ABSTRACT
Python has become the de facto language for scientific computing. Programming in Python is highly productive, mainly due to its rich science-oriented software ecosystem built around the NumPy module. As a result, the demand for Python support in High Performance Computing (HPC) has skyrocketed. However, the Python language itself does not necessarily offer high performance. In this work, we present a workflow that retains Python’s high productivity while achieving portable performance across different architectures. The workflow’s key features are HPC-oriented language extensions and a set of automatic optimizations powered by a data-centric intermediate representation. We show performance results and scaling across CPU, GPU, FPGA, and the Piz Daint supercomputer (up to 23,328 cores), with 2.47x and 3.75x speedups over previous-best solutions, first-ever Xilinx and Intel FPGA results of annotated Python, and up to 93.16% scaling efficiency on 512 nodes.

CCS CONCEPTS
• Software and its engineering → Parallel programming languages; Distributed programming languages; Data flow languages; Source code generation.

KEYWORDS
Data-Centric, High Performance Computing, Python, NumPy

1 INTRODUCTION
Python is the language to write scientific code [30]. The capability to write and maintain Python code with ease, coupled with a vast number of domain-specific frameworks and libraries such as SciPy, Matplotlib, scikit-learn [62], or pandas [74], leads to high productivity. It also promotes collaboration with reproducible scientific workflows shared using Jupyter notebooks [42]. Therefore, numerous scientific fields, ranging from machine learning [2, 61] to climate [72] and quantum transport [75] have already adopted Python as their language of choice for new developments.
We propose a way to bridge the gap between the three Ps for Python programming using a data-centric paradigm. In particular, we empower Python users with an automatic optimization and specialization toolbox, which spans the entire Python/HPC ecosystem (Fig. 1)—from the code, through communication distribution, to hardware mapping. At the core of the toolbox, we use the Stateful Dataflow multiGraphs (SDFG) [13] data-centric intermediate representation, which enables these optimizations in the form of multi-level data movement transformations. With a data-centric representation, as opposed to library bindings, all data dependencies and potential overlap are inferred statically from the code, and interpreter overhead is mitigated. Compared with implicit and lazy evaluation approaches, we also provide a set of extensions that give power users complete control over parallelism and partitioning schemes, using pythonic principles and syntax (e.g., returning “local view” objects from global data, but allowing users to operate on the “global view” as well).

We demonstrate a wide variety of benchmarks using the automatic toolbox over annotated Python code, both on individual nodes and the Piz Daint supercomputer. For the former, we show that it consistently outperforms other automatic approaches on multicore CPUs and GPUs, and for the first time show automatic Python HPC compilation results for both Xilinx and Intel FPGAs, which vastly differ in architecture and partitioning language. In distributed-memory environments, we show high scaling efficiency and absolute performance compared with distributed tasking. Thus, we realize all three Ps within a single system.

The paper makes the following contributions:

- **(Productivity)** Definition of high-performance Python, a methodology to translate it to a data-centric IR, and extensions to improve said conversion via explicit annotation.
- **(Portability)** A set of automatic optimizations for CPU, GPU and FPGA, outperforming the best prior approaches by 2.47× on CPU and 3.75× on GPU on average (geometric mean [1]).
- **(Performance)** Automatic implicit MPI transformations and communication optimizations, as well as explicit distribution management, with the former scaling to 512 nodes with up to 93.16% efficiency.

## 2 DATA-CENTRIC PYTHON

The central tenet of our approach is that understanding and optimizing data movement is the key to portable, high-performance code. In a data-centric programming paradigm, three governing principles guide development and execution:

1. Data containers must be separate from computations.
2. Data movement must be explicit, both from data containers to computations and to other data containers.
3. Control flow dependencies must be minimized, they shall only define execution order if no implicit dataflow is given.

In the context of SDFGs, examples of data containers are arrays and scalar data, which have a NumPy-compatible data type, such as int32 or float64.

Python is an imperative language and, therefore, not designed to express data movement. Its terseness makes the process of understanding dataflow difficult, even when comparing to other languages like C and FORTRAN, as the types of variables in Python code cannot be statically deduced.

Below, we define high-performance Python programs, discuss the decorators that we must add to Python code to make the dataflow analyzable, and then detail how we translate them into the SDFG data-centric intermediate representation.

### 2.1 High Performance Python

Our approach supports a large subset of the Python language that is important for HPC applications. The focus lies on NumPy arrays [36] and operations on such arrays. In addition to the low-overhead data structures NumPy offers, it is central to many frameworks focused on scientific computing, e.g., SciPy, pandas, Matplotlib. As opposed to lazy evaluation approaches, high-performance Python must take control flow into account to auto-parallelize and avoid interpreter overhead. This tradeoff between performance and productivity is necessary because Python features such as co-routines are not statically analyzable and have to be parsed as “black-boxes”. To combat some of these Python quirks, we propose to augment the language with analyzable constructs useful for HPC.

### 2.2 Annotating Python

The Data-Centric (DaCe) Python frontend parses Python code and converts it to SDFGs on a per-function basis. The frontend will parse only the Python functions that have been annotated explicitly by the user with the @dace.program decorator. DaCe programs can then be called like any Python function and perform Just-in-Time (JIT) compilation.

