A Computational Stack for Cross-Domain Acceleration

Sean Kinzer Joon Kyung Kim Soroush Ghodrati Braharendra Yatham
Alric Althoff* Divya Mahajan† Sorin Lerner Hadi Esmaielzadeh

Abstract—Domain-specific accelerators obtain performance benefits by restricting their algorithmic domain. These accelerators utilize specialized languages constrained to particular hardware, thus trading off expressiveness for high performance. The pendulum has swung from one hardware for all domains (general-purpose processors) to one hardware per individual domain. The middle-ground on this spectrum—which provides a unified computational stack across multiple, but not all, domains—is an emerging and open research challenge. This paper sets out to explore this region and its associated tradeoff between expressiveness and performance by defining a cross-domain stack, dubbed PolyMath. This stack defines a high-level cross-domain language (CDL), called PMLang, that in a modular and reusable manner encapsulates mathematical properties to be expressive across multiple domains—Robotics, Graph Analytics, Digital Signal Processing, Deep Learning, and Data Analytics. PMLang is backed by a recursively-defined intermediate representation allowing simultaneous access to all levels of operation granularity, called sDFG. Accelerator-specific or domain-specific IRs commonly capture operations in the granularity that best fits a set of Domain-Specific Architectures (DSAs). In contrast, the recursive nature of the sDFG enables simultaneous access to all the granularities of computation for every operation, thus forming an ideal bridge for converting to various DSA-specific IRs across multiple domains. Our stack unlocks multi-accelerator for end-to-end applications that cross the boundary of multiple domains each comprising different data and compute patterns.

Evaluations show that by using PolyMath it is possible to harness accelerators across the five domains to realize an average speedup of 3.3× over a Xeon CPU along with 18.5× reduction in energy. In comparison to Jetson Xavier and Titan XP, cross-domain acceleration offers 1.7× and 7.2× improvement in performance-per-watt, respectively. We measure the cross-domain expressiveness and performance tradeoff by comparing each benchmark against its hand-optimized implementation to achieve 83.9% and 76.8% of the optimal performance for single-domain algorithms and end-to-end applications. For the two case studies of end-to-end applications (combining algorithms from multiple domains), results show that accelerating all kernels offers an additional 2.0× speedup over CPU, 6.1× improvement in performance-per-watt over Titan Xp, and 2.8× speedup over Jetson Xavier compared to only the one most effective single-domain kernel being accelerated. Finally, we examine the utility and expressiveness of PolyMath through a user study, which shows, on average, PolyMath requires 1.9× less time to implement algorithms from two different domains with 2.5× fewer lines of code relative to Python.

Index Terms—Accelerator design, Machine learning systems, Algorithms/System Co-design, Domain Specific Language

I. INTRODUCTION

End-to-end applications ranging from delivery drones [1] to smart speakers [2] cross multiple domains. One such application senses the environment, (1) pre-processes the sensory data, feeds it to a (2) perception module that in turn invokes a (3) decision making process to determine actions. Perception is currently reigned by Machine Learning (ML), which has attracted significant attention, but applications are not just ML. Sensory data processing relies on algorithms from Digital Signal Processing (DSP) while Control Theory and Robotics bring forth the final action that may also feed the perception module. Even though these domains work in tandem to realize an entire application, they are becoming isolated by the current push towards Domain-Specific Accelerators (DSAs). One the one hand, these accelerators tradeoff generality for performance and energy efficiency by restricting programmability to a single domain [3]. On the other hand, the traditional general-purpose computational stack cannot meet the computational demands of emerging applications [4][6]. Although, Domain-Specific Accelerators (DSA) bridge this performance gap, but make implementation an arduous task of dealing with isolated programming interfaces. Thus, expressiveness is becoming limited, making the composition of an end-to-end application a major challenge for execution on accelerators. As a
consequence, users seeking to create compute-intensive applications composed of algorithms from different domains must choose between either using a lower-performance, general-purpose processor or bear the burden of manually stitching together various domain-specific accelerators.

This emerging challenge creates a new tradeoff between performance and expressiveness, illustrated in Figure 1. On one extreme, we have General-Purpose Processors that allow expressing all domains at the cost of performance and/or efficiency. On the other extreme, are domain-specific accelerators that can only support a single domain to be executed on one particular specialized architecture, thus are very performant. Even though certain DSAs offer computational stacks, composing an end-to-end application that crosses the boundary of many domains requires intimate knowledge of multiple different interfaces and various hardware accelerators to obtain high performance. Recent efforts such as, Graphicionado, RoboX, TVM, Tabla, aim to unify high-level coding within a single domain, cross-domain stacks for accelerators still remain an open challenge.

As Figure 1 illustrates, by addressing this challenge, PolyMath defines a new point in the Performance vs. Expressiveness design space. The ingredients of our approach are:

1) Exploiting the mathematical similarities across domains to design a modular and reusable language, PMLang, that is expressive across Robotics, Graph Analytics, DSP, Data Analytics, and Deep Learning. It offers one-to-one mapping between code and mathematical formulation while retaining modularity, thus making it familiar to both domain experts and software engineers. PMLang offers lightweight type modifiers based on domain semantics to enable accelerators to handle on-chip and off-chip data allocation, storage, and transfer. PMLang is not an abstraction over existing domain specific languages, rather it explores a novel dimension of designing a unified, stand-alone language across multiple domains. This unique dimension defines a class of languages that we refer to as Cross Domain Languages (CDLs). To enable cross-domain acceleration, PMLang and its associated compilation stack takes an initial step to bridge CDLs with domain-specific architectures, that are constrained to a single domain.

2) To preserve expressiveness and provide flexibility for compilation to different accelerators, we devise an intermediate representation that is a recursively-defined Dataflow Graph, providing simultaneous access to all levels of operation granularity. The compute granularity of the kernels is not uniform across different accelerators, required for cross-domain settings. Thus, we define sDGF recursively in that its nodes are also sDFGs. As such, sDGF uniquely offers simultaneous access to various levels of computation granularity within a single program, thus enabling leveraging different accelerators. This capability enables cross-domain multi-acceleration—acceleration of a cross-domain application on different accelerators. The flexible, recursive nature of our IR shown in Figure 2 whose edges preserve the type modifier metadata.

3) PolyMath uses a modular compilation framework that conveniently enables creation and application of pipelined compilation passes on the sDGF. To convert sDGFs to executable accelerator code, PolyMath offers a graph lowering algorithm with a conversion strategy which uses metadata embedded in the sDGF edges to flexibly translate the graph nodes. The lowering algorithm applies transformations which produce a new sDGF made up of compute kernels at the same granularity of the target accelerator. Once lowered, the metadata associated with the sDGF edges is translated to the accelerator’s own IR for final binary generation through its own scheduling and mapping framework.

4) Last but not least, PolyMath will be the very first extensible, modular, and open-source computation stack to enable the community to innovate and explore the impending challenges of cross-domain acceleration at a time when domain-specific compilation stacks, except for DNNs, are elusive. We have made the entire codebase available in a public repository (https://github.com/he-actlab/polymath).

