Automatically Harnessing Sparse Acceleration

Automatically Harnessing Sparse Acceleration

Philip Ginsbach
School of Informatics
The University of Edinburgh
United Kingdom
philip.ginsbach@ed.ac.uk

Bruce Collie
School of Informatics
The University of Edinburgh
United Kingdom
bruce.collie@ed.ac.uk

Michael F. P. O'Boyle
School of Informatics
The University of Edinburgh
United Kingdom
mob@ed.ac.uk

Abstract

Sparse linear algebra is central to many scientific programs, yet compilers fail to optimize it well. High-performance libraries are available, but adoption costs are significant. Moreover, libraries tie programs into vendor-specific software and hardware ecosystems, creating non-portable code.

In this paper, we develop a new approach based on our *specification Language for implementers of Linear Algebra Computations* (LiLAC). Rather than requiring the application developer to (re)write every program for a given library, the burden is shifted to a *one-off* description by the library implementer. The LiLAC-enabled compiler uses this to insert appropriate library routines without source code changes.

LiLAC provides automatic data marshaling, maintaining state between calls and minimizing data transfers. Appropriate places for library insertion are detected in compiler intermediate representation, independent of source languages.

We evaluated on large-scale scientific applications written in FORTRAN; standard C/C++ and FORTRAN benchmarks; and C++ graph analytics kernels. Across heterogeneous platforms, applications and data sets we show speedups of 1.1× to over 10× without user intervention.

CCS Concepts • **Software and its engineering → Compilers**; **Specification languages**.

Keywords sparse linear algebra, domain specific languages, library integration, declarative langauges, data marshalling

ACM Reference Format:
Philip Ginsbach, Bruce Collie, and Michael F. P. O'Boyle. 2020. Automatically Harnessing Sparse Acceleration. In *Proceedings of the 29th International Conference on Compiler Construction (CC '20), February 22–23, 2020, San Diego, CA, USA.* ACM, New York, NY, USA, 12 pages. https://doi.org/10.1145/3377555.3377893

CC '20, February 22–23, 2020, San Diego, CA, USA

ACM ISBN 978-1-4503-7120-9/20/02…$15.00
https://doi.org/10.1145/3377555.3377893

1 Introduction

Linear algebra is an important component of many applications and a prime candidate for hardware acceleration. While there has been significant compiler effort in accelerating dense algebra [23, 36, 40], there has been less success with sparse codes. This is largely due to indirect memory access, which challenges compiler analysis [32]. Sparse-based algorithms are, however, increasingly important as the basis of graph algorithms and data analytics [28].

We currently see the wide-scale provision of fast sparse libraries [2, 3, 5, 55]. They deliver excellent performance, but require significant programmer intervention and are rarely portable across platforms. Alternatives, such as the SLinGen/LGen system [45, 46], provide specialized code generators for linear algebra, but again require code modification by the programmer and focus only on dense computations.

Program modification is particularly problematic when the targets are hardware accelerators that require careful data marshaling. Such modifications are often program-wide and severely reduce the portability of the program. Furthermore, they require a commitment to specific hardware vendors, resulting in codebases that quickly become obsolete. In order to mitigate this, many projects have to keep multiple execution paths, resulting in arcane build systems and unmaintainable code. In this time of rapid hardware innovation, such a vendor lock-in is undesirable. In fact, the difficulty of efficient portable integration is a key impediment to the wider use of accelerator libraries and hardware.

In this paper, we reexamine how compilers and libraries can be used to achieve performance without programmer effort. Highly tuned and platform specific-libraries invariably remain the fastest implementations available. However, we show that we can automatically integrate these libraries without polluting the source code. This is performed as a compiler transformation step, leaving the original source code intact and portable.

To achieve this, we develop a new specification language for implementers of libraries, the *specification Language for implementers of Linear Algebra Computations* (LiLAC). Using LiLAC, library implementers specify with a few lines of code, *what* a library does and *how* it is invoked. Our compiler then determines where the library specification matches user code and automatically transforms it to utilize the library. The language has two complementing parts.