**Static symbolic typing.** To enable Ahead-of-Time (AOT) compilation, which is of key importance for FPGAs and for reusing programs across different inputs, SDFGs should be statically typed. Therefore, the function argument data types are given as type annotations, providing the required information as shown below:

```python
N = dace.symbol()
@dace.program
def jacobi_1d(TSTEPS: dace.int32,
            A: dace.float64[N],
            B: dace.float64[N]):
    for t in range(1, TSTEPS):
```

The Python method jacobi_1d has three arguments; TSTEPS is a 32-bit integer scalar, while A and B are double precision floating-point vectors of length N. The symbolic size N, defined with `dace.symbol`, indicates that the vector sizes can be dynamic (but equal). All subsets are then symbolically defined (e.g., the subset B[1:-1] becomes B[1:N-1], and symbolic manipulation can then be performed in subsequent data-centric transformations.

**Parametric parallelism.** An important feature that has no direct expression in Python is a loop that can run in parallel. Our approach
We turn to present the SDFG intermediate representation (IR) and a which can be used in Python code as a substitute to the Python LoopToMap

Alternatively, the DaCe framework provides a mapping of Python syntax and constructs to SDFG.

From Python to DaCe

We turn to present the SDFG intermediate representation (IR) and a novel data-centric Python translation procedure in tandem. While previous work [13] converted a restricted, low-level Python definition of the SDFG IR, here we aim to cover the majority of the Python/NumPy language constructs via static analysis and fallback for unsupported features. We summarize the equivalence between Python constructs and SDFG counterparts in Table 1, and present the generation of an SDFG from a Python program using the gemm kernel as an example:

```
@dace.program
def gemm(alpha, beta, C, A, B):
    C[:] = alpha * A + beta * C
```

<table>
<thead>
<tr>
<th>Python</th>
<th>SDFG Equivalent</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Declarations and Types</strong></td>
<td></td>
</tr>
<tr>
<td>Primitive data types</td>
<td>Scalar data container</td>
</tr>
<tr>
<td>NumPy array</td>
<td>Array data container</td>
</tr>
<tr>
<td><strong>Assignments</strong></td>
<td></td>
</tr>
<tr>
<td>Assignments</td>
<td>Tasklet or map scope with incoming and outgoing memlets for read/written operands</td>
</tr>
<tr>
<td>Array subscript</td>
<td>Memlet (dataflow edge)</td>
</tr>
<tr>
<td><strong>Statements</strong></td>
<td></td>
</tr>
<tr>
<td>Branching (if)</td>
<td>Branch conditions on state transition edges</td>
</tr>
<tr>
<td>Iteration (for)</td>
<td>Conditions, increments on state transition edges</td>
</tr>
<tr>
<td>Control-flow (break, continue, return)</td>
<td>Edge to control structure or function exit state</td>
</tr>
<tr>
<td><strong>Functions</strong></td>
<td></td>
</tr>
<tr>
<td>Function calls (with source) and decorator for argument types</td>
<td>Nested SDFG for content, memlets reduce shape of inputs and outputs</td>
</tr>
<tr>
<td>External/Library calls</td>
<td>Tasklet with callback or Library Node</td>
</tr>
</tbody>
</table>

Table 1: Mapping of Python syntax and constructs to SDFG.

The first step in the above code multiplies each element of A with alpha. SDFGs view data containers separately from the computations the data are part of, as per the first data-centric tenet. These containers are represented by oval Access nodes. In the first statement, these refer to tmp0, alpha, and A (see Fig. 2a).

In SDFGs, connections to data containers are called memlets, and they describe the data movement — the edge direction indicates whether it is read or written, and its contents refer to the part of the data container that is accessed. Computations consume/produce memlets and can be divided into multiple types:

1. Stateless computations (Tasklets, shown as octagons), e.g., representing scalar assignments such as a = 1.
2. Calls to external libraries (Library Nodes, folded rectangles), that represent calls to functions that are not in the list of functions decorated with @dace.program. Matrix-matrix multiply is a common and important operation tmp1 = tmp0 @ B and is contained in a Library Node called MatMul.
3. Calls to other SDFGs (Nested SDFGs, rectangles), which represent calls to functions decorated with @dace.program.
4. Maps are a particular type of Nested SDFGs matching the language augmentation previously discussed in Section 2.2 and express that the content can be processed in parallel.

Element-wise array operations automatically yield Map scopes. Similarly, assignments to arrays yield a Map scope, containing a Tasklet with the element-wise assignment. Augmented assignments such as C += 1 are a special case where the output is also an input when no data races are detected.

Parallel maps can be augmented to express how Write-Conflict Resolution (WCR) should determine the value of data when multiple sources write to it concurrently. If data races are found, the outgoing edges are marked as dashed. E.g., the following program requires WCR (SDFG representation is shown in Fig. 2b):
By connecting the inputs and outputs of computations with the containers explicitly, the data-centric representation of a program can be created. To represent control dependencies, we encapsulate code driven by pure data dependencies in States (bright blue regions) in the SDFG. The states are connected with State Transition Edges (in blue) that can be annotated with control flow, representing loops, branch conditions, and state machines in general. The following Python program is represented by the SDFG in Fig. 3:

```python
for i in range(NI):
    C[i] += 1
```

![Figure 3: SDFG representation of a Python for-loop. There are two states (left: guard, right: body) connected by control flow (state transition) edges that define the loop range.
](image)

The above loop also is an excellent use case for the LoopToMap transformation (Section 2.2) since its iterations are independent.