Although there are many accelerators in the literature, many of their compilation stacks are not available. By making PolyMath open-source and extensible, the community can add other domains which align with the core mathematical constructs in PMLang.

We show PolyMath’s balance of expressiveness and performance by compiling twelve different algorithms across robotics, graph analytics, digital signal processing, data analytics, and deep learning. These workloads achieve an overall speedup of 3.3× over a Xeon CPU along with 18.1× reduction in energy. In comparison to Titan Xp and Jetson Xavier GPUs, cross-domain acceleration offers 7.2× and 1.7× in energy reduction. We next measure the tradeoff of cross-domain algorithm expression and find that PolyMath can achieve 83.9% of the performance of the same algorithms implemented in their native stack’s language. We also study two end-to-end applications that cross multiple domains. Accelerating all the kernels offers an additional 2.0× speedup over CPU, 6.1× additional improvement in energy requirements over Titan Xp, and 2.8× speedup over Jetson Xavier vs when only one most effective kernel was accelerated. The results show that when only a part is accelerated, the slower non-accelerated kernels dictates the overall improvement, whereas, when all the algorithms in the application are accelerated Amdahl’s burden reduces, and the improvement of all the domains is magnified. To evaluate the usability and expressiveness of PMLang relative to Python, we conduct a user study and found that on average PMLang required 1.9× less time to implement algorithms with 2.5× fewer lines of code. Finally, end-to-end performance of Polymath is 76.8% of the performance of two manually implemented end-to-end applications. Given the fact that PolyMath offers greater ease of programming compared to Python, the automation overhead of 23.1% (=100%-76.8%) is a fair bargain.
II. PMLang: Mathematical Programming Interface

PMLang is designed to encapsulate the mathematical properties of these domains, as they are tied together by similar operations on multi-dimensional data, include minimal control flow, and share use-cases such as cyber-physical systems. Consider the following application in its entirety: An end-to-end neuroscience application requires multiple domains to study the impact of deep brain stimulation on movement disorders and goes through the following steps: (1) convert raw electrocorticographic (ECoG) brain signals to frequency domain using fast Fourier transform (FFT); (2) apply logistic regression to classify these frequency domain signals into various biomarkers; (3) based on the classification, use model predictive control to send an optical stimulation back to the brain. This application crosses three domains, DSP, Data Analytics, and Control Theory in each iteration to generate deep brain stimulation signals.

There are numerous domain specific architectures for each of these algorithms/domains individually; however, using them for this application would require writing each part of the application in a different DSL, compiling them separately, and manually joining their executables. Instead, PMLang allows users to write their application as a single program, thus, eliminating the overhead of stitching together stacks to execute the program across multiple domain specific architectures.

Keeping the properties of target domains in mind, PMLang is designed to reduce the time to code a mathematical expression into a formula-based textual format, enabled by language constructs for modularity and light-weight type modifiers. Moreover for code organization and reduction in implementation time, PMLang includes reusable execution code blocks called components that perform operations on flows of data. These components encapsulate a task comprised of either other components and/or mathematical expressions which use traditional, imperative syntax to facilitate familiarity for experienced programmers. For modularity and reusability, components have distinct boundaries and arguments which are distinguished by type modifiers consisting of input, output, state, and param; each of which is associated with how the component will use the argument, shown in Table I. By using type modifiers in component arguments to explicitly identify data semantics PMLang binds operations to data being operated on for accelerators to take advantage of.

The remainder of the section will delve into details of PMLang constructs through an example. For brevity, we show the PMLang program for Model Predictive Control (MPC) from Control Theory used for Robotics. MPC attempts to solve a constrained optimization problem over a finite sequence of inputs. MPC can also be used for the aforementioned brain stimulation application. We provide MPC in the context of a mobile two-wheeled robot performing trajectory tracking, shown in Figure 3. Here, the sensors send the current state of the robot as inputs to a model that predicts the next location, then optimizes a sequence of control signals for moving the robot to the predicted location. The program finds the optimal sequence of control signals, \( ctrl_{mdl} \), over a finite period of time to match a reference trajectory, \( pos_{ref} \), that specifies the position and orientation of the robot. At each point in time, the actual position and orientation of the robot is input to the algorithm, which optimizes the \( ctrl_{mdl} \) and sends the next control signal, \( ctrl_{sgnl} \), back to the robot. This process is summed up in the following steps:

**Input:** \( pos \rightarrow (x,y,\theta) \) orientation of the robot.
**Output:** \( ctrl_{sgnl} \rightarrow (\nu,\omega) \) control consisting the velocity and angular velocity to be sent to the robot.
**State:** \( u \rightarrow (v,\omega) \) linear and angular velocities across a predetermined time horizon \( h \).

**Step 1: Make a prediction** Using input \( pos \) and cost matrices \( P \) and \( H \), predict the position and angle of the robot across horizon \( h \).

**Step 2: Compute the gradient of the objective function** Calculate the error on the predicted position and orientation, \( pos_{pred} \), using pre-computed gradient coefficient matrices \( HQ_{g} \) and \( R_{g} \).

**Step 3: Update the control model and send the output signal** Using gradient, \( g \), update the control model, and send the output control signal, \( ctrl_{sgnl} \).

A. Components

Components form the building blocks of PMLang, and are used to delineate different parts of the program into multiple levels of execution. To delineate the access semantics for each argument of the components, PMLang uses type modifiers ((input), output, state, param). Using type modifiers relieves programmers from concern about the underlying accelerator-specific mechanisms for data exchange between different components. An input argument is used to feed data into the component, and is read-only; an output argument is used to return data from a component, and can only be written to; a state argument can be read or written.
to, and represents data that is part of the state of the component, thus is preserved across invocations/iterations; and a param argument is a constant that is used to parameterize the component. These type modifiers describe whether or not the data will be re-used (state), kept unchanged (param), or used once and discarded (input/output). As an example, line 1 shows that argument pos is an input to the predict_trajectory component. As another example, line 41 shows a state argument named ctrl_mdl which indicates that ctrl_mdl is used, updated, shared across invocations of main, which matches the MPC semantics of optimizing the control model over a series of time steps. Type modifiers also enable custom accelerators to place input data such as pos in Read-only FIFO buffers to reduce data communication overhead and hardware memory logic, or store state data such as ctrl_mdl on-chip on the accelerator for fast repeated data accesses. Using a single set of type modifiers to describe data semantics across multiple domains unifies program implementation for end-to-end applications. This is exemplified in robotics and deep learning, where one domain uses “model” and the other “weight” to describe the same data semantics, both of which are described as state data in PMLang.

