

Fortran performance optimisation and auto-parallelisation by leveraging MLIR-based domain specific abstractions in Flang

Nick Brown
n.brown@ed.ac.uk
EPCC at the University of Edinburgh
Edinburgh, UK

Maurice Jamieson
EPCC at the University of Edinburgh
Edinburgh, UK

Anton Lydike
The School of Informatics, University
of Edinburgh
Edinburgh, UK

Emilien Bauer
The School of Informatics, University
of Edinburgh
Edinburgh, UK

Tobias Grosser
The School of Informatics, University
of Edinburgh
Edinburgh, UK

ABSTRACT

MLIR has become popular since it was open sourced in 2019. A sub-project of LLVM, the flexibility provided by MLIR to represent Intermediate Representations (IR) as dialects at different abstraction levels, to mix these, and to leverage transformations between dialects provides opportunities for automated program optimisation and parallelisation. In addition to general purpose compilers built upon MLIR, domain specific abstractions have also been developed.

In this paper we explore complimenting the Flang MLIR general purpose compiler by combining with the domain specific Open Earth Compiler's MLIR stencil dialect. Developing transformations to discover and extract stencils from Fortran, this specialisation delivers between a 2- and 10-times performance improvement for our benchmarks on a Cray supercomputer compared to using Flang alone. Furthermore, by leveraging existing MLIR transformations we develop an auto-parallelisation approach targeting multi-threaded and distributed memory parallelism, and optimised execution on GPUs, without any modifications to the serial Fortran source code.

KEYWORDS

LLVM, MLIR, xDSL, stencil based computation, HPC

1 INTRODUCTION

Since it was first released open source by Google in 2019 and then merged into the LLVM codebase as a sub-project thereafter, MLIR [14] has gained significant popularity. Enabling hierarchies of Intermediate Representations (IR) to be expressed and mixed in a structured manner, and for lowerings between these abstraction levels to be provided, numerous tools and technologies have been developed that leverage MLIR in their flow.

One such class of tools that can strongly benefit from MLIR is that of Domain Specific Languages (DSLs), such as [13], [17], [4], which provide domain specific abstractions that enable programmes to express their problem in a high level, abstract, fashion. It is possible, using MLIR, to develop IR dialects that closely match these domain specific abstractions, and for these dialects to then be lowered to more general purpose abstractions which themselves benefit from existing lowerings to LLVM-IR and ultimately the LLVM backends. Given the rich semantic information that can

often be found in domain specific abstractions about a programmer's intentions, it is often possible for the compiler framework to drive key decisions around performance and parallelism using this high level information far more effectively than, for example, working with the lower-level LLVM-IR directly.

General purpose compilers, such as Flang [10] for Fortran, Polygeist [16] for C++, and Pylir for Python [19], have been developed which sit atop the MLIR ecosystem. Whilst these leverage the central ideas of MLIR, these general purpose language compilers do not tend to fully exploit the domain specific abstractions that are found in some MLIR dialects.

In this paper we explore the potential of combining MLIR-based general purpose compilers with domain specific MLIR abstractions by enriching the Flang compiler with the MLIR stencil dialect from the Open Earth Compiler [12]. By automatically identifying and translating appropriate Fortran constructs into the stencil dialect during compilation, we aim to explore whether improved performance and new capabilities can be unlocked by leveraging this domain specific specialisation built atop MLIR. This paper is structured as follows; Section 2 describes the background to this work, exploring the central building blocks such as MLIR, the stencil dialect, Flang and xDSL that we use in this work. Our stencil-based Flang compilation flow is then described in Section 3, where we highlight our approach to Fortran code stencil identification, explore how existing MLIR passes can be leveraged by our approach to target a variety of architectures and work around some of the challenges present in Flang. Performance of our approach is compared against that of using Flang alone and the Cray compiler (for CPUs) and Nvidia HPC SDK (for GPUs) in Section 4 across single CPU core, multi-threaded and distributed memory parallelism, and GPUs. Section 5 then surveys how our contribution relates to existing work, before Section 6 draws conclusions and discusses further work.

The contributions of this paper are as follows:

- We demonstrate that, by exploiting stencil-based domain specific information one can deliver improved performance compared to using general purpose Flang alone for stencil codes.
- Illustrate that domain specific MLIR abstractions and the wider ecosystem can enable the targeting of multiple HPC architectures and scales of parallelism in an automated fashion without requiring source code modifications.

- Providing a wider study around the benefits that the composability of MLIR dialects can provide to compliment general purpose compilers for high performance workloads.

2 BACKGROUND

At its core LLVM provides numerous language frontends and architecture specific backends which are connected via LLVM-IR. An LLVM frontend, such as Clang, generating LLVM-IR can therefore target any backend, and backends for a wide range of architectures including CPUs, GPUs, and FPGAs have been developed. However, LLVM-IR itself is low level, requiring significant work by the frontends in lowering to this level and resulting in potential redundancies between them.

MLIR aims to address this issue by providing a series of IR dialects and transformations between these, so that frontends can instead translate to more suitable, higher level intermediate representations. MLIR is also a framework where developers can add their own dialects and transformations, and dialects can be mixed and manipulated at different levels of abstraction, enabling progressive lowering of the abstraction level to LLVM-IR. Much of this lowering is undertaken by existing dialects and transformations, thus significantly reducing the overall software effort in developing compilers by promoting reuse between them.

MLIR has become popular since it became a sub-project of LLVM, and has the potential to revolutionise compiler development. Providing many dialects as standard, such as *arith* for arithmetic operations, *scf* for structured control flow which provides serial and parallel loops, *memref* for memory management and data access, *openmp* for OpenMP parallelism, *gpu* for GPU execution, and *vector* for vectorisation. Transformations exist which will manipulate dialects and lower between them. For instance, there are lowerings from the dialects listed above to the *llvm* dialect which corresponds to LLVM-IR. Once lowered, it is possible to generate LLVM-IR from the *llvm* dialect and for this to be provided to the LLVM backends.

2.1 xDSL

Arguably one of the challenges faced by MLIR is the steep learning curve associated with the technology. Requiring the developer to leverage C++, understand LLVM concepts, and work with the Tablegen format in order to describe dialects raises the overhead involved in development.

By contrast, xDSL [22] is a Python based compiler design toolkit which is 1-1 compatible with MLIR. Providing not only the majority of MLIR dialects, but also numerous additional experimental dialects too, these are expressed in IRDL [8] format within Python classes. xDSL enables a rapid exploration and prototyping of MLIR dialects and concepts, with a view to then committing the mature dialects and transformations into the main MLIR codebase once the concepts are proven. As xDSL is 1-1 compatible with MLIR, one is able to arbitrarily go between the two technologies during compilation.