In the conversion to the SDFG IR, we also replace calls to library functions (e.g., np.linalg.solve and object methods (a.view())) with custom subgraphs or Library Nodes, which users can extend for other libraries and object types via a decorated function. Following this initial conversion, the resulting SDFG contains a state per statement and a nested SDFG per function call.

### 2.4 Dataflow Optimization

The direct translation to SDFGs creates a control-centric version of the code, which matches the Python semantics but does not contain dataflow beyond a single statement (similarly to compilers’ -O0). To mitigate this, we run a pass of IR graph transformations that coarsens the dataflow, exposing a true data-centric view of the code (similar to -O1). The transformations include redundant copy removals, inlining Nested SDFGs, and others (14 in total), which only modify or remove elements in the graph, such that they cannot be applied indefinitely.

To understand this pass, we showcase one transformation — state fusion — which allows merging two states where the result does not produce data races. For example, the two states containing the assignments below can be merged:

```
    tmp0 = alpha * A
    tmp1 = tmp0 @ B
```

Internally, the transformation matches a subgraph pattern of two states connected together and compares source and sink Access nodes using symbolic set intersection. If no data dependency constraints are violated, Access nodes are either fused (if they point to the same memory, see Fig. 4) or set side by side, creating multiple connected components that can run in parallel.

All transformations in the DaCe framework follow the same infrastructure, either matching a subgraph pattern or allowing the user to choose an arbitrary subgraph [13]. While the dataflow coarsening pass happens automatically as part of our proposed toolbox, one can also apply transformations manually and separately, without changing the original Python source code (we color such “performance engineering codes” in cyan):

```python
sdfg = gemm.to_sdfg()
sdfg.apply(StateFusion)
```

![Figure 4: State fusion of tmp0 = alpha * A and tmp1 = tmp0 @ B.
](image)

### 2.5 Python Restrictions

Some features available in Python are incompatible with our definition of high performance Python, and are discussed below. This does not exclude programs using the full feature set of Python from analysis, but calls to functions containing unsupported features will not benefit from our optimization. The restricted features are:

1. Python containers (lists, sets, dictionaries, etc.) other than NumPy arrays as arguments, as they are represented by linked lists and difficult to analyze from a dataflow view. Note that this does not preclude internal containers (which can be analyzed) and list comprehensions.
2. Dynamic typing: fixed structs are allowed in SDFGs, so fields can be transformed and their class methods into functions with the `dace.program` decorator. However, dynamic changes to classes or dynamic types are unsupported as their structure cannot be statically derived.
3. Control-dependent variable state (i.e., no looping rules), e.g., the following valid Python:

   ```python
   x = ...
   if x > 5:
       y = np.ndarray([5, 6], dtype=np.float32)
   # use y (will raise exception if x <= 5)
   ```

4. Recursion. This is a limitation of the data-centric programming model [13], as recursion is a control-centric concept that is not portable across platforms.

After performing the full translation and coarsening, the resulting `gemm` kernel can be seen in Figure 5 (left-hand side). This data-centric representation can now use the SDFG IR capabilities to further optimize and map the original Python code to different architectures and distributed systems.
3 PORTABILITY AND PERFORMANCE

The translated data-centric Python programs can now be optimized for performance on different hardware architectures. In this section, we propose a novel set of data-centric passes to auto-optimize and specialize SDFGs to run at state-of-the-art performance on CPUs, GPUs, and FPGAs, all from the same source IR. The programming portability is very high since our approach starts from the same Python codes parsed to the same SDFGs. Although the same Python codes are used, the optimized IRs may differ, the process can be automatic, and the user needs only to select the architecture to specialize for. In our evaluation, we only discuss results produced in an automated fashion.

3.1 Automatic Optimization Heuristics

As mentioned above, DaCe provides a user-extensible set of graph transformations. Yet, the framework does not perform them automatically [13], to endow performance engineers with fine-grained control and promote separation of concerns. Furthermore, DaCe includes tools and graphical interfaces to assist users with manual optimization without the explicit need for an expert. For productivity purposes, however, we believe that prototyping fast data-centric Python programs should be possible with minor code modifications.

By observing the common pitfalls in generated code from SDFGs vs. what a performance engineer would write, we propose a set of transformation heuristics for SDFGs that yield reasonable performance in most cases (≈03 compiler equivalent). This pass can be performed automatically (configurable) or using the following decorator:

```python
@dace.program(auto_optimize=True, device=...)
```

where device can be DeviceType.(CPU,GPU,FPGA).

Our auto-optimizer performs the following passes in order:

1. **Map scope cleanup**: Remove "degenerate" maps of size 1, repeatedly apply the `LoopToMap` transformation (Section 2.2), and collapse nested maps together to form multidimensional maps. The latter also increases the parallelism of GPU kernels as a by-product.

2. **Greedy subgraph fusion**: Collect all the maps in each state, fusing together the largest contiguous subgraphs that share the same (or permuted) iteration space or the largest subset thereof (e.g., fusing the common three dimensions out of four). We use symbolic set checks on memlets to ensure that the data consumed is a subset of the data produced.

3. **Tile WCR maps**: Tile (configurable size) parallel maps with write-conflicts that result in atomics, in order to drastically reduce such operations.

4. **Transient allocation mitigation**: Move constant-sized and small arrays to the stack, and make temporary data containers persistent (i.e., allocated upon SDFG initialization) if their size only depends on input parameters. This nearly eliminates dynamic memory allocation overhead.