In addition to being reusable, these components allow users to conceive their program as a collection of sub-steps at varying levels of granularity making it adaptable for compilation to different accelerators. To instantiate a component, the programmer specifies its name and arguments. An example of component instantiation is shown in line 49-51 where predict_trajectory, compute_ctrl_grad, and update_ctrl_model is instantiated. Each instantiation creates a copy of the component, as if it were inlined. This is in contrast to conventional languages that rely on a function call stack which is sequential in nature. Instead, inlining enables the program to be mapped to our sDFG IR, which preserves opportunities for parallelism based on data flow dependencies.

B. Index Variables

PMLang is based on mathematical notations that do not use for loops and instead use indices (e.g., \( \sum_{i=0}^{n-1} A_i \)). To simplify programming based on formulae, PMLang uses index variables to concisely specify operations performed over ranges of multi-dimensional data without using explicit for loops. In its most basic form, an index variable represents a range of integers, specified by its lower and upper bounds. Line 48 shows two such index variable declarations: i and j. This approach reflects the inherent parallelism in mathematical formulae since operations expressed using this approach are naturally vectorizable without performing any loop transformations. For example, below is a PMLang statement iterating over all \( j \)s, each of which can be performed in parallel.

```plaintext
err[j] = pos_ref[j] - pos_pred[j];
```

Strided indexing. To support strided/non-sequential indexing (e.g., convolution), PolyMath also supports arithmetic operations on index variables as shown below.

```plaintext
ctrl_mdl[i] = ctrl_prev[(i+1)*h] - g[(i+1)*h];
```

Boolean conditional over indexing. Unlike a domain-specific language such as TABLA [17] that focuses solely on data analytics, PMLang allows Boolean conditionals to be applied to indices, which provides support for other domains such as graph analytics and robotics. For instance, the following computes the sum of the non-diagonal parts of the matrix \( A \):

```plaintext
index i[0:N-1], j[0:M-1];
raz = sum[i][j] \!: i \!: j \!: A[i][j];
```

Support for Boolean conditionals and non-sequential index variables flexibly incorporates common as well as specific-to-domain characteristics of algorithms across robotics, graph analytics, DSP, data analytics, and deep learning. These features distinguish PMLang from DSLs which either use (1) concise operations expressed through a fixed API (e.g., named functions such as “dense” in TVM [16]), or (2) simply do not support these construction since the specific target domain does not require them (e.g., TABLA [17].)

C. Mathematical Operations

Index variables allow for a nearly one-to-one mapping between mathematical notation and PMLang code. PMLang offers standard mathematical operators to be used with multi-dimensional data, expressed in a single statement by using


index variables. PMLang’s syntax for math expression of

\[ C_j = \sum_{i=0}^{n} A_{j,i} \times B_i \]

Non-Linear operations. PMLang includes a set of built-in functions to be used in math expressions commonly used across the multiple target domains, including non-linear operations such as cosine/sine (DSP, robotics), gaussian (robotics, DSP, data analytics), sigmoid/ReLU (deep learning, data analytics), etc. Including non-linear operations as part of PMLang simplifies algorithm expression and allows PolyMath to leverage the performance benefits of non-linear compute units in custom accelerators.

Reduction operations. PMLang is also equipped with built-in group reduction operations such as sum, prod, max etc., to calculate the summation (\( \sum \)), product (\( \Pi \)), or maximum value of a sequence of numbers. These group reductions operations are converted to \( \text{srDFG} \) with two levels of granularity: (1) the outer group \( \text{DFG} \) node that (2) encapsulates the scalar inner operations (nodes). This multi-granular representation enables the compiler to map the node encompassing node to a dedicated unit if the accelerator harbors it. Otherwise, the inner basic nodes are mapped to individual ALUs. This crucial flexibility is a unique feature of PolyMath and enables it to cross domain and target different accelerators.

Custom reduction operations. PMLang also supports custom group reduction operations, as they are commonly used in graph analytics and DSP algorithms. Custom reduction operations can be defined in PolyMath by specifying the arithmetic for a given set of input arguments. Below is an example of the definition of the \( \text{min} \) reduction function and using it to find the minimum value for the matrix \( A \):

\[
\text{reduction \_min}(a, b) = a < b \ ? a : b;
res = \text{min}[i][j][A[i][j]];
\]

D. Domain Annotations

PMLang uniquely targets multiple domains, each of which is eventually accelerated with a Domain-Specific Architecture. As such, PMLang offers a light-weight mechanism to specify the target domain for \( \text{only top-level} \) component instantiations without tying it to a specific accelerator. All of the code within a component also inherits the same domain, which alleviates the programmer from having to annotate all component instances in their program. This is done by simply adding one of the five keywords: RBT (Robotics), GA (Graph Analytics), DSP (Digital Signal Processing), DA (Data Analytics), and DL (Deep Learning), as demonstrated below:

\[
\text{RBT: \_predict\_trajectory}(\text{pos, ctrl\_mdl}, F, H, \text{pos\_pred});
\]

III. SIMULTANEOUS-RECURSIVE DATAFLOW GRAPH

Accelerator-specific or domain-specific IRs commonly capture operations in the granularity that best fits the target architecture. One of the major challenges that PolyMath faces is targeting multiple domains each of whose accelerators operate on different granularities of computation. Even within a single domain, various architectures accept computation in different granularities. To address this challenge, we designed \( \text{srDFG} \), an intermediate representation which is a \( \text{recursively-defined} \) dataflow graph, and provides \( \text{simultaneous} \) access to each level of recursion, \( \text{srDFG} \). Our \( \text{srDFG} \) enables the compiler to simultaneously access all the granularities of computation for every component, thus forming the ideal bridge to convert to various accelerator-specific IR. Furthermore, \( \text{srDFG} \) enable multi-acceleration for end-to-end applications that cross the boundary of multiple domains with different data and compute patterns across Robotics, Graph Analytics, DSP, Data Analytics, and Deep Learning. Next, we describe the \( \text{srDFG} \) structure using Figure 3, a visual representation of the MobileRobot algorithm described in Section II.

A. \( \text{srDFG} \) Definitions

An \( \text{srDFG} \) is defined as a pair, \( \langle N, E \rangle \), of nodes \( N \) representing PMLang operations, and edges \( E \) representing input or output operands. An \( \text{srDFG} \) node \( n \in N \) is a pair \( \langle \text{name}, \text{srdfg} \rangle \) of a string representing the name of an operation, and its lower-granularity operations \( \text{srDFG} \) composition. Each numbered box in Figure 5 represents the \( \text{srdfg} \) for different nodes at varying granularities within the MobileRobot algorithm. As shown in 2, both the the subtraction operation and the \( \text{mmul} \) component are nodes. An edge \( e \in E \) is a tuple of source \( \text{src} \) and destination \( \text{dst} \) nodes, and the edge metadata \( \text{md} \): \( \langle \text{src}\_\text{dst}\_\text{md} \rangle \). Edge metadata consists of the type, type modifier, and shape of the operand associated with the edge. For math operations, input and output edges represent operands and results, whereas adjacent edges in component instantiations represent state, input and output arguments. This is illustrated in 1, where \( \text{pos} \) is the input argument, \( \text{ctrl}\_\text{sgnl} \) is the output argument, and \( \text{ctrl}\_\text{mdl} \) is the state argument that creates a cycle for multiple iterations. Given a node \( n \), we denote the name and \( \text{srdfg} \) as \( n\_\text{name} \) and \( n\_\text{srdfg} \), and similarly denote \( \text{src}\_\text{dst}\_\text{md} \) in an edge \( e \) as \( e\_\text{src} \), \( e\_\text{dst} \), and \( e\_\text{md} \). Lastly, the domain annotations previously described are translated to the \( \text{srdfg}\_\text{domain} \) attribute for each \( \text{srdfg} \).