In addition to providing many of the standard MLIR dialects ¹, xDSL also provides additional dialects such as Distributed Memory Parallelism (DMP) which, in a technology agnostic manner, expresses parallelism across nodes. DMP can then be lowered to the

¹Providing a complete set of MLIR dialects is planned by the xDSL developers

MPI xDSL dialect, which specialises DMP to leverage MPI which is then lowered to interaction with the MPI library via the *func* dialect.

2.2 Stencil dialect

```

1  do i = 2, 255
2    do j = 2, 255
3      data(j,i) = (data(j,i-1)+data(j,i+1)+data(j-1,i)+data(j+1,
4                    i)) * 0.25
5    enddo
6  enddo

```

Listing 1 Sketch of Fortran stencil code example which averages all neighbouring values across a grid

A dialect provided by xDSL is the *stencil* dialect which was initially developed by ETH Zurich as part of the Open Earth Compiler [12]. A stencil is a geometric arrangement of a group of neighbouring grid cells that, by using a numerical approximation routine, relate to a specific grid cell of interest. The stencil dialect expresses stencil calculations such as that illustrated in Listing 1 which is calculating an average of values across neighbouring grid cells in two dimensions. Driven by nested loops, two in Listing 1, the loop variables are used as array indices typically with some offset applied such as *data(j,i-1)* which accesses the grid cell at the same location in the first dimension and one step before in the second dimension (arrays index precedence is from left to right in Fortran).

Stencil-based calculations are extremely common in computational simulation codes, for instance to solve systems of PDEs, [3] [21] [6], and Listing 2 sketches the corresponding IR in Static Single Assignment (SSA) form using the stencil dialect for the calculation of Listing 1. It can be seen that the nested loops at lines 1 and 2 of Listing 1 have been transformed into the *stencil.apply* operator at line 1 of Listing 2 which accepts a *stencil.temp* field as an input argument. There are additional stencil operators to convert from a *memref* to this *stencil.temp* type, but these have been omitted for brevity. The *stencil.access* operations at lines 4 to 7 of Listing 2 correspond to accesses on the *data* array at line 3 of Listing 1 and each of these operations loads a specific neighbouring grid cell. Lines 8 to 11 then use the standard *arith* dialect to undertake the calculation which is returned from the stencil apply operator block at line 12. It is important to highlight that this *stencil.apply* operator is running across the entire grid, whose lower and upper bounds are determined by the types of the input and output fields, effectively executing lines 3 to 12 for every grid cell.

It should be stressed that the code provided in this section is just an example and there is no existing lowering from Fortran to the stencil dialect. Indeed it is the objective of the work reported in this paper to be able to generate SSA such as that of Listing 2 from Fortran code. Our hypothesis is that, by transforming and simplifying applicable loops and their calculations into the stencil dialect, we unlock the potential for richer and more complex optimisations by the compiler framework. Driven by existing transformations, some of which bespoke for the stencil dialect and others standard MLIR passes, it is then possible to target different architectures such as

```

1  %result = "stencil.apply"(%18) ({
2  ^0(%data : !stencil.temp<[-1,255]x[-1,255]xf64>):
3    %c0 = arith.constant 2.500000e-01 : f64
4    %d0 = "stencil.access"(%data) {"offset" = #stencil.index<0, -1>} : (!stencil.temp<[-1,255]x[-1,255]xf64>) -> f64
5    %d1 = "stencil.access"(%data) {"offset" = #stencil.index<0, 1>} : (!stencil.temp<[-1,255]x[-1,255]xf64>) -> f64
6    %d2 = "stencil.access"(%data) {"offset" = #stencil.index<-1, 0>} : (!stencil.temp<[-1,255]x[-1,255]xf64>) -> f64
7    %d3 = "stencil.access"(%data) {"offset" = #stencil.index<-1, 0>} : (!stencil.temp<[-1,255]x[-1,255]xf64>) -> f64
8    %t0 = arith.addf %d3, %d2 : f64
9    %t1 = arith.addf %t0, %d1 : f64
10   %t2 = arith.addf %t1, %d0 : f64
11   %t3 = arith.mulf %t2, %c0 : f64
12   "stencil.return"(%t3) : (f64) -> ()
13 }) : (!stencil.temp<[-1,255]x[-1,255]xf64>) -> !stencil.temp<[0,254]x[0,254]xf64>

```

Listing 2 Sketch of corresponding SSA-based IR leveraging the stencil dialect to represent the stencil calculation of Listing 1

CPUs and GPUs, as well as implicit shared memory and distributed memory parallelism in an efficient manner.

2.3 Flang

Flang is the LLVM Fortran frontend, and whilst there was a previous Flang compiler for several years, known as classic Flang, this was never an official LLVM project and has been recently replaced with a ground-up rewrite built on-top of MLIR. This new Flang compiler is an official component of LLVM, and the objective is to support the full range of standard Fortran, including future versions of the language. However, at the time of writing support for Fortran at or beyond 2003 is still work in progress and yet to reach full maturity.

After lexing and parsing of Fortran code, this is then lowered to the Fortran IR (FIR) [9] dialect² which provides IR level expression of Fortran constructs and concepts. However, the developers of Flang have adopted a surprising design decision where FIR is then lowered directly to LLVM-IR by Flang, rather than lowering first to more general MLIR dialects such as scf and then leveraging existing MLIR passes to optimise these and ultimately lower to the llvm dialect.

Furthermore, only a small subset of the standard MLIR dialects are registered in Flang, and conversely *mlir-opt* is unaware of the FIR dialect. Consequently interoperability between FIR and many of the standard dialects is limited, which was a challenge this work and is explored further in Section 3.

3 COMBINING FLANG WITH STENCIL DIALECT ABSTRACTIONS

Figure 1 illustrates the overall architecture of our approach, where Fortran source code is processed by Flang and, using the *-fc1 -emit-mlir* flags, we output the corresponding SSA-based IR in the FIR MLIR dialect.

The SSA-based IR is loaded by xDSL which provides the FIR dialect, and our *discover stencils* transformation is then run on this FIR IR to identify any stencil calculations that are being undertaken. Listing 3 sketches the algorithm used by this pass, where

all the loops in the code are first gathered and then all FIR *store* operations in the module are identified and iterated over. It is first determined, for a store operation, whether it is indexed by loops in the program (line 5 of Listing 3) and this involves walking the IR to find the corresponding *fir.coordinate_of* operation which provides the indices into the array. The expressions which contribute to each index of *fir.coordinate_of* operation are then walked to extract the calculations involved, for instance the array access *data(i,j)* of the example in Listing 1 will identify that the first dimension is indexed by variable *i* and the second dimension by *j*. These extracted variable indices are then compared against the gathered loops to determine whether the store is being driven by nested loops and if so *is_indexed_by_loops* returns true.