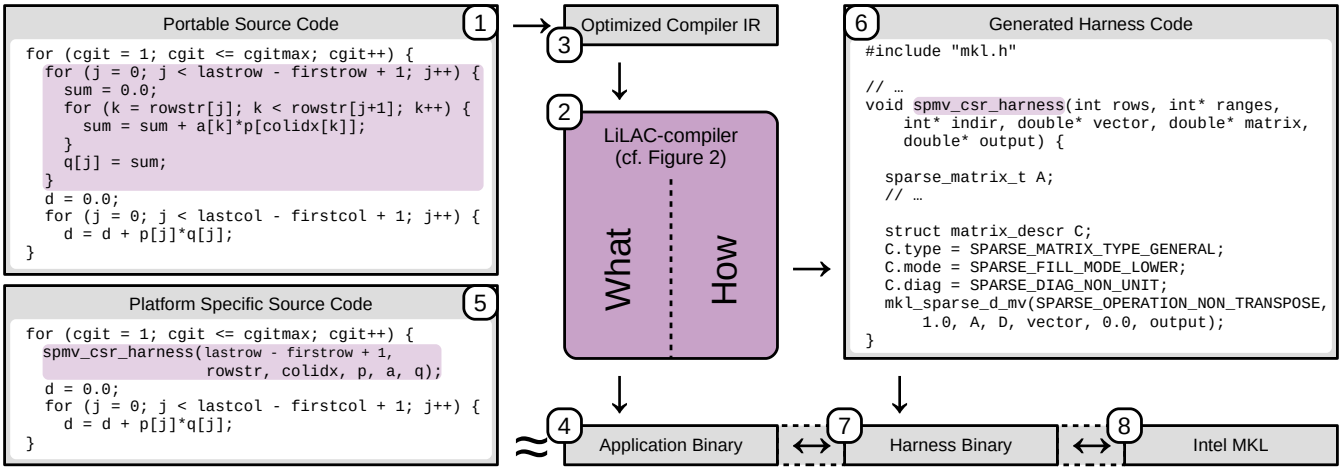


Figure 1. LiLAC applied to NPB Conjugate Gradient: Code (1) that matches the LiLAC-What specification (cf. Figure 2) is replaced by calls to a harness (5) during compilation (2), resulting in an application binary (6) that corresponds to (hypothetical) platform-specific source code (4). The harness is generated from the LiLAC-How specification (cf. Figure 2) to utilize Intel MKL.

LiLAC-What is a high-level language to describe sparse and dense linear algebra computations. The LiLAC compiler uses it to detect such functionality in user applications at compiler intermediate representation level. It is powerful enough to formulate linear algebra routines, yet remains independent of compiler internals and is easy to understand and program. **LiLAC-How** specifies how libraries can be used to perform a LiLAC-What-specified computation. Besides generating setup code and handling hardware context management, it crucially enables efficient memory synchronization. It uses memory protection mechanisms to automatically track data changes and transfers memory only when necessary.

The research contribution of this paper is a combination of three techniques for the acceleration of sparse linear algebra:

- Accelerate unchanged source code by identifying sparse linear algebra computations with backtracking search.
- Avoid vendor lock-in with an extensible specification language that adapts to new accelerator libraries.
- Achieve program-wide memory synchronization with only local transformations using memory protection.

Together, these techniques result in a system that works on existing and novel software. It offers the full performance of fast libraries, avoids vendor lock-in, and keeps the source code easy to maintain and free from pollution.

2 Overview

Figure 1 shows the LiLAC-enabled compiler from the user perspective. In the top left corner (1), we see unmodified application source code. This is *conjugate gradient* from the NAS-PB suite. To achieve good performance on Intel processors, the compiler (2) has been configured to offload native sparse code to Intel MKL. Using a specification of *What* computations MKL supports, it recognizes the highlighted loop as a suitable sparse matrix-vector product. Instead of passing it on to the compiler backend for code generation, it inserts a call to a *harness* function. This is performed on intermediate code (3) and results in a program (4). In the bottom left (5) is an equivalent source-level representation.