B. \( \text{srDFG} \) Semantics

As an example, the \( \text{srDFG} \) shown in 3 will begin operations when the data in edges \( \text{R}\_g \) and \( \text{ctrl}\_\text{mdl} \) are
C. Enabling Different Accelerators

Custom accelerators support unique set of operations performed on a variety of different typed and shaped inputs and outputs. To ensure flexibility, each sr\textsuperscript{4}DFG includes operations as nodes, \( n \), as well as the more fine-grained operations to define the node, \( n.dfg \). To illustrate this point, each sr\textsuperscript{4}DFG in Figure 5 represents a possible operation supported by an accelerator. If \( 2 \) is supported by the target accelerator, the sr\textsuperscript{4}DFG can be transformed to consist only of the operations represented by each node in \( 1 \). If \( 2 \) is not supported but \( 3 \) is supported, then the operations in both \( 1 \) and \( 2 \) will be selected for compilation. PolyMath’s base unit of lowering is a node, and if the nodes in the sr\textsuperscript{4}DFG cannot be lowered to a specific hardware because of unsupported nodes, the compilation fails for that accelerator.

Each of these accelerator operations is closely tied to the types and shapes of its operands. The sr\textsuperscript{4}DFG uses the edge metadata to specify the operand information when performing compilation of a node to an acceleration operation. For example, an accelerator might support the element-wise multiplication in \( 4 \), but requires the number of elements being multiplied to perform the operation. Each input edge to \( 3 \) includes the shape as part of its metadata, which allows for compilation of \( 4 \). Domain-specific accelerators differentiate how data is stored by receiving this information from the programmer on how the variables are used. For example, the ctrl\_mdl edge in \( 1 \) has the state type modifier which causes the accelerator to store the data local to the component.

IV. Compilation Framework

PolyMath performs compilation in three steps: (1) compilation from PMLang to an sr\textsuperscript{4}DFG; (2) lowering the sr\textsuperscript{4}DFG to the granularity of the different domains and converting it to the accelerator’s IR; (3) invoking the accelerator’s provided compiler to generate the final binaries. Figure 4 illustrates these phases for a PMLang graph analytics implementation which is compiled to an sr\textsuperscript{4}DFG, and then lowered and converted to GRAPHICIONADO’s pipeline IR.

A. sr\textsuperscript{4}DFG Generation

For each PMLang program, compilation forms an Abstract Syntax Tree (AST) using syntax analysis. The program AST is then traversed and a symbol table \( S \) is created, storing information contained in each component. The component information consists of variable names and variable metadata \( md \) (e.g., edge type, type modifier, and shape). For each component, a DFG is formed by stitching statements together using static single assignment. Lower-level, sr\textsuperscript{4}DFG operations are formed for element-wise operations or group operations, which means edges may represent both scalar and multi-dimensional values.

After completing AST traversal generating sr\textsuperscript{4}DFGs from PMLang statements, a single sr\textsuperscript{4}DFG is generated starting with the highest level component \( main \). Previously, component statements were skipped because all component sr\textsuperscript{4}DFGs were not created. When the \( main \ sr\textsuperscript{4}dfg \) is traversed, a compo-
nent node \( n_e \) is created using previously skipped component statements and their argument type modifiers, preserving the domain annotation as \( n_e, \text{dfg}_d, \text{domain} \) attribute. Edges adjacent to \( n_e \) are added to \( \text{srdfg} \) by using the type modifiers for arguments in the component signature of \( n_e \), where type modifiers are (1) in/out edge sfor input/output, and (2) edges such that \( e = (\text{src}, \text{dst}, \text{md}) \) where \( \text{src} = \text{dst} \) for state. The \( \text{srDFG} \) is repeated for each component statement recursively, creating nodes and edges to generate the lower levels of operation granularity for each component, which inherit the top-level domains.

B. Example \( \text{srDFG} \) Passes

PolyMath implements a modular framework and set of APIs that enable custom, target-independent passes over the IR. These passes take an \( \text{srDFG} \) as an input and produce a transformed \( \text{srDFG} \). This feature conveniently enables applying pipelines of passes on the same IR. Also, traditional passes such as constant propagation, constant folding, etc. are supported via this PolyMath pass infrastructure. We cover one such compiler pass below.

Algebraic combination. Transformation passes in PolyMath benefit from simultaneous access to all levels of operation granularity for a program. This bolsters traditional compiler passes such as algebraic simplification that are typically limited by single granularity IRs which hide opportunities for simplification. In contrast, a PolyMath pass can identify hidden algebraic simplifications which span multiple levels of granularity which would remain obscured in other flat IRs. As an example, if an \( \text{srDFG} \) with a top-level matrix-vector multiplication is added to the output of another matrix-vector operation contained in another node’s subgraph, the matrix vector operations can be fused together by concatenating their inputs. This transformation opportunity remains unidentified in flat IRs, but PolyMath uniquely reveals these transformation prospects by preserving a program’s multi-granularity in the \( \text{srDFG} \) and supporting transformations crossing granular boundaries.

C. Compilation from \( \text{srDFG} \) to Accelerator IR

Algorithm 1: \( \text{srDFG} \) Lowering Algorithm

```
function \( \text{Lower} (\text{srdfg}, \text{O}_\text{m}) \)
let \( (N,E) = \text{srdfg} \), \( \text{subDfg} \) for each \( n \in N \) do
if \( \text{name} \notin \text{O}_\text{m} \) then
let \( \text{subDfg} = \text{Lower}(n, \text{O}_\text{m}) \)
\( \text{srdfg} \leftarrow \text{srdfg} \rightarrow \text{subDfg} \)
end
return \( \text{srdfg} \)
```

Compiling a \( \text{srDFG} \) to a domain-specific architecture consists of (1) lowering \( \text{srDFG} \) operations supported by the target accelerator (Algorithm 1) and (2) forming valid accelerator IR by translating and combining each \( \text{srDFG} \) node (Algorithm 2). Algorithm 1 is a function \( \text{Lower} \) which takes as input a dataflow graph \( \text{srdfg} \) which is a pair \( (N,E) \) of nodes \( N \) and edges \( E \) defined in Section III-A. PolyMath lowers \( \text{srDFG} \) operations to different domains with accelerator targets that support different granularity operations. \( \text{Lower} \) uses a map, \( \text{O}_\text{m} \), with domain names as keys, and lists of domain-specific accelerator operation names, \( \text{O}_d \), as values to lower \( \text{srDFG} \) nodes to the correct granularity.