The SSA-based IR is then walked to locate array read operations which are on the right hand side of the calculation. In addition to capturing the array variable being accessed, array indices are walked in a manner similar to the store operation however in addition to capturing the index variables, this also identifies offsets where, for example, *data(j,i-1)* would record *j* indexing the first dimension and *i* minus 1 as the second dimension. Using this information, the list of arrays that are accessed as part of this calculation are determined at line 7 and then, for each of these, the corresponding stencil loading operations are then generated at line 10, with the operations and SSA value returned, the later then used as input to the *stencil.apply* operation at line 15.

The loops which drive this stencil calculation are then identified at line 13 of Listing 3, and from these the lower and upper bounds in each dimension. The *stencil.apply* operation is generated at line 15 and the *generate_stencil_apply* function first creates a *stencil.apply* based upon the lower and upper bounds and with each array load SSA result as an input, before then walking the SSA IR to identify the mathematical expressions and data dependencies that contribute to the stencil calculation. These mathematical expressions are removed from the original FIR code and placed into the stencil, with the data dependencies being replaced by the *stencil.apply* operation which sets the field offset based upon the offsets already gathered in *rhs_read_ops*, for instance 0 for dimension zero and minus 1 for dimension one in the example above.

Operations to store the result of applying the stencil back to the memref are generated at line 16 of Listing 3, before the applicable

²There is an HL-FIR dialect in development which provides higher level FIR operations although at the time of writing the default compilation flow is still via FIR

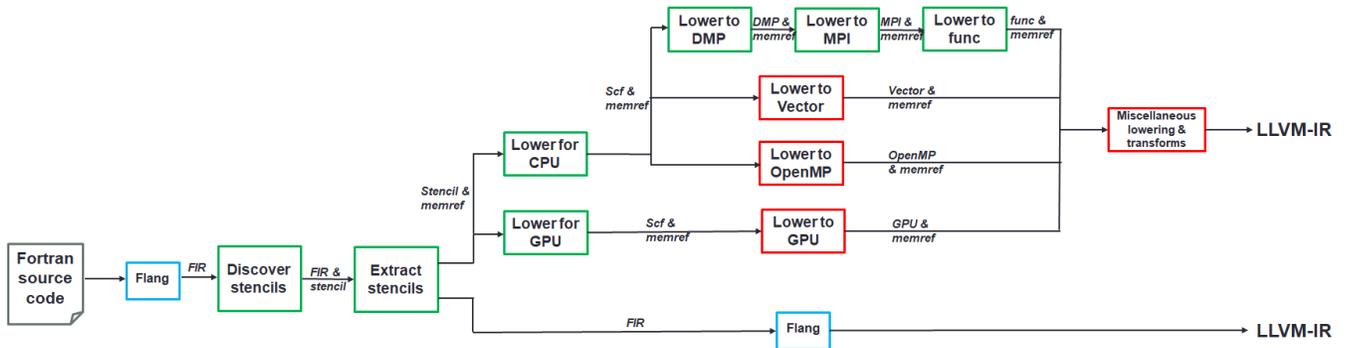


Figure 1: Architectural pipeline diagram of our approach, identifying and extracting stencils from Fortran code, by operating on the Flang generated FIR, and then leveraging appropriate transformations. Boxes represent transformations and annotated arrows the dialects. Blue boxes are activities in Flang, Green boxes are transformation passes executing in xDSL, and red boxes are passes running inside MLIR.

```

1  identified_stencils=[]
2  loops = gather_program_loops(module)
3
4  for store_op in module:
5      if is_indexed_by_loops(store_op, loops):
6          rhs_read_ops=get_array_read_data_ops(store_op)
7          unique_array_names=get_unique_array_names(rhs_read_ops)
8          load_ops=[], load_ssa_results=[]
9          for name in unique_array_names:
10             specific_ops, specific_ssa_results=generate_stencil_field_load(name)
11             load_ops=load_ops.append(specific_ops)
12             load_ssa_results=load_ssa_results.append(specific_ssa_results)
13             applicable_loops=get_applicable_loops(store_op, loops)
14             lb, ub=get_loop_bounds_per_dim(applicable_loops)
15             stencil_apply_ops=generate_stencil_apply(load_ssa_results, lb, ub, store_op)
16             store_ops=generate_stencil_store(store_op)
17             identified_stencils.append((applicable_loops, load_ops+stencil_apply_ops+store_ops))
18
19  for stencil in identified_stencils:
20     applicable_loops=stencil[0]
21     ops=stencil[1]
22     top_level_loop=find_top_level_loop(applicable_loops)
23     insert_ops_before(top_level_loop, ops)
24
25  for loop in loops:
26     if loop is empty:
27         remove_loop(loop)
28
29  merge_stencils_if_possible(module)

```

Listing 3 Sketch of stencil discovery algorithm used in the first transformation of our pipeline of Figure 1

loops that this stencil involves along with the generated instructions are stored in the *identified_stencils* list. The next step is then to add these stencils into the IR, which are added directly preceding the outer most loop involved in the stencil calculation. This is handled between lines 19 and 23, where the top level loop for the

stencil is identified at line 22 and the stencil operations added into the list of operations directly preceding this at line 23.

The bodies of all FIR loops are then walked at lines 25 to 27 and those loops which are now empty, because their operations have been transformed into the stencil dialect and extracted, are removed. Lastly, a pass is undertaken across the SSA-based IR to

merge any stencils that are located next to each other and share the same lower and upper bounds.

It should be noted that Listing 3 is a sketch of the algorithm, and there are some specific complexities omitted for brevity. For instance, differences in how arrays are represented in FIR by Flang if they are stack or heap allocated mean that there are different possible routes when walking backwards from *fir.store* and *fir.load* operations. Other features, such as accessing loop indexes, constants and non-stencil based variables within a stencil calculation are also supported and translated to the corresponding operations in the stencil dialect.