Beyond the above general-purpose heuristics, we apply more transformations depending on the chosen device: For CPUs, we try to increase parallelism by introducing the OpenMP `parallel` clause. For GPU and FPGA, we perform the `(GPU,FPGA)` `TransformSDFG` automatic transformations [13], which introduce copies to/from the accelerator and convert maps to accelerated procedures.

On the FPGA, we perform a few further transformations that diverge from the “traditional” fused codes in HPC: we create separate connected components (regions on the circuit) to stream off-chip (DRAM) memory in bursts to the program. Between computations, we try to modify the graph’s structure to be composed of separate pipelined units that stream memory through FIFO queue Access nodes (we call this transformation `StreamingComposition`). This also enables further transformations to the graph to create systolic arrays during hardware specialization.

From this point, the only remaining step to lower the SDFG is to specialize the Library Nodes to their respective fastest implementations based on the target platform.
In the following, we show results for single node shared memory parallel programs created using data-centric Python for CPU, GPU and FPGA, and compare these with other frameworks: NumPy over the CPython interpreter, Numba, Pythran, and CuPy. We collect a set of existing Python codes from different scientific and HPC domains [3, 5, 8, 9, 15, 20, 37, 41, 49, 51, 60, 67, 70–72, 75], as well as a NumPy version of Polybench [63] ported from the C benchmark. In this adaptation, we strive to express the algorithms of the original benchmark in a way that is natural to a Python programmer. E.g., in `gemm`, for the matrix-matrix product is implemented with @, Python’s dedicated operator for matrix multiplication [29].

All the data used are double-precision floating point numbers or 64-bit integers for CPU and GPU, while the FPGA tests use single-precision floating point numbers and 32-bit integers.

3.4 Experimental Setup. The CPU and GPU evaluations are performed on a machine running CentOS 8, with 1.5 TB of main memory, two Intel Xeon Gold 6130 CPUs (2x16 cores), and an NVIDIA V100 GPU (CUDA version 11.1) with 32 GB of RAM. We use CPython 3.8.3 as part of an Anaconda 3 environment. We test NumPy 1.19.2 with Intel MKL support, Numba 0.51.2 with Intel SVML support, the latest Pythran version from their GitHub repository [33] (commit ID 09349c5), and CuPy 8.3.0. For all frameworks that need a separate backend compiler, we use GCC 10.2.0, with all the performance flags suggested by the developers. To put high-performance Python into the perspective of low-level C implementations, we also compare the applications adapted from Polybench with the original Polybench/C [63] benchmark, compiled with GCC and the Intel C Compiler with automatic parallelization enabled (icc –o3 –March=native –mtune=native –parallel).

We evaluate FPGA performance on two different boards from either major FPGA vendor; a Bitware 520N accelerator with an Intel Stratix 10 2800 GX FPGA and the Xilinx Alveo U250 accelerator board. Intel FPGA kernels are built with the Intel OpenCL SDK for FPGA and Quartus 20.3 targeting the p520_max_sg280h shell, and Xilinx kernels are built with the Vitis 2020.2 compiler targeting the xilinx_u250_xdma_201830_2 shell.

For the DaCe Python versions, we annotate types and symbolic shapes on the decorated functions to enable AOT compilation and work with FPGAs. We do not annotate loops as `dace.maps` and keep them in their original form, leaving parallelization for the automatic heuristics (Section 3.1).

We compare the performance of the different frameworks and compilers using runtime as our primary metric of execution. Unless otherwise mentioned, we run each benchmark ten times and report the median runtime and 95% nonparametric confidence interval (CI) [39].

3.4.1 Benchmarking Results. CPU results are presented in Fig. 7. The right-most column contains NumPy’s execution runtime for each of the benchmarks annotated on the y-axis. Each of the other columns contains the speedup (green tint and upward arrow) or slowdown (red tint and downward arrow) of execution compared to NumPy for each of the competing Python frameworks and Polybench C versions (compiled with ICC or GCC). Furthermore, we compute the 95% CI using bootstrapping [27] and annotate its size (as superscript in brackets) as a percentage of the median; values less than 1% are omitted (Fig. 7 uses vector graphics, and readers can zoom in with a PDF viewer if any values are not readable due to their size). The upper part of the chart aggregates the benchmarks.
using the geometric mean of the speedups over NumPy. Numba and Pythran have fallback modes for Python code that fails to parse. In such cases, we measure the runtime of the fallback mode.

The figure shows an overall improvement in DaCe Python’s performance, both over the Python compilers and the optimizing C compilers. Specifically, the subgraph fusion transformation capabilities surpass those of Numba. This is especially apparent in stencils, where the difference can be in orders of magnitude. In applications such as crc16, all compiled implementations successfully eliminate interpreter overhead. Shorter kernels benefit from the C versions due to runtime and timing overhead mitigation. It also appears that with control-flow heavy codes (e.g., nussinov), simple C code can be better optimized by GCC and ICC over generated code. It is worth noting that on some applications, NumPy is faster than the C versions: this is because of the performance benefits of vectorized NumPy code compared with explicit, sequential loops. An exceptional case is 3mm, which consists of three matrix multiplications. ICC pattern matches the matrix-matrix product and links to MKL, achieving similar performance to the Python frameworks. On the other hand, GCC does not compile the unoptimized C code to an executable that uses MKL, leading to lower performance.