The algorithm consists of first using the \( \text{srdfg}_d, \text{domain} \) attribute as a key to determine the correct granularity of operations for lowering, storing the set of supported operations in \( \text{O}_d \). For each node \( n \), if \( \text{name} \) is not included in \( \text{O}_d \), \( n, \text{srdfg} \) inherits the \( \text{srdfg}_d \) domain, and lowers \( n \) in \( \text{srdfg} \) by replacing it with a \( \text{srDFG} \) comprised of only supported operations. Replacing \( n \) in the \( \text{srdfg} \) consists of substituting \( \text{src} \) or \( \text{dst} \) in adjacent edges \( (\text{src}, \text{dst}, \text{md}) \) with a node in the \( \text{srDFG} \) if \( \text{src} = \text{dst} = n \). Once each node \( n \in N \) has been replaced with supported operations based on \( O_d \), by preserving different levels of granularity in the \( \text{srDFG} \), the same \( \text{srDFG} \) is capable of generating dfg’s with operations supported by a variety of custom accelerators. For instance, the hierarchy of RoBoX begins at the System level, followed by finer grained Task computations all the way down to varying operation granularities in its macro dataflow graph, such as Vector, Scalar, and Group operations.

Once a \( \text{srdfg} \) has been lowered, it is compiled to an IR suitable for the accelerator using Algorithm 2. Accelerator IR for a domain \( d \), denoted with \( \pi_d \), is comprised of accelerator IR fragments, each of which is a basic operator and its arguments. To generate each \( \pi_d \) for the different targets, Algorithm 2 takes as input a lowered \( \text{srDFG} \) produced by Algorithm 1 and accelerator specifications for the targets corresponding to each domain. Acceleration specifications for each domain are stored in \( \text{AccSpec} \), and define how \( \text{srDFG} \) nodes are translated and merged to form accelerator IR. A specification for domain \( d \) is a pair \( (m_d, +d) \) where:

- \( m_d \) is a map from operator names to a translation function for that operator. The translation functions, \( t \), works as follows: given a \( \text{srdfg} \) and a node \( n \), \( t(\text{srdfg}, n) \) returns the accelerator IR fragment \( \pi_d \) representing the accelerator operation for \( n \).
- \( +d \) is an operator that combines an accelerator IR \( \pi_d \) and an accelerator IR fragment produced by \( t_d \).
Having defined the necessary variables, we can describe Algorithm 2. The algorithm extracts the nodes and edges in the srDFG, then applies a translation function to each node, creating an accelerator IR fragment. Each IR fragment for each of the program’s domains is separately accumulated into complete representations of an accelerator program IR, and are returned by the algorithm.

The most complicated part of the compilation are the translation functions, t. The translation function does two things: (1) identify the correct accelerator IR operation, and (2) assign the correct arguments for that operator. Assigning the correct arguments uses these steps:
1) Convert types to the equivalent accelerator type
2) Use edges with input type modifier as input arguments
3) Use edges with output type modifier as output arguments
4) Initialize IR variables for edges with the state type modifier
5) Add constants for arguments with param type modifier
If the accelerator IR fragment requires the shape of operation arguments, it is also included as part of the arguments to the operator or declaration operation. To ensure data is transferred between domain boundaries, load and store IR fragments are created when there are sources and destinations with different domains than a node. As a final step, accelerator provided compilers are used to create binaries from the generated IR.

Compilation flexibility. The combination of Algorithm 2 and Table II enables compilation to different types of domain-specific accelerator because of two key properties. First, simultaneous access to each level of srDFG recursion allows supported accelerator operations to be translated. Unsupported srDFG nodes on the particular accelerator are refined and transformed to the appropriate level of granularity through recursion which enables identification of the accelerator-supported operations. Second, the metadata stored in the srDFG allows the IR generation to be parameterized based on the target accelerator. As a result, users can create different accelerator specifications for different accelerators and these same algorithms will do the appropriate mapping. Each algorithm can be instantiated for a number of different mappings without changes to the high-level algorithm.

V. EVALUATION

Table II illustrates the difference between conventional general-purpose stacks, domain-specific stacks in the literature, and PolyMath. As shown, PolyMath represent a middle ground between domain-specificity and generality, enabling cross-domain multi-acceleration.

A. Experimental Setup

1) Algorithms and Datasets.: Table III shows workloads from Robotics, Graph Analytics, DSP, Data Analytics, and Deep Learning domains and the lines of code (LOC) for the PMLang implementation. Table IV shows a break down of end-to-end application domains, algorithms, configurations, and PMLang LOC.

Single domain workloads. In Robotics domain, we have two benchmarks, MobileRobot [21] and Hexacopter [22]. Section II discusses the two-wheeled MobileRobot in detail. Hexacopter is a six-rotor micro UAV that uses motion planning and orientation control to determine trajectory. For both these workloads the physical robot and task specification is expressed in PMLang. For Data Analytics we have Low Rank Matrix Factorization (LRMF) and Kmeans clustering. LRMF converts a large matrix into two smaller matrices, which if taken product of, represent the original matrix. For LRMF we use two Movielens [23] datasets. Kmeans clustering partitions data into k-clusters. For one Kmeans workload cluster hand written digits with mnist [24] dataset. The second benchmark uses data from the UCI repository [25] to cluster households with similar electricity consumption. In the Digital Signal Processing domain we have four benchmarks, two for each Fast Fourier Transform (FFT), and Discrete Cosine Transform (DCT). The FFT implementation is a fine-grained butterfly and bit-reversal to transform a signal to frequency domain. The DCT algorithm applies a filter kernel to an input image and is used for compression. For Deep Learning, we use two popular convolutional neural networks, ResNet-18 [26] and MobileNet [27] for object classification. For Graph Analytics, we implement and apply Breadth-First Search on two graphs, one of Twitter users and followers [28] and another of Wikipedia links [29].

End-to-end cross-domain applications. Table V shows the end-to-end applications for our case study, the different domains they comprise of, and specification of each algorithm. The brain stimulation application, BrainStimul, is described in Section II. The stock market application, called OptionPricing, predicts call option price in stock market and uses two data analytics algorithms. This application first performs sentiment analysis through logistic regression on news articles to understand market signals and then Black-Scholes to predict the price.

2) Optimized CPU and GPU implementations.: Table VI and VII shows the optimized CPU and GPU framework their specifications. The robotics’ CPU implementation uses ACADO Toolkit [30] to implement optimized, self-contained C code and uses cuBLAS [31] libraries for GPUs. For graph analytics, we used Intel GraphMat [32] for CPU implementations and Enterprise [33] for GPU. The DSP workloads use

<table>
<thead>
<tr>
<th>Domain</th>
<th>Robotics</th>
<th>Graph Analytics</th>
<th>DSP</th>
<th>Data Analytics</th>
<th>Deep Learning</th>
<th>Genomics</th>
<th>SAT Solvers</th>
</tr>
</thead>
<tbody>
<tr>
<td>PMLang</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>Darwin</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>DNNWeaver</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>TVM</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>TABLA</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>RoboX</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>DeeCo7</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>BCP Acc</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>PolyMath</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
</tbody>
</table>
TABLE III: Benchmarks and workloads used to evaluate PolyMath.