Once stencils have been identified, the SSA-based IR is now in the form of the FIR and stencil dialects being mixed together. However, this is problematic because, as explained in Section 2.3, Flang is unaware of many of the standard MLIR dialects, and likewise the MLIR driver tool, *mlir-opt*, is unaware of FIR. Consequently the FIR portions of the IR must be separated from that of the stencil dialect. This is performed by the *extract stencil* pass which lifts the stencil components out as a function into a separate MLIR module, which will then be called from FIR by a function call. As these are then separate MLIR modules they can then be compiled by different flows, i.e. one using Flang and the other *mlir-opt*, and linked together at runtime. Due to this requirement of the stencil dialect IR not containing any FIR dialect IR, during the phase in Listing 3 where we build up the *stencil.apply* operator by extracting the mathematical expressions from FIR into the stencil, we need to convert FIR operations into standard dialects. This is simplified significantly by the fact that Flang leverages the standard arith and math dialects for arithmetic and mathematical operations, which are registered with *mlir-opt* and-so can be handled. However, FIR data conversions and miscellaneous operations, such as *fir.no_reassoc* which is used to prevent operator reassociation, do need to be converted into their standard MLIR dialect counterparts.

A further challenge is in providing data interoperability between FIR, which contains its own bespoke data representation operations and types, and the stencil dialect which uses the standard memref dialect. The only way to support this is to convert the FIR data to an FIR *llvm_ptr* type, and pass this to the stencil function which then builds the memref from this. In-fact FIR is entirely isolated from a typing perspective, as one can only reduce to its own, FIR dialect, representation of an *llvm_ptr*, and not the *llvm_ptr* in the LLVM dialect. Furthermore, it is not possible in the FIR-based IR module to leverage an unrealized conversion cast, which would enable one to go from an FIR *llvm_ptr* to an LLVM *llvm_ptr* from a typing perspective, because the builtin dialect is not registered with Flang. However, *llvm_ptr* in the FIR and LLVM dialects is semantically identical and passing an argument of type FIR *llvm_ptr* to a function that accepts LLVM *llvm_ptr* is allowed when linking the resulting object files.

As per Figure 1, the IR module containing the stencil dialect portion is then transformed either for CPU or GPU using the xDSL stencil lowering. In fact these lowerings are the same source code, but driven by a command line option to tune for the architecture in question, for instance the CPU lowering converts the top level loop into *scf.parallel* and nested inner loops into *scf.for*, whereas the GPU lowering attempts to coalesce the loops into a single *scf.parallel*

loop. For GPU, shared memory parallelism and single core execution the lowered IR is then transformed by existing MLIR scf transformation passes, for instance *convert-scf-to-openmp* to lower to the OpenMP dialect, *convert-parallel-loops-to-gpu* for GPU execution, and *scf-for-loop-specialization* for vectorisation. Alternatively, if the programmer wishes to leverage distributed memory parallelism then they can apply the *lower to DMP* transformation which will transform to the DMP dialect in xDSL, which can then be lowered to MPI and the corresponding function calls via two subsequent passes.

Irrespective of the specific passes and targets, this IR is then transformed through a series of existing MLIR miscellaneous passes, such as lowering dialects such as math and memref to LLVM, reconciling unrealized conversion casts, and canonicalization. The MLIR pass pipeline for GPU target is reported in Listing 4 and there are several aspects to highlight. Firstly, for GPUs, we found that this was very sensitive and slight modifications to the pipeline would silently fail to generate GPU target binary code and run only on the CPU. As the GPU binary is embedded in the single generated MLIR CPU file, this is very easy to miss. Secondly, there is a sensitivity to the loop tiling factors on the GPU, provided to the *scf-parallel-loop-tiling* pass as arguments, as these can make both an impact on performance and furthermore some values can result in runtime failures on the GPU. We have had to find these optimal values empirically, although the values illustrated in Listing 4 perform well across a range of kernels, and an improved MLIR pass that avoids the need for these specific numbers would be beneficial. Thirdly, several transformations by default lower to opaque pointers, whereas in our flow we favour including the type and size. Consequently, we provide the option to use transparent pointers instead.

```
mlir-opt --pass-pipeline="builtin.module(test-math-
algebraic-simplification,scf-parallel-loop-tiling{
parallel-loop-tile-sizes=32,32,1},canonicalize,test-
expand-math,func.func(gpu-map-parallel-loops),
convert-parallel-loops-to-gpu, fold-memref-alias-
ops,finalize-memref-to-llvm{index-bitwidth=64 use-
opaque-pointers=false},lower-affine, gpu-kernel-
outlining,func.func(gpu-async-region),canonicalize,
convert-arith-to-llvm{index-bitwidth=64},finalize-
memref-to-llvm{index-bitwidth=64 use-opaque-
pointers=false},convert-scf-to-cf,convert-cf-to-llvm{
index-bitwidth=64},finalize-memref-to-llvm{use-
opaque-pointers=false}, gpu.module(convert-gpu-to-
nvvm,reconcile-unrealized-casts,canonicalize,gpu-to-
cubin),fold-memref-alias-ops,lower-affine,gpu-to-
llvm{use-opaque-pointers=false},finalize-memref-to-
llvm{index-bitwidth=64 use-opaque-pointers=false},
reconcile-unrealized-casts)" stencil.mlir
```

Listing 4 Command issued to lower transformed stencil IR to GPU via *mlir-opt*

Once the stencil has been lowered to the LLVM MLIR dialect, this is then transformed into LLVM-IR, compiled into an object file

by Clang. The isolated FIR IR module is compiled by Flang into an object file and these are then linked together into an executable. In the manner described in this section we have intercepted the FIR-based IR representation of a Fortran code, transformed applicable parts into stencils and then applied specialist optimisations and transformations on these.

4 RESULTS AND EVALUATION

4.1 Experimental environment

In this section we use two stencil-based codes as benchmarks, the first solving LaPlace’s equation for diffusion in three dimensions via a Gauss Seidel solver. An iterative algorithm containing an outer loop operating across the entire domain and generating a progressively improving sequence approximate solutions, for each grid cell at each iteration this benchmark works with a 7-point stencil, the orthogonal neighbours in three dimensions, and averages values across the six neighbouring cells. Consequently, there are six floating point operations required per grid cell.

The second benchmark is the Piacsek and Williams advection scheme [18], commonly used by Met Office codes such as the MONC high-resolution atmospheric model [3], for calculating the movement of quantities through the atmosphere due to kinetic effects (i.e. wind). This is more complex than the first benchmark, containing three separate stencil computations across three fields which are then fused by our stencil transformation into a single stencil region. There are 63 floating point operations required per grid cell.

CPU based benchmarks conducted in this paper have been run on ARCHER2, a Cray-EX which is the UK national supercomputer. Each node contains two 64-core AMD EPYC 7742 (Rome, Zen 2) CPUs with 256 GB of memory. There are 16 cores per NUMA region, providing a total of 8 NUMA regions per node. Nodes are interconnected via HPE Cray Slingshot, delivering two 100 Gbps bi-directional links per node.