Fig. 8 presents the runtime of applications that were successfully transformed to run on GPU. As with CPU, auto-optimizing DaCe consistently outperforms or is equivalent to CuPy, 3.75x (geometric mean) faster. The auto-optimization passes contribute to these results, mainly attributing to subgraph fusion and avoiding intermediate allocations on shorter applications. Due to redundant copy removal and view semantics being native to the SDFG, we see a particular improvement on stencils, e.g., heat3d. Although CuPy-optimized code could potentially employ similar transformations [64], as far as we know, this cannot be performed out-of-the-box. The user must explicitly define element-wise or reduction based kernels, significantly changing the code. There is one instance where CuPy outperforms DaCe — resnet. This is due to a suboptimal vectorized representation of convolution in the Python source code, which translates to a loop of summations. In our generated code, this automatically results in many unnecessary atomic operations, even if tiled. The issue can be easily mitigated with further manual transformations (changing the maps’ schedules) after the fact.

FPGA results can be seen in Fig. 9, where there is no comparison point as no other framework compiles high-performance Python directly. Although both platforms use different languages and features (e.g., accumulators), applications can be synthesized for both from the same annotated Python code. There is a noticeable difference in performance, especially on stencil-like applications, likely resulting from Intel FPGA’s compiler toolchain superior stencil pattern detection. However, this can also be mitigated with subsequent manual transformations on the SDFG or augmenting the automatic heuristics decision-making process to transform stencils explicitly. Library Node expansions take device-specific features into account. For example, when there are accumulations (e.g., GEMV), we take

![Figure 8: CuPy and DaCe GPU runtime (lower is better).](image-url)
advantage of hardened 32-bit floating-point accumulation on Intel FPGAs, which allows single-precision numbers to be directly summed into an output register. On Xilinx FPGAs, and, in general, for 64-bit floating-point accumulation, such native support does not exist. Therefore, we perform accumulation interleaving [24] across multiple registers to avoid loop-carried dependencies.

3.4.3 Discussion. While DaCe already outperforms existing libraries, this is not the end of the optimization process. Instead, the generated SDFGs can be starting points for optimization by performance engineers. The provided transformations and API can be used to eliminate sources of slowdown in the above applications or extended to use new, domain-specific transformations [25, 40]. Furthermore, the applications can also be adapted to distributed memory environments, where the productivity and performance benefits can be even greater.

4 SCALABLE DISTRIBUTED PYTHON

The data-centric representation of Python programs can serve as a starting point for creating distributed versions. These distributed SDFGs abandon the global view of the data movement in favor of a local one. Like Message Passing, the flow of data in distributed memory is explicitly defined through Library Nodes. This approach allows for fine-grained control of the communication scheme and better mapping of SDFGs to code using optimized communication libraries.

In this Section, We show how to design transformations that specialize parallel map scopes to support distributed memory systems. We then show how to optimize such distributed data-centric programs. Finally, we show how developers can take control of the distribution entirely by expanding the original Python code with distributed communication while still allowing data-centric optimization to occur.

4.1 Transforming for Scale

Leveraging the data-centric representation, we can create transformations that convert specific shared-memory parallel kernels into distributed memory. The advantage of this approach is that once such a transformation is available, we can apply it to any subgraph in any SDFG that matches the same pattern. Furthermore, transformations can be compounded, building on each other to achieve complex results. We focus again on the gemm kernel for illustrating the transformations. As we shall show, the transformations extend beyond and automatically distribute other kernels as well.

Distributing global view element-wise operations. We distribute these by following a scatter-gather pattern, broadcasting (scalars) or scattering (arrays) input data containers from the root rank to the machine nodes, performing local computation, and gathering or reducing the outputs. By the nature of element-wise operations, careful selection of the array distribution parameters is not always necessary. The primary constraint is that each rank receives matching subsets of data, allowing it to perform local computation without further communication. Therefore, the most efficient distribution for contiguous arrays is to treat them as uni-dimensional and scatter them with MPI_Scatter (1-D block distribution). However, in cases where the result of an element-wise operation is consumed, e.g., a matrix-matrix product or a stencil computation, it is beneficial to preserve the dimensionality of the arrays. For this reason, the transformation has optional parameters for the block sizes per dimension, leading to block or block-cyclic distributions. We offer implementations that use PBLAS [55] methods, such as p?gemr2d and p?tran, and MPI derived data types, which have previously demonstrated performance benefits [68]. Applying the above transformation with block distributions on the operation `tmp8 = alpha * A` transforms the SDFG subgraph as shown in Fig. 10. We emphasize that these transformations are applied to an SDFG without changing the data-centric program code:

```python
sdg.apply(DistributeElementWiseArrayOp)
```

Figure 9: FPGA runtime, Large instance, single precision.

Figure 10: Distribution of element-wise array operation.

Distributing Library Nodes. We also create expansions for Library Nodes to distributed SDFG subgraphs. For example, the matrix-matrix and matrix-vector products expand to the aforementioned PBLAS library calls, along with the corresponding distribution of inputs and gathering outputs. Using PBLAS requires the definition of a process grid. The DaCe PBLAS library environment handles this automatically using BLACS [54]. The grid’s parameters are free symbols that can be chosen by the user or take default values.