<table>
<thead>
<tr>
<th>Domain</th>
<th>Benchmark</th>
<th>Algorithm</th>
<th>Config/Dataset</th>
<th>LOC</th>
</tr>
</thead>
<tbody>
<tr>
<td>Robotics</td>
<td>Mobile Robot</td>
<td>Model Predictive Control</td>
<td>Trajectory Tracking, Horizon = 1024</td>
<td>52</td>
</tr>
<tr>
<td></td>
<td>Hexacopter</td>
<td>Model Predictive Control</td>
<td>Attitude Control, Horizon = 1024</td>
<td>197</td>
</tr>
<tr>
<td>Graph Analytics</td>
<td>Twitter Followers</td>
<td>Breadth-First Search</td>
<td>#Vertices=61.57M, #Edges=1468.36M</td>
<td>14</td>
</tr>
<tr>
<td></td>
<td>Wikipedia Links</td>
<td>Breadth-First Search</td>
<td>#Vertices=3.56M, #Edges=84.75M</td>
<td>14</td>
</tr>
<tr>
<td></td>
<td>LiveJournal</td>
<td>Single Source Shortest Path</td>
<td>#Vertices=4.84M, #Edges=68.99M</td>
<td>14</td>
</tr>
</tbody>
</table>

TABLE IV: Algorithmic composition of end-to-end applications.

<table>
<thead>
<tr>
<th>Benchmark</th>
<th>Algorithm</th>
<th>Domain</th>
<th>Config/Dataset</th>
<th>LOC</th>
</tr>
</thead>
<tbody>
<tr>
<td>Fast-Fourier Transform (FFT)</td>
<td>DSP</td>
<td>1D FFT; 4096 real; 8192x1 input</td>
<td>12</td>
<td></td>
</tr>
<tr>
<td>Logistic Regression (LR)</td>
<td>Data Analytics</td>
<td>4096 features</td>
<td>8</td>
<td></td>
</tr>
<tr>
<td>Model Predictive Control (MPC)</td>
<td>Robotics</td>
<td>Horizon = 1024</td>
<td>64</td>
<td></td>
</tr>
<tr>
<td>Black-Scholes (BLKS)</td>
<td>Data Analytics</td>
<td>8192 options</td>
<td>10</td>
<td></td>
</tr>
<tr>
<td>Logistic Regression (LR)</td>
<td>Data Analytics</td>
<td>129549 words</td>
<td>8</td>
<td></td>
</tr>
</tbody>
</table>

TABLE V: Domains and accelerators used for evaluations.

<table>
<thead>
<tr>
<th>Domain</th>
<th>PolyMath Accelerator</th>
<th>Baseline Framework</th>
</tr>
</thead>
<tbody>
<tr>
<td>Robotics</td>
<td>RoBoX (ASIC)</td>
<td>ACADO/cuBLAS</td>
</tr>
<tr>
<td>Graph Analytics</td>
<td>GRAPHICIONADO (ASIC)</td>
<td>Intel GraphMat/Enterprise</td>
</tr>
<tr>
<td>Data Analytics</td>
<td>HyperStreams(FPGA) / TABLA (FPGA)</td>
<td>MLPack/OpenBlas/CUDA</td>
</tr>
<tr>
<td>DSP</td>
<td>DECO (FPGA)</td>
<td>FFTW3/cuFFT/NVIDIA-DCT</td>
</tr>
<tr>
<td>Deep Learning</td>
<td>TVM-VTA (FPGA)</td>
<td>TVM/Tensorflow</td>
</tr>
</tbody>
</table>

C subroutine libraries [34, 35] for CPU and Nvidia implementations [36, 37] for GPU. For data analytics, we use mlpack [38], a fast and flexible C++ ML library built on top of OpenBLAS [39], NVBLAS [40], and Armadillo [41]. The Deep Learning workloads are compiled using optimized Tensorflow [42] for CPU and GPU. All of our experiments were performed on an Intel Xeon E7, Titan Xp GPU, and low-power Jetson Xavier AGX.

3) Domain-Specific accelerators: Table V shows the accelerator used for each domain. To evaluate each benchmark on domain-specific accelerators, PolyMath was used to compile programs to the target accelerator IR, and the target accelerator’s compiler was used to generate executable binaries. RoBoX’s [15] Macro DFG from srDFG as it offers programmable ASIC for system, tasks, and penalties for control algorithms optimized using MPC. We use GRAPHICIONADO [14], an ASIC accelerator for graph analytics algorithms expressed as vertex programs, as the target for the Graph Analytics workloads. We compile FFT and DCT srDFG representation to DECO [7], a DSP block based FPGA accelerator, by translating to it’s DFG, which is then compiled as executable binaries. For LRMF and kmeans we convert srDFG to TABLA [43], an open source template-based FPGA accelerator for machine learning, by compiling a srDFG to a DFG. We use TVM-VTA [44], a programmable deep learning FPGA accelerator, as the target for Deep Learning workloads, as it is state-of-the-art and open-source. Each DSA requires specific levels of operation granularity: single operation [7, 17], coarse DNN layers [44], and coarse time snapshots [15], enabled by each srDFG’s multi granularity, for mapping of kernels to the accelerator.

Multi-acceleration. For BrainStimul, we compile parts to DECO [7] (FFT-4096), TABLA [17] (Logistic Regression), and RoBoX [15] accelerators. For OptionPricing, we execute logistic regression based sentiment analysis on TABLA [17] and Black Scholes on HyperStreams [45]. All accelerators are cascaded as a single System On Chip (SOC), comprised of memory and a host. A light-weight manager executes on the host, ensuring data dependencies between different accelerators and initiating DMA transfers between DRAM and local accelerator memory. This setting is similar to prior work [46] that also uses an array of micro-accelerators.
B. Experimental Results

The goal of PolyMath is to facilitate the use of the wide variety of custom accelerators across end-to-end cross domain applications. It does so by abstracting away hardware level details through a versatile, extensible, and a modular stack that can maintains the required levels of kernel granularities best suited for each design. In this section we compare domain-specific accelerators executing PMLang code with optimized CPU and GPU implementations to better understand the portability of PolyMath. We then perform case studies through two end-to-end applications and observe that PolyMath allows cross domain multi-acceleration.