GPU based benchmarks are run on Cirrus, an HPE/SGI 8600 HPC system, where the GPU nodes provide Nvidia Tesla V100-SXM2-16GB (Volta) GPUs and two 20-core Intel Xeon Gold 6248 (Skylake) CPUs with 384 GB of memory. Nodes are interconnected via an FDR single infiniband (IB) fabric.

All experiments are averaged over five runs, and we use version 15 of the Cray Compilation Environment (CCE), version 16 of LLVM, version 0.14.0 of xDSL, and version 22.11 of the Nvidia HPC SDK. Performance is reported in terms of throughput, using the metric millions of grid cells per second (MCells/s). Because the Gauss Seidel benchmark is simpler, requiring only 6 floating point operations per grid cell compared to 63 FP operations for the PW advection benchmark, this will naturally deliver the higher throughput of the two benchmarks and therefore a throughput comparisons between benchmarks is less interesting than comparisons for different sizes and configurations of each individual benchmark.

4.2 Single node CPU performance

Figure 2 reports performance, as throughput, achieved on a single core of ARCHER2 for our two benchmarks at different problem sizes. We are comparing approaches leveraging the Cray compiler

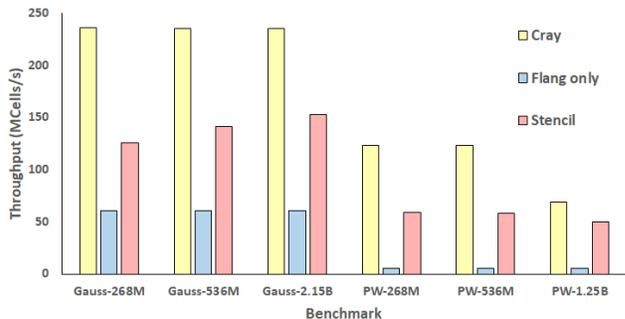


Figure 2: Single core performance comparison for Gauss Seidel and PW advection benchmarks using different problem sizes across Cray, Flang only and our stencil approach.

(Cray in Figure 2), Flang on its own (*Flang only* in Figure 2), and our stencil approach (*Stencil* in Figure 2) described in Section 3. It can be seen that the Cray compiler provides excellent performance, which is inline with previous community experience, and using Flang directly delivers the lowest performance. Our stencil approach achieves performance that lies between the Cray compiler and Flang, and when profiling we found that the executable generated by the Cray compiler undertakes considerably more vectorisation than our stencil approach, even though we use the *scf-parallel-loop-specialization* transformation pass in our CPU flow. It can be seen clearly however, that by leveraging our stencil optimisation the programmer is able to obtain significantly higher performance than Flang alone, especially for the PW advection benchmark.

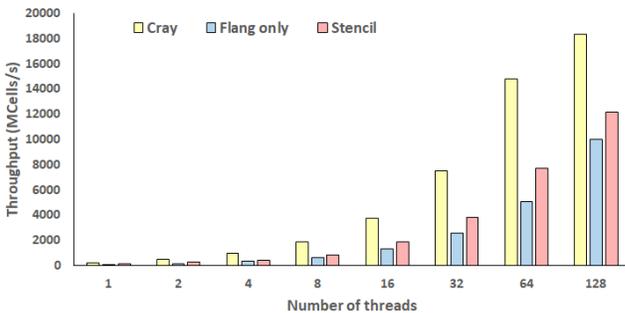


Figure 3: Multithreaded performance comparison for Gauss Seidel benchmark using OpenMP with a problem size of 2.1 billion grid cells across Cray, Flang only and our stencil approach.

Figures 3 and 4 report multithreaded performance for the Gauss Seidel and PW advection benchmarks respectively when run on the largest problem size of 2.1 billion grid cells. Similarly to single core performance, it can be seen that generally the Cray compiler delivers the best performance, with vanilla Flang the lowest and our stencil approach between these two. However, it can be seen

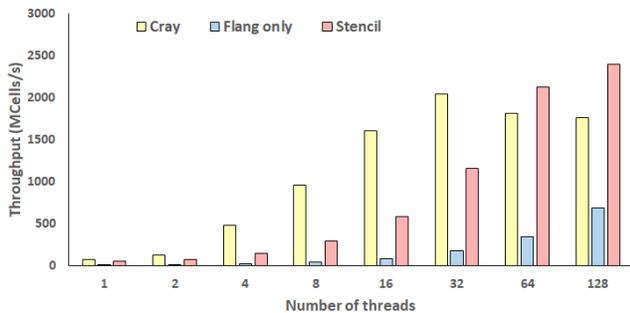


Figure 4: Multithreaded performance comparison for PW advection benchmark using OpenMP with a problem size of 2.1 billion grid cells across Cray, Flang only and our stencil approach.

in Figure 4 that for the PW advection benchmark our stencil approach delivers the highest performance at 64 and 128 thread. Furthermore, it should be highlighted that the Cray and vanilla Flang multithreaded experiments are being executed with hand written OpenMP code, requiring the programmer to modify their code to run multithreaded. By comparison, our stencil approach is leveraging unchanged serial code with the OpenMP parallelism added automatically by the compiler.

4.3 GPU performance

Figure 5 reports the log scale performance of our two benchmarks, at different problem sizes, running on a Nvidia V100 GPU. For each configuration there are three numbers reported, firstly the performance obtained from a manual porting of the code to the V100 GPU using OpenACC and compiled with the Nvidia compiler (*OpenACC with Nvidia* in Figure 5). There are two results for our stencil approach where the initial approach we adopted, (*Stencil (initial data approach)* in Figure 5), was for the GPU dialect to manage all data movement via the *gpu.host_register* operation on all stencil data arrays. However, we found that this delivered very poor performance and when profiling discovered it was due to excessive movement of data between the host and GPU over PCI express. Effectively, this *gpu.host_register* operation is allocating data on the host and moves it across on demand, without effective caching which was causing significant runtime overhead.

Consequently, we developed our own approach, (*Stencil (optimised data approach)* in Figure 5), to managing the memory by developing a bespoke transformation pass. This walks the SSA-based IR just after the stencil extraction pass described in Section 3, and identifies what data must be placed on the GPU and when. Additional functions are added into the extracted stencil module to call operations in the GPU dialect for data allocation, movement, and deallocation and these are called from FIR using the same approach as when calling stencil execution. The FIR IR holds references to the GPU allocated data as FIR LLVM pointers which are provided to the stencil execution functions as arguments and can also be used, in combination with the FIR data reference, to copy data between the host and device. It was found that this transformation was highly effective and by undertaking this direct management of

data via a transformation pass we were able to achieve significant performance as can be seen in the log scale results of Figure 5.