4.2 Optimizing Communication

Creating distributed versions of the operations separately from each other will perform correctly but poorly on real applications. Thus,
we can use the data-centric aspect of the SDFG IR, which can track access sets through memlets to remove such communication bottlenecks automatically. Doing so separately from the distribution transformations allows users to write more pattern-matching distribution transformations without worrying about inter-operation communication, on the one hand; and on the other hand, allows the system to find such optimizations in any input code (e.g., if manually-written with communication redundancy).

One example of such a transformation is redundant gather-scatter removal. Consider the SDFG representation of `gemm`, shown in Fig. 5. We distribute all three element-wise array operations and we expand the `MatMult` node to a call to `pdgemm`. Due to the produced scatter and gather operations, the outputs `tmp1` and `tmp2` of `pdgemm` and `tmp2` of `beta * C` will be connected to the last element-wise array operation `tmp1 + tmp2` as shown in the following pseudo-code:

```python
pdgemm(..., 1tmp1, ...)
1tmp1 = gather(1tmp1)
1tmp1 = scatter(1tmp1)
1tmp2 = gather(1tmp2)
1tmp2 = scatter(1tmp2)
1C = 1tmp1 + 1tmp2
...
```

The above sequence of operations yields redundant communication on `tmp1` and `tmp2` and can therefore be omitted (the transformation for `tmp1` is shown in Fig. 11). By following the data movement and inspecting other Access nodes in the state, data dependencies of the global array can be inferred. Furthermore, since we know the Scatter and Gather node semantics, we can check whether the data distributions match. We note that users can use the DaCe API to define transformations to, e.g., optimize the re-distribution of data when the distributions do not match. If the distributions are 2D block-cyclic, such a transformation could, among other solutions, utilize PBLAS and a `p2gemm2d` Library Node to efficiently bypass the Scatter and Gather operations.

Combining the above transformations, the shared-memory `gemm` program from Section 2.3 can be converted to distributed-memory as follows, again without altering the code of the data-centric Python program (type annotations omitted for brevity):

```python
@dace.program
def gemm(alpha, beta, A, B):
    C[:,] = alpha * A @ B + beta * C
```

```python
dist_sdfg = gemm.to_sdfg()
dist_sdfg.apply(DistributeElementwiseArrayOp)
dist_sdfg.expand_library_nodes('PBLAS')
dist_sdfg.apply(RemoveRedundantComm)
```

4.3 Assuming Direct Control via Local Views

The implicit, global view approach works well for a plethora of Python programs that make heavy use of high-level array operations. However, as highly-tuned HPC applications often use specific partitioning schemes, our data-centric toolbox also provides explicit control via Python annotations. As opposed to the existing tools that manage communication implicitly, the aim of the interface is to use (num)pythonic concepts to retain productivity while maximizing performance. E.g., the `jacobii_2d` stencil below would yield unnecessary Scatter and Gather collectives at every timestep:

```python
@dace.program
def jacobii_2d(TSTEPS: dace.int32, A: dace.float64[N, N], B: dace.float64[N, N]):
    for t in range(1, TSTEPS):
        B[:1:-1,1:-1] = 0.2 * (A[:1:-1,1:-1] + A[:1:-1,2:] + A[:1:-2,:-1] + A[:1:-2,2:])
        A[:1:-1,1:-1] = 0.2 * (B[:1:-1,1:-1] + B[:1:-1,2:] + B[:1:-2,:-1] + B[:1:-2,2:])
```

For this reason, our data-centric approach allows the user to express arbitrary communication patterns by integrating explicit communication directly into Python. The above shared-memory data-centric Python program can be modified to be distributed:

```python
@dace.program
def half_step(inbuf: dace.float64[1Nx+2, 1Ny+2], outbuf: dace.float64[1Nx+2, 1Ny+2]):
    req = np.empty((8,), dtype=mpi4py.MPI_Request)
    dace.comm.Isend(inbuf[1, 1:-1], 0, req[0])
    dace.comm.Irecv(outbuf[:, 0], 0, req[1])
    dace.comm.Waitall([req])
```

In the above program, we distribute the arrays A and B into 2D blocks at the beginning. The local views IA, IB are then computed and communicated via explicit halo exchange, using `Isend`, `Irecv`, and `Waitall` MPI calls in every time-step.

While this approach is similar to the mpi4py bindings, there are two distinct advantages to the data-centric approach with an explicit local view. First, the MPI calls are integrated into the program’s dataflow with Library Nodes, enabling the above transformations and other automatic code generation features such as overlapping. Second, explicit `Isend` and `Irecv` calls communicate strided data using the MPI vector datatype, avoiding extraneous copies. The latter is also an example of using symbolic information on the graph to assert that high performance is attained — our MPI...
derived data type creation code, which is created once for each data type, relies on the symbol values not changing over the run. E.g., the initialization of the `inbuf[1:1, 1]` data type:

```c
MPI_Type ntype;
MPI_Type_vector(1Nx, 1, 1Ny+2, MPI_DOUBLE, &ntype);
MPI_Type_commit(&ntype);
```

would raise a performance warning in DaCe Python if the sizes may change at runtime. This avoids potential mistakes that even experienced performance engineers can make in large HPC codes.