1) Performance and Energy Comparisons: Single kernel comparisons. Figure 7 and Figure 8 show the speedup of PolyMath compiled programs listed in Table III to domain specific accelerators over Xeon E-2176G CPU, Titan Xp GPU, and Jetson Xavier AGX as the baseline, respectively. On average, PolyMath translated implementations outperform Xeon E-2176G and Jetson Xavier by 3.3× and 1.2× in terms of runtime and offer 7.2× and 1.7× more Performance-per-Watt over Titan Xp and Jetson Xavier GPUs. Smaller benchmarks such as MovieLens-100K and ElecUse are unable to fully utilize Titan Xp, thus cannot obtain higher benefits in comparison to Jetson but incur higher PPW. The average still demonstrates an increase in both performance and energy efficiency in the cross-domain setting. PolyMath implementations also offer 18.1× more energy efficiency over CPUs, but due to many lower power accelerators. This is especially true for discrete cosine transform and deep learning benchmarks; DCT due to its high coarse granular matrix multiplications for which DeCO a programmable FPGA accelerator is not as effective as Titan Xp and deep learning models because our backend for CNNs is VTA that is designed as a low-power accelerator but is also being compared to a high-end GPU for uniformity. Note that PolyMath does not contribute any overhead specifically for deep learning acceleration because it offers direct conversion of sP+DFG to the TVM nodes.

Optimal performance comparison. The accelerators used for execution of PolyMath implementations also offer custom stacks that are built for their target architecture. We compare PolyMath to implementations in their native stack, which represent optimal executions to demonstrate the cross-domain overhead of our stack to native implementations. Figure 9 compares the performance of PolyMath implementations with the optimal performance that can be reached by programs written by experts for each of the accelerators. The figure shows that PolyMath achieves 83.9% the optimal runtime.

The performance of PolyMath relative to optimal implementations is dependent on the domain and the algorithm because each stack differs in the level of data semantics that can be complex to compile to from a more expressive language. For instance, accelerators specializing in Deep Learning often utilize three primary type modifiers for variables in neural network graphs: input, output, and weights—each of which can be directly mapped to type modifiers in PMLang, thus incur zero overhead. In contrast, Robotics algorithms contain unique data semantics, such as task penalties, constraints variables, time varying references, etc., which do not differentiated with PolyMath type modifiers, thus implementations do not reach optimal performance. Accelerators such as DeCO require specific topologies for their graph-based IR, i.e. balanced DFGs, because they rely on stage-based computation, which results in reduced execution time relative to PolyMath translations. In the case of Data Analytics, we see a low percentage of optimal performance for ElecUse, because the benchmark is small, which makes any extra operations included in the sP+DFG have a more significant impact on performance relative to the optimal implementation. In contrast, DigitCluster uses the same algorithm but on a larger dataset, thus can amortize the overhead more effectively. It is important to note that the optimal performance is reached by a specialized program.
on each accelerator written by an expert, whereas PolyMath offers support for multiple domains constituting end-to-end applications that can be expressed as single comprehensive program, coupled with sDFG to pave way for compilation to multiple accelerators.

2) End-to-End Application Case Study: PolyMath offers means to express cross domain applications as a single program which can be compiled to multiple accelerators pertaining to each of these domains. Figure 10 shows the runtime and energy improvement of the end to end applications in comparison to CPU. Figure 11 illustrates the runtime and performance-per-Watt improvement in comparison to Titan Xp and Jetson. Figure 10a and Figure 11a shows these results for BrainStimul and Figure 10b and Figure 11b for OptionPricing application. In these graphs we provide entire application improvement for all possible acceleration combinations, from one domain algorithm accelerated to cross-domain where all algorithms are accelerated. Each end-to-end result incorporates data communication overheads from data transfer between hardware. Stand-alone kernel acceleration, as shown in , can offer very high speedups. However, when these kernels are incorporated within a more comprehensive application, those speedups do not manifest in the entire application because the non-accelerated kernel becomes a bottleneck. For instance, the gap between the highest benefit obtained from the best single-domain acceleration and cross-domain end-to-end acceleration, is 1.85× for BrainStimul and 2.06× for OptionPricing (Figure 10). Every kernel that is added for acceleration not only benefits itself from specialized execution but also reduces the Amdahl’s burden and magnifies other accelerated component’s impact. The benefits of individual kernel acceleration are present despite end-to-end communication runtime overheads of 23.4% and 17.0% and energy overheads of 21.8% and 12.4% for BrainStimul and OptionPricing, respectively.

Figure 11a and 11b show the BrainStimul and OptionPricing results for both Titan Xp and Jetson. Figure 11a shows that PolyMath offers 1.2× runtime improvement over Titan Xp compared to 1.8× over Jetson. In contrast, PolyMath improves performance-per-watt by 8.3× over Titan Xp compared to 2.8× for Jetson due to its lower power consumption. As Figure 11b shows, the OptionPricing benchmark underutilizes the Titan Xp, only offering 1.5× and 9.2× improvement in performance and performance-per-watt compared to 1.4× and 1.9× over Jetson. This is caused by the difference in levels of coarse parallelism in the algorithms. Overall, the PolyMath implementation of OptionPricing still outperforms both GPUs for both runtime and performance-per-watt.

Lastly, Figure 12 shows end-to-end Polymath implementations achieve 76.7% optimal performance for BrainStim and 76.9% for OptionPricing compared to entirely manual implementations of each application. Given the fact that PolyMath offers greater ease of programming compared to Python (Figure 13), the automation overhead of 23.1% is a fair tradeoff.
3) **User Study:** To determine the usability of PMLang, we conducted a user study with 20 programmers who are either professional software engineers or PhD students in computer science. The goal of the user study is to measure the expressiveness of PMLang by comparing it to Python, an intuitive programming language which is commonly used in three of the focus domains of PolyMath: DSP, Data Analytics, and Deep Learning. The study tasked each user in the study with implementing either a DSP or Analytics algorithm in Python or PMLang. The users were divided into four groups; Kmeans in Python, Kmeans in PMLang, DCT in Python, and DCT in PMLang. For fairness and ease in expression of tensor algebraic operations, we allow the users to import Python modules such as numpy [47]. Each participant in the user study has varying levels of expertise (from beginner to proficient) in the target domains, and is proficient in Python. Every participant went through the following process:

1) Participants were introduced to PMLang with a short, six-minute video which walked through the language and small examples.
2) To avoid any algorithm knowledge bias, participants were randomly assigned either the DSP or ML algorithm to implement in either Python and PMLang.
3) To minimize variation in algorithm understanding, users were not allowed to begin their first implementation before having read and confirmed their understanding of the algorithm.
4) We timed participants during their implementations and measured their Lines of Code (LOC) after completion.

**Results.** Figure 13 compares the LOC between Python and PMLang as a ratio of Python LOC to PMLang LOC in (a), and the implementation time of Python and PMLang as a ratio of Python implementation time to PMLang implementation time in (b). The results show that PMLang required 2.5× fewer lines of code on average and 1.9× less implementation time on average. The Kmeans implementation averaged 3.3× fewer LOC, whereas for DCT the average reduction of LOC is 1.8×. In general, the Kmeans algorithm is more verbose than DCT and on average required 47.6% more lines of Python code than DCT. Because there were more lines of code included in Kmeans, there was more opportunity to reduce multi-line operations to a single PMLang statement, which explains the difference in LOC reduction.