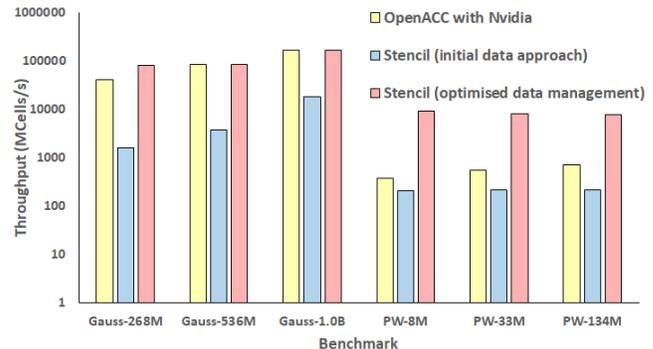


Figure 5: Log scale GPU performance comparison on Nvidia V100 GPU for both benchmarks at different problem sizes. Comparing OpenACC compiled with the Nvidia compiler against our stencil approach. Stencil approach reports figures for both our initial and optimised data management strategies.

It can be seen in Figure 5 that the data optimised version of our stencil approach is very competitive against the hand written OpenACC compiled with the Nvidia compiler. For the Gauss Seidel benchmark our approach outperforms the Nvidia compiler with hand written OpenACC for the smallest problem size and is comparable for the two larger problem sizes. Whilst there is considerably more work per grid cell, and hence a lower throughput in terms of grid cells processed per second, our optimised approach outperforms the manually written OpenACC for all sizes with the PW advection benchmark. When profiling the OpenACC code we found that, due to the use of unified memory, there were numerous data access stalls. The overhead is far less than that of the MLIR managed data approach, but it does still add considerable overhead. Whilst it would likely be possible to obtain increased performance with the manual code by controlling all the data movement directly in code, this would further increase the complexity. By contrast, in our approach the programmer’s Fortran source code is unchanged to run on the GPU.

4.4 Distributed memory performance

As described in Section 3, there is also a lowering in xDSL from the stencil dialect to MPI via an intermediate Distributed Memory Parallelism (DMP) dialect. Whilst this is somewhat experimental, it is worth exploring the performance that we can obtain when running on an unmodified Fortran code compared to a hand-crafted MPI parallelisation to explore the potential for automatic distributed memory parallelisation of serial code. Of our two benchmarks, we selected Gauss Seidel only for this experiment because it operates in iterations, requiring a halo swap between iterations. The PW advection benchmark by comparison is a kernel called from a larger code base which then undertakes halo swapping before the next timestep when other calculations have completed.

Figure 6 reports the distributed memory performance of the Gauss Seidel benchmark running on ARCHER2, where there are

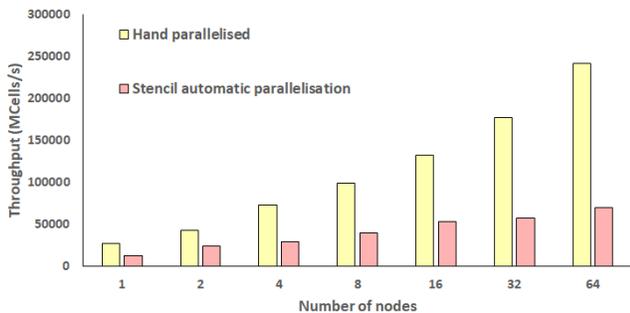


Figure 6: Distributed memory performance of Gauss Seidel benchmark across nodes of ARCHER2, using one MPI process per core (128 cores per node). Global problem size of 17 billion grid cells comparing a hand-parallelised version compiled with the Cray compiler to the distributed memory version auto-generated from our flow.

128 cores per node and we are mapping one MPI process to each core. We decompose the 3D space into two dimensions, and the results compare a hand parallelised version with the Cray compiler, (*Hand parallelised* in Figure 6), against the automatic parallelisation obtained via lowering our stencil dialect IR to the xDSL MPI dialect via the DMP dialect (*Stencil automatic parallelisation* in Figure 6). It can be seen that the hand parallelised version outperforms our automatically parallelised version which is for two reasons. Firstly, as reported for single core runs, the Cray compiler is very good at vectorisation and delivering high single core performance which is relevant for the baseline performance. Secondly, the hand parallelised version scales better than the automatic parallelisation via the DMP dialect.

Whilst this demonstrates that there is still work to be done in the DMP and MPI xDSL dialects to obtain optimal performance, it should be noted that we are still able to take an unmodified Fortran serial code and for it to run across 8192 CPU cores and obtain a throughput of around 70,000 million grid cells per second. We believe that this demonstrates the potential of our approach, even if the individual components still need further enhancement to reach a level of maturity where they can match the performance of hand parallelised code.

5 RELATED WORK

There are several existing frameworks and DSLs which aim to optimise the execution of stencil-based computations. For instance, the Pochoir [20] stencil compiler enables programmers to write their code in C++ and use bespoke templates to drive a higher level description of their stencil computation. This is then provided to the Pochoir compiler which undertakes source to source translation as a preprocessing step, the results of which are then compiled by the Intel compiler. Leveraging Cilk [2], which provides multithreading parallelism capabilities, their source to source translation tool generates Cilk based code from the programmer’s source code.

ExaStencils [15] is a stencil-based framework, and this supports a wider range of execution targets than Pochoir. Similar to our approach, ExaStencils supports shared and distributed memory parallelism as well as execution on GPUs. However, similarly to Pochoir users must learn new abstractions and explicitly port their code to this technology, whereas our approach leverages unmodified serial Fortran code. The major disadvantage with both Pochoir and ExaStencils is the siloed nature of their compilation stacks, where these have been developed as bespoke compilers and share no underlying infrastructure with other projects. This is especially important given the effort undertaken to optimise stencil execution by these projects, which not only requires significant investments of time to initially develop, but then a continuing maintenance burden to support new architectures and fix bugs. Adopting DSLs like these results in a risk for users, as they can not be sure about the long term future of these technologies, especially as many are developed during research projects with a fixed end date.

By contrast, our approach leverages the popular LLVM/MLIR ecosystem which is actively developed and maintained by a large community including individuals, academic organisations, and commercial companies. Furthermore, the bespoke nature of our contribution is a small part in the overall compilation flow where, as described in Section 3, we provide stencil discovery and extraction transformations for FIR, with the generated SSA-based IR then transformed by passes developed by other groups, many of which are in the main MLIR codebase. Consequently, programmers can have much more confidence leveraging an approach that sits within the MLIR ecosystem because of the significant investment that has, and is being made into this project.