### 4.4 Evaluation

To measure the performance of distributed data-centric programs, we conduct scaling experiments on the multi-core partition of the Piz Daint supercomputer. Each node has two 18-core Intel E5-2698v3 CPUs and 64GB of memory. The nodes are connected through a Cray Aries network using a Dragonfly topology. We benchmark a subset of the Polybench kernels from Section 3.4, which could be automatically transformed to use distributed memory (Section 4.1): adi, biCG, doItgen, gemm, gemver, gesummv, k2mm, k3mm, mvt, jacobi_1d, and jacobi_2d. We compare this work with Dask [23] v2.31 and Legate [10] (commit ID febd3bf [57]), two state-of-the-art distributed tasking Python frameworks, in weak scaling from 1 to 1,296 processes (648 nodes). Dask Array [4] scales a variety of NumPy workflows, including element-wise array operations, reductions, matrix-matrix products, and linear algebra solvers, among others. Legate is providing a drop-in replacement for the NumPy API to accelerate and distribute Python codes.

We select initial problem sizes that fit typical HPC workloads without having excessive runtime. The kernels’ scaling factors and the initial problem sizes for each framework are presented in Tab. 2. For kernels with computational complexity ranging from $O(n)$ to $O(n^2)$, we target a runtime in the order of 100ms. For kernels with higher complexity, we target a runtime in the order of 1s. We note that with the problem sizes selected for benchmarking data-centric Python and Legate, Dask either runs out of memory or exhibits unstable performance. Thus, we halved the problem sizes for Dask but still encountered out-of-memory errors at larger node counts. Furthermore, several issues rendered testing Legate at scale difficult. With the assistance of the developers, many of those problems were solved; however, others remained. We could only run each benchmark up to a fraction of the total nodes available, either due to runtime errors or because a single execution did not finish within 10 minutes of allocated time. We annotate Fig. 12 with these errors.

We ignore the time needed for initializing and distributing data type: relies on the symbol values not changing over the run. E.g.,

```c
MPI_Datatype ntype;
MPI_Type_commit(& ntype);
MPI_Type_vector (lNx , 1, lNy +2 , MPI_DOUBLE , & ntype);
```

Table 2: Distributed benchmarks, initial problem sizes for the different frameworks (F), and scaling factors (S.F.) as a function of the number of processes $S$.

<table>
<thead>
<tr>
<th>Benchmark</th>
<th>F</th>
<th>Initial Problem Size</th>
<th>S.F.</th>
</tr>
</thead>
<tbody>
<tr>
<td>atax</td>
<td>DaCe/Legate Dask</td>
<td>20000, 25000</td>
<td>all √$S$</td>
</tr>
<tr>
<td>biCG</td>
<td>DaCe/Legate Dask</td>
<td>25000, 20000</td>
<td>all √$S$</td>
</tr>
<tr>
<td>doItgen</td>
<td>DaCe/Legate Dask</td>
<td>128, 512, 512</td>
<td>(S, √$S$)</td>
</tr>
<tr>
<td>gemm</td>
<td>DaCe/Legate Dask</td>
<td>8000, 9200, 5200</td>
<td>all √$S$</td>
</tr>
<tr>
<td>gemver</td>
<td>DaCe/Legate Dask</td>
<td>10000, 5000</td>
<td>√$S$</td>
</tr>
<tr>
<td>gesummv</td>
<td>DaCe/Legate Dask</td>
<td>22400, 11400</td>
<td>√$S$</td>
</tr>
<tr>
<td>jacobi_1d</td>
<td>DaCe/Legate Dask</td>
<td>10000, 24000</td>
<td>(−, S)</td>
</tr>
<tr>
<td>jacobi_2d</td>
<td>DaCe/Legate Dask</td>
<td>1000, 1300</td>
<td>(−, √$S$)</td>
</tr>
<tr>
<td>k2mm</td>
<td>DaCe/Legate Dask</td>
<td>6400, 7200, 4400</td>
<td>all √$S$</td>
</tr>
<tr>
<td>k3mm</td>
<td>DaCe/Legate Dask</td>
<td>3200, 3600, 2200</td>
<td>all √$S$</td>
</tr>
<tr>
<td>mvt</td>
<td>DaCe/Legate Dask</td>
<td>22000, 11000</td>
<td>√$S$</td>
</tr>
</tbody>
</table>

Figure 12: Distributed runtime (dashed lines) and scaling efficiency (solid lines) on 23,328 cores of Piz Daint.
whereas in others we observe slowdowns of 1.7–15

Approaches similar to our own targeting Python code have already

Therefore, an HPC project must force the domain scientists to sacri-

fice productivity and work directly on the lower-level languages or

in C, C++, and Fortran, among other device-specific languages.

is amenable to low-level optimizations for the underlying archi-

tectures. Traditionally, this translates to writing these applications

in Python, many other research compilers are based on the

LLVM [47] infrastructure and IR. There is an ongoing movement

in the compiler community towards Multi-Level IRs [48], in which

a multitude of IR dialects can retain domain- and platform-specific

information, in turn enabling domain-specific optimizations [35].

MLIR performs optimization passes on each dialect to compile pro-

grams, followed by lowering passes to subsequent dialects, down

to hardware mapping. This feature could be utilized to implement

DaCe Library Nodes. The data-centric transformation API is also

shared by languages such as Halide [65], which enables users to

invoke schedule optimizations separately from program definition.

5 PRODUCTIVITY

Python is already a very productive language, especially for domain

scientists, due to its rich ecosystem described in Section 1. Data-

Centric Python is Python with extensions that themselves are valid

Python syntax. For example, the \texttt{dace.program} decorator follows

the PEP 318 standard [28]. Therefore, Data-Centric Python essen-

tially inherits the Python language’s programming productivity.

