Workshop on the LLVM Compiler Infrastructure in HPC (LLVM-HPC)*, 2021, pp. 27–38.

[47] W. S. Moses, I. R. Ivanov, J. Domke, T. Endo, J. Doerfert, and O. Zinenko, "High-performance gpu-to-cpu transpilation and optimization via high-level parallel constructs," in *Proceedings of the 28th ACM SIGPLAN Annual Symposium on Principles and Practice of Parallel Programming*, ser. PPoPP '23. New York, NY, USA: Association for Computing Machinery, 2023, p. 119–134. [Online]. Available: https://doi.org/10.1145/3572848.3577475

[48] Modular Inc, "Mojo programming manual," 2023. [Online]. Available: https://docs.modular.com/mojo/programming-manual.html

[49] ——, "A journey to 68,000x speedup over python - part 3," 2023. [Online]. Available: https://www.modular.com/blog/mojo-a-journey-to-68-000x-speedup-over-python-part-3

[50] J. Ragan-Kelley, A. Adams, D. Sharlet, C. Barnes, S. Paris, M. Levoy, S. Amarasinghe, and F. Durand, "Halide: Decoupling algorithms from schedules for high-performance image processing," *Commun. ACM*, vol. 61, no. 1, p. 106–115, dec 2017. [Online]. Available: https://doi.org/10.1145/3150211

[51] Y. Ikarashi, G. L. Bernstein, A. Reinking, H. Genc, and J. Ragan-Kelley, "Exocompilation for productive programming of hardware accelerators," in *Proceedings of the 43rd ACM SIGPLAN International Conference on Programming Language Design and Implementation*, ser. PLDI 2022. New York, NY, USA: Association for Computing Machinery, 2022, p. 703–718. [Online]. Available: https://doi.org/10.1145/3519939.3523446

[52] exo lang, "The exo language," 2022. [Online]. Available: https://exo-lang.dev/

[53] Google, "Google brands carbon language as "experimental successor to c++"," 2022. [Online]. Available: https://devclass.com/2022/07/20/google-brands-carbon-language-as-experimental-successor-to-c/

[54] P. Tillet, H. T. Kung, and D. Cox, "Triton: An intermediate language and compiler for tiled neural network computations," in *Proceedings of the 3rd ACM SIGPLAN International Workshop on Machine Learning and Programming Languages*, ser. MAPL 2019. New York, NY, USA: Association for Computing Machinery, 2019, p. 10–19. [Online]. Available: https://doi.org/10.1145/3315508.3329973