LLVM is used extensively in HPC as a growing number of compilers from vendors are now built upon the framework. Prominent HPC compiler teams, such as those at Cray, Intel, ARM, AMD and Nvidia have made significant investments in LLVM and MLIR, and many of these organisations produce products build upon the technologies. There are additional community activities exploring optimisation opportunities provided by LLVM and MLIR, for instance [7] which explored the use of MLIR to optimise the execution of stencil-based codes. Developing their own *cfv* dialect which contained a stencil operation, although this is significantly more limited than the Open Earth Compiler’s stencil dialect used in this work, the authors were able to demonstrate that the MLIR ecosystem is beneficial for optimising these calculations in a multi-threaded environment. However they did not integrate their work with existing, general purpose, MLIR compilers or attempt to undertake the stencil discovery and extraction from code as explored in this paper. Furthermore, by only considering multi-threaded CPU execution of their dialect they limit the benefits to be gained from the MLIR ecosystem such as GPU execution.

Lastly, DaCE [23] is a parallel programming framework that operates over codes typically written in Python and converts these into a dataflow Directed Acyclical Graph (DAG) IR representation. This is then used by DaCE to generate optimal code for a variety of targets including shared and distributed memory CPUs and GPUs. Providing interoperability with MLIR via the DaCE dialect [1], the ability to move between DaCE dataflow and MLIR control flow representations can deliver improved executable performance. Furthermore, Stencilflow [5] provides the ability to translate stencils

into the DaCE dataflow DAG IR with the primary objective of enabling these to be efficiently executed on FPGAs. However, at the time of writing, Stencilflow is not actively maintained, and furthermore does not undertake the automated discovery and extraction of stencils from existing code as described in this paper. It would also likely be impractical to integrate DaCE with MLIR-based compilers, such as Flang, due to its extensive nature and the additional maintenance dependency. By contrast, our approach is far more lightweight and could be integrated by the addition of our two additional compiler passes into these tools.

6 CONCLUSIONS AND FURTHER WORK

In this paper we have explored the potential to enrich general purpose compiler flows by leveraging domain specific information via MLIR dialects and transformations. By discovering and extracting stencils from the FIR intermediate representation generated by Flang, we have demonstrated that it is possible to then exploit this domain specific information and transformations to target numerous architectures and forms of parallelism. Flang provides an unusual approach to leveraging MLIR, by translating the FIR dialect directly into LLVM-IR rather than into the existing standard MLIR dialects which themselves are lowered. This does decrease composability between dialects, but as demonstrated in this paper can be worked around by extracting dialects that are not registered with Flang into separate modules and compiling independently.

Our main objective was to explore whether a fusion of domain specific abstraction with the general purpose Flang compiler would enable us to deliver improved performance compared with Flang alone and, using two benchmarks, we demonstrated that our approach delivers around a two time speed up for the iterative Gauss Seidel solver and approximately a 10 times speed up for the atmospheric PW advection scheme on a single CPU compared to Flang. We found that the Cray compiler delivers impressive performance, considerably outperforming Flang and our stencil approach. However, it should be borne in mind that the Cray compiler is a mature product and the result of many years of development, furthermore it is not available for general release and can only be found on Cray supercomputers. Nevertheless, our stencil approach did outperform the Cray compiler for the PW advection benchmark when run shared memory at 64 and 128 threads. Consequently we have demonstrated that to help close the performance gap between Flang and compilers such as the Cray compiler, then leveraging other existing parts of the MLIR ecosystem can be beneficial.

On a V100 GPU our stencil flow, automatically porting the Fortran code to GPUs from the programmer’s perspective, delivered similar performance to a hand written OpenACC implementation for the Gauss Seidel benchmark, and our approach was on average approximately 15 times faster than the hand written OpenACC for the PW advection benchmark. Furthermore, we explored lowering our stencils via the existing xDSL distributed memory and MPI dialects on up to 8192 cores of ARCHER2. Whilst hand parallelised distributed memory code did perform and scale better than the automatic parallelised version, the fact that we were able to scale to 8192 cores is encouraging given that the same, unchanged, Fortran source code was used for this experiment as for the single CPU, multi-threaded CPU, and GPU experiments.

There are five avenues that would be interesting to explore as further work. Firstly, single CPU performance optimisations in the scf dialect lowering, especially targeting improved vectorisation, would be worthwhile. Secondly, the xDSL DMP and MPI dialects are currently in active development and the results of this paper demonstrate that whilst they are currently usable they would benefit from additional optimisation and tuning to match the performance of existing hand crafted codes. Thirdly, we believe that it would be useful to pursue combining the stencil optimisation reported in this paper with general purpose MLIR compilers so that they can benefit from this optimisation. As this work has been developed with Fortran as the focus, then Flang integration would likely be the easiest to achieve, although we would need to consider whether to integrate with FIR or wait for the new HL-FIR to be fully matured. However our central algorithm and associated transformations could also be adapted to other languages and benefit a wider set of MLIR based compilers such as Polygeist and Pylir. Fourthly, we believe that it would be worth exploring the potential of lowering FIR into the standard MLIR dialects rather than directly to LLVM-IR. This could reduce the maintenance burden as lowering to LLVM-IR from the standard dialects would exploit more shared passes, and furthermore would also aid in bringing additional dialects into the Flang ecosystem. Lastly, combining distributed memory parallelism with GPU execution, enabling multi-node GPU execution, potentially in combination with a communication technology such as NVLink [11], would be worthwhile.

We conclude that this study highlights that there is a significant potential for MLIR to enrich our existing programming languages and development, not only by leveraging domain specific knowledge to deliver improved performance, but to also potentially support delivering new capabilities for codes such as automatic parallelisation and architecture portability. Using the existing building blocks of MLIR, we have found that it can be very effective to develop bespoke components and then integrate these with the existing ecosystem.