Since performance is the most important metric in HPC, scientific

applications must eventually be lowered to a representation that

is amenable to low-level optimizations for the underlying architec-

tures. Traditionally, this translates to writing these applications

in C, C++, and Fortran, among other device-specific languages.

Therefore, an HPC project must force the domain scientists to sacri-

fice productivity and work directly on the lower-level languages or

maintain two different code-bases. DaCe, and other frameworks that

accelerate Python, increase HPC productivity by bridging the code

that domain scientists want to write with the code that achieves

high performance.

6 RELATED WORK

Approaches similar to our own targeting Python code have already

been introduced and compared with in Sections 1, 3, and 4. In this

section, we further discuss relevant frameworks, libraries, and

approaches towards the three Ps.

Productivity. The complexity of optimizing applications, com-
bined with the repetitive nature of performance engineering for

specific domains, has given rise to a wide variety of Domain-Specific

Languages (DSLs) [19, 38, 59, 69] and embedded DSLs, particularly

in Python [46, 72]. In the latter category, a notable example is

deep learning frameworks, which use Python’s various capabili-

ties to construct readable code that performs well. PyTorch [61]

uses object-oriented programming to construct deep neural net-

works as modules, relying on reflection to detect parameters and

nonblocking calls for asynchronous execution to avoid interpreter

overhead. TensorFlow [2] used Python’s weak typing system to

construct graphs from Python functions but has recently transi-
tioned to “eager” execution to improve productivity, making codes

more readable and simplifying debugging.

Portability. In the past three decades, compilers have undergone

a transition from all-pairs solutions (between source languages and

hardware platforms) to funneling through Intermediate Represen-
tations (IR), on which they can perform language- and platform-

agnostic optimization passes. Although DaCe is currently devel-
poped in Python, many other research compilers are based on the

LLVM [47] infrastructure and IR. There is an ongoing movement

in the compiler community towards Multi-Level IRs [48], in which

a multitude of IR dialects can retain domain- and platform-specific

information, in turn enabling domain-specific optimizations [35].

MLIR performs optimization passes on each dialect to compile pro-

grams, followed by lowering passes to subsequent dialects, down

to hardware mapping. This feature could be utilized to implement

DaCe Library Nodes. The data-centric transformation API is also

shared by languages such as Halide [65], which enables users to

invoke schedule optimizations separately from program definition.

Performance portability. Aimed at keeping a consistent ratio of

performance to peak performance across hardware [73], it is the

core premise of several standards [6, 21, 31, 32, 58]. In directive-

based frameworks [21, 58], \texttt{pragma} statements are added to C/C++

and FORTRAN programs to introduce parallelism, similarly to our

proposed annotations. In kernel-based frameworks [6, 31, 32], ker-

nels are constructed as functions with a limited interface and of-

floated to target devices, such as CPUs, GPUs, or FPGAs. As each

platform requires its own set of directives, kernel parameters, or

sometimes kernel implementations, programs often contain multi-

ple codes for each target. To resolve such issues, HPC languages

such as Chapel [18] and HPF [66] define high-level implicit ab-

stractions used as parallel primitives. Also popular in the HPC

world are performance-portable libraries, embedded within C++,

notably Kokkos [26], RAJA [50], and Legion [11] (which powers

Legate [10]). These allow integrating heterogeneous and distributed

systems through task-based abstractions, data dependency analysis

(e.g., Legion’s logical regions), and common parallel patterns. Such

patterns can also be found in NumPy, and the SDFG can be seen as

a generalization of these graphs with symbolic data dependencies.

7 CONCLUSION

Discussion. While distributed SDFGs forego the global view of

data movement to facilitate the design of custom communication

schemes, future work could explore the trade-offs between a more
pythonic approach to communication and extracting the best performance. Moreover, improvements to existing transformations, e.g., Vectorization, and implementation of new ones could increase the out-of-the-box performance, reducing the need for manual optimization.

We present Data-Centric Python—a high-performance subset of Python with annotations that produces supercomputer-grade HPC codes. Based on the SDFG intermediate representation, we show how a pipeline of static code analysis and dataflow transformations can take input NumPy code, leverage its vectorized nature, and map it efficiently to CPUs, GPUs, and FPGAs, outperforming current state-of-the-art approaches on each platform by at least 2.4x.

The resulting data-centric programs effectively eliminate the performance pitfalls of Python, including interpreter overheads and lack of dataflow semantics for library calls, the latter being crucial for running at scale. Many even outperform baselines written in C code. A key feature of the data-centric toolbox is giving users explicit control when necessary, rather than making assumptions at the framework level. We evaluate Data-Centric Python on a distributed environment and show that the parallel efficiency remains above 90%, even on hundreds of nodes. These promising results indicate that productive coding with Python can scale and map to heterogeneous compute architectures, setting the once-scripting language at the same level as FORTRAN, C, and other HPC giants.

8 ACKNOWLEDGEMENTS

This project received funding from the European Research Council (ERC) under the European Union’s Horizon 2020 program (grant agreements DAPP No. 678880, EPiGRAM- HS No. 801039, and DEEP-SEA No.955606). The Swiss National Science Foundation supports Tal Ben-Nun (Ambizione Project No. 185778). The authors would like to thank Mark Klein and the Swiss National Supercomputing Centre (CSCS), Paderborn University (DaceML-FPGA project), and Xilinx (XACC program) for support and access to computational resources.

REFERENCES