The greater complexity of Kmeans appears to have an effect in the speedup of implementation time as well, where the average speedup for Kmeans was 2.6× and the average speedup for DCT was 1.2×. Part of this speedup can be attributed to typing more LOC in PMLang, but it is also a result of being able to directly translate mathematical notation to the equivalent PMLang statement. These results indicate that the more complicated the mathematical program is, the more the programmer will benefit from implementing the program in PMLang. Further, PMLang is expressive enough for programmers unfamiliar with the language to write algebraic expressions more efficiently than they would write the same expressions in a language they are familiar with, Python.

VI. **Related Work**

**DSLs for custom architectures.** There are various domain-specific languages designed to facilitate the use of hardware accelerators. These languages are mostly designed for a single domain [48–51] or like Spatial [52], they focus on conveniently expressing lower level hardware-centric information. Another language, Halide [50], allows expression of image processing pipelines and contains constructs for filter-based algorithms. Lime [53] focuses on high-level synthesis from Java, thereby enabling execution of Java programs on both FPGAs and CPUs. Instead, PolyMath offers a Cross-Domain Language (CDL) and compute stack to explore the emerging tradeoff between expressiveness and performance while leveraging currently isolated, domain-specific accelerators.

**Mathematical and scientific computing environments.** There are numerous scientific and numerical programming environments [54–57], and frameworks [47, 58]. PolyMath uniquely provides a 1-to-1 mapping of mathematical expressions to its statements and leverages the natural parallelism in formulas without any explicit annotations for vectorization. In contrast, MATLAB [54], Julia [55], or R, require manual effort from the user to identify the parallelism across different computations, vectorize its code, and determine column/row arrangements for matrix operations. Moreover, these languages do not delineate between the semantics of data in their programs and do not offer a multi-granular representation, as offered by PolyMath, to enable usage of various domain-specific accelerators.

**Intermediate Representations.** A number of intermediate representations [59, 60] provide abstractions to enable program analysis using virtual resources. Both LLVM and JVM operate at the granularity of a single CPU instruction, which
is highly inefficient for domain-specific architectures. Some works \cite{61} have adapted LLVM to guarantee independence between parallel operation threads by using a dataflow graph structure intended for heterogeneous platforms. A number of other works \cite{16, 62, 66} focus on the domain of machine learning and have implemented an end-to-end approach for optimization on heterogeneous platforms after performing optimizations from a high-level language. These works supported limited algorithm domains \cite{16, 62-63}, and rely on C/C++ or other general-purpose programming languages \cite{61, 66}, requiring the programmer to express complex mathematical expressions in unintuitive ways. MLIR \cite{66} is a hierarchical, high level IR, but is general-purpose and as such is on the end of the expressiveness curve (Figure \ref{fig:expressiveness}). Whereas PolyMath is restricted by its mathematical language (PMLang) to a limited set of domains, falling in the middle of the spectrum of expressiveness. Furthermore, MLIR does not have any compilation stack to support variety of accelerators from different domains as PolyMath does and is practically demonstrated in the evaluations. Tiramisu \cite{67} introduces a scheduling language with novel commands to explicitly manage the complexities that arise when targeting multicores, GPUs, and distributed machines. Tiramisu offers an IR based on the polyhedral model to allow fine-grained optimization. As such, Tiramisu can serve as a potential backend for PolyMath that deals with the higher-level complexity of expressing cross-domain application and not low-level fine-grained optimization.

**Acceleration frameworks and toolchains.** TensorFlow \cite{42} is an end-to-end open source platform for expressing ML algorithms in Python. Deep learning accelerators (e.g., TPU \cite{13}) leverage Tensorflow. Similarly, a variety of deep learning frameworks \cite{19, 68, 69} allow users to run their DNNs on FPGA based hardware designs. Full stack solutions such as TABLA \cite{17} and RoBox \cite{15} support classical supervised machine learning and model predictive control in robotics, respectively. Other toolchains \cite{70, 73} aim to simplify running deep neural networks on hardware accelerators by performing design space exploration to find the best configuration for their particular design. These solutions, however, are bound to their own custom architectures for particular platforms (FPGAs or ASICs). In contrast, the sDFG offers a flexible hook that can be translated to these toolchains and frameworks as well as to future accelerator designs and platforms. The cross-domain nature of PolyMath that supports Robotics, Graph Analytics, DSP, Data Analytics and Deep learning sets it apart from these domain-specific stacks.

**VII. CONCLUSION**

As domain-specific accelerators are becoming prevalent, there is an emerging tradeoff between expressiveness and performance. This paradigm—a pendulum swing from general-purpose processing to the opposite direction—creates implicit programming silos between different domain. This paper set out to explore the region between these extremes and explore the new expressiveness-performance tradeoff. To that end, we defined a cross-domain computational stack, Poly-Math, that bridges the expressiveness gap between multiple domains, Robotics, Graph Analytics, DSP, Deep Learning, and Data Analytics. The results from user study and performance evaluations showed that PolyMath strikes an effective balance between expressiveness and performance while enabling cross-domain multi-acceleration. It is time to look beyond the timely, yet temporary, success of domain-specific accelerators and devise a future that enables end-to-end applications. The current approach towards acceleration excludes significant opportunities by restricting the domain. To harness these untapped opportunities, a new paradigm needs to emerge that breaks the boundaries of domains, but also preserves the benefits of domain-specificity. PolyMath takes the initial step in breaking this new ground.

**VIII. ACKNOWLEDGMENTS**

This work was in part supported by generous gifts from Qualcomm, Google, Microsoft, Xilinx, Leidos as well as the National Science Foundation (NSF) awards CNS#1703812, ECCS#1609823, CCF#1553192, Air Force Office of Scientific Research (AFOSR) Young Investigator Program (YIP) award #FA9550-17-1-0274, National Institute of Health (NIH) award #R01EB028350, and AirForce Research Laboratory (AFRL) and Defense Advanced Research Project Agency (DARPA) under agreement number #FA8650-20-2-7009 and #HR0011-18-C-0020. The U.S. Government is authorized to reproduce and distribute reprints for Governmental purposes not withstanding any copyright notation thereon. The views and conclusions contained herein are those of the authors and should not be interpreted as necessarily representing the official policies or endorsements, either expressed or implied of Qualcomm, Google, Microsoft, Xilinx, NSF, AFOSR, NIH, AFRL, DARPA or the U.S. Government.

**REFERENCES**


Nadav Rotem, Jordan Fix, Saleem Abdulrasool, Summer Deng, Roman Dzhabarov, James Hegeman, Roman Levenstein, Bert Maher, Nadathur Satish, Jakob Olsen, Jongsoo Park, Artem Rakhov, and Misha Smelyanskiy. Glow: Graph lowering compiler techniques for