ACKNOWLEDGMENTS

This work has been funded by the xDSL ExCALIBUR EPSRC project. This work used the ARCHER2 UK National Supercomputing Service (<https://www.archer2.ac.uk>). This work used the Cirrus UK National Tier-2 HPC Service at EPCC (<http://www.cirrus.ac.uk>) funded by the University of Edinburgh and EPSRC (EP/P020267/1)

REFERENCES

- [1] Tal Ben-Nun, Berke Ates, Alexandru Calotoiu, and Torsten Hoefer. 2023. Bridging control-centric and data-centric optimization. In *Proceedings of the 21st ACM/IEEE International Symposium on Code Generation and Optimization*. 173–185.
- [2] Robert D Blumofe, Christopher F Joerg, Bradley C Kuszmaul, Charles E Leiserson, Keith H Randall, and Yuli Zhou. 1995. Cilk: An efficient multithreaded runtime system. *ACM SigPlan Notices* 30, 8 (1995), 207–216.
- [3] Nick Brown, Michele Weiland, Adrian Hill, Ben Shipway, Chris Maynard, Thomas Allen, and Mike Rezny. 2020. A highly scalable Met Office NERC Cloud model. *arXiv preprint arXiv:2009.12849* (2020).
- [4] Valentin Clement, Sylvaine Ferrachat, Oliver Fuhrer, Xavier Lapillonne, Carlos E Osuna, Robert Pincus, Jon Rood, and William Sawyer. 2018. The CLAW DSL: Abstractions for performance portable weather and climate models. In *Proceedings of the Platform for Advanced Scientific Computing Conference*. 1–10.
- [5] Johannes de Fine Licht, Andreas Kuster, Tiziano De Matteis, Tal Ben-Nun, Dominic Hofer, and Torsten Hoefer. 2021. StencilFlow: Mapping large stencil programs to distributed spatial computing systems. In *2021 IEEE/ACM International*

- Symposium on Code Generation and Optimization (CGO)*. IEEE, 315–326.
- [6] Raúl de la Cruz and Mauricio Araya-Polo. 2015. Modeling stencil computations on modern HPC architectures. In *High Performance Computing Systems. Performance Modeling, Benchmarking, and Simulation: 5th International Workshop, PMBS 2014, New Orleans, LA, USA, November 16, 2014. Revised Selected Papers 5*. Springer, 149–171.
- [7] Mohamed Essadki, Bertrand Michel, Bruno Maugars, Oleksandr Zinenko, Nicolas Vasilache, and Albert Cohen. 2023. Code Generation for In-Place Stencils. In *Proceedings of the 21st ACM/IEEE International Symposium on Code Generation and Optimization*. 2–13.
- [8] Mathieu Fehr, Jeff Niu, River Riddle, Mehdi Amini, Zhendong Su, and Tobias Grosser. 2022. IRDL: an IR definition language for SSA compilers. In *Proceedings of the 43rd ACM SIGPLAN International Conference on Programming Language Design and Implementation*. 199–212.
- [9] FIR. 2023. *FIR Language Reference*. Retrieved Aug 16, 2023 from <https://flang.llvm.org/docs/FIRLangRef.html>
- [10] Flang. 2023. *Flang Documentation*. Retrieved Aug 16, 2023 from <https://flang.llvm.org/docs/>
- [11] Denis Foley and John Danskin. 2017. Ultra-performance Pascal GPU and NVLink interconnect. *IEEE Micro* 37, 2 (2017), 7–17.
- [12] Tobias Gysi, Christoph Müller, Oleksandr Zinenko, Stephan Herhut, Eddie Davis, Tobias Wicky, Oliver Fuhrer, Torsten Hoefler, and Tobias Grosser. 2021. Domain-specific multi-level IR rewriting for GPU: The Open Earth compiler for GPU-accelerated climate simulation. *ACM Transactions on Architecture and Code Optimization (TACO)* 18, 4 (2021), 1–23.
- [13] Michael Lange, Navjot Kukreja, Mathias Louboutin, Fabio Luporini, Felipe Vieira, Vincenzo Pandolfo, Paulius Velesko, Paulius Kazakas, and Gerard Gorman. 2016. Devito: Towards a generic finite difference dsl using symbolic python. In *2016 6th workshop on python for high-performance and scientific computing (PyHPC)*. IEEE, 67–75.
- [14] Chris Lattner, Mehdi Amini, Uday Bondhugula, Albert Cohen, Andy Davis, Jacques Pienaar, River Riddle, Tatiana Shpeisman, Nicolas Vasilache, and Oleksandr Zinenko. 2021. MLIR: Scaling compiler infrastructure for domain specific computation. In *2021 IEEE/ACM International Symposium on Code Generation and Optimization (CGO)*. IEEE, 2–14.
- [15] Christian Lengauer, Sven Apel, Matthias Bolten, Shigeru Chiba, Ulrich Rüde, Jürgen Teich, Armin Größlinger, Frank Hannig, Harald Köstler, Lisa Claus, et al. 2020. Exastencils: advanced multigrid solver generation. In *Software for Exascale Computing-SPEXA 2016-2019*. Springer International Publishing, 405–452.
- [16] William S Moses, Lorenzo Chelini, Ruizhe Zhao, and Oleksandr Zinenko. 2021. Polygeist: Raising C to polyhedral MLIR. In *2021 30th International Conference on Parallel Architectures and Compilation Techniques (PACT)*. IEEE, 45–59.
- [17] Gihan R Mudalige, IZ Reguly, Satya P Jammy, Christian T Jacobs, Michael B Giles, and Neil D Sandham. 2019. Large-scale performance of a DSL-based multi-block structured-mesh application for Direct Numerical Simulation. *J. Parallel and Distrib. Comput.* 131 (2019), 130–146.
- [18] Steve A Piacsek and Gareth P Williams. 1970. Conservation properties of convection difference schemes. *J. Comput. Phys.* 6, 3 (1970), 392–405.
- [19] Pylir. 2023. *Pylir Documentation*. Retrieved Aug 16, 2023 from <https://zero9178.github.io/Pylir/>
- [20] Yuan Tang, Rezaul Alam Chowdhury, Bradley C Kuszmaul, Chi-Keung Luk, and Charles E Leiserson. 2011. The pochoir stencil compiler. In *Proceedings of the twenty-third annual ACM symposium on Parallelism in algorithms and architectures*. 117–128.
- [21] Mariano Vazquez, Guillaume Houzeaux, Seid Koric, Antoni Artigues, Jazmin Aguado-Sierra, Ruth Aris, Daniel Mira, Hadrien Calmet, Fernando Cucchietti, Herbert Owen, et al. 2014. Alya: towards exascale for engineering simulation codes. *arXiv preprint arXiv:1404.4881* (2014).
- [22] xDSL. 2023. *A Python Compiler Design Toolkit*. Retrieved Aug 16, 2023 from <https://github.com/xdslproject/xdsl>
- [23] Alexandros Nikolaos Ziogas, Timo Schneider, Tal Ben-Nun, Alexandru Calotoiu, Tiziano De Matteis, Johannes de Fine Licht, Luca Lavarini, and Torsten Hoefler. 2021. Productivity, portability, performance: Data-centric Python. In *Proceedings of the International Conference for High Performance Computing, Networking, Storage and Analysis*. 1–13.