LiLAC also generates the corresponding harness code (6), which gets compiled into a shared library (7) that is linked with the application binary. This harness interfaces with the underlying library implementation, Intel MKL (8).

2.1 Implementation Overview

Figure 2 shows the internals of the LiLAC system. It is fully integrated into the build system of the established LLVM compiler framework, extending the clang compiler.

On the left is the LiLAC specification - just 16 lines of code. It is independent of the user application and can be provided by the library implementer. It consists of a *What* and a *How* part. These two parts are processed by the LiLAC system and result in a runtime library and a generated detection function, which is incorporated into the clang compiler.

LiLAC-What specifies the functionality that is provided by a library, in this example *spmv-csr* (cf. Figure 2). From this, a function that detects the computation in normalized LLVM IR code is generated and the harness interface is determined. The detection functions are based on a backtracking search algorithm, as elaborated in section 4. The detection function is linked directly into the LiLAC-compiler, either statically or dynamically at (compiler) run time.

LiLAC-How specifies how the library, Intel MKL in this case, is invoked to perform the specified calculation. This involves boilerplate code, but also advanced features. These include efficient data synchronization and the caching of invariants. In the given example, the *columns* variable is such an invariant. It is required for the library call, but not

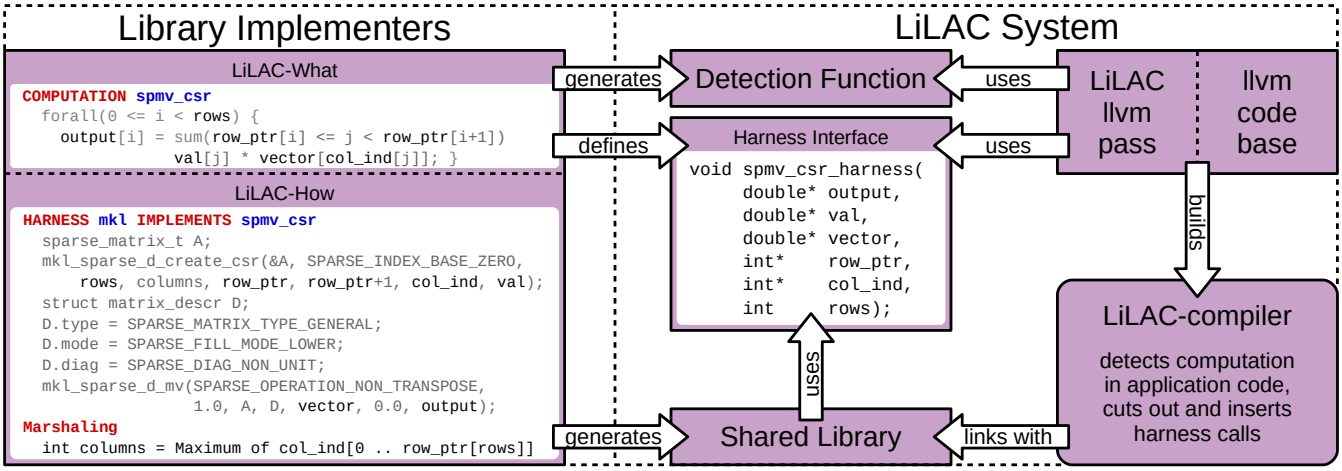


Figure 2. Overview of LiLAC internals: On the left is the complete LiLAC program that the library implementer has to provide. At compile time of LLVM, this program is parsed and incorporated into a modified clang C++ compiler, behaving as in Figure 1.

statically available. Therefore, it has to be computed at runtime. Using *Marshaling*, LiLAC automatically generates the harness such that this is only recomputed if the values in *row_ptr* change. Such changes are captured with generated memory protection code using *mprotect*, managed by LiLAC.

On the right of the figure, we can see how the components generated from the LiLAC specification are used to build the LiLAC-compiler. The detection function is compiled and used directly by the LiLAC-Compiler, linked either statically or dynamically. Interacting with the internals of LLVM, it implements a transformation pass that is executed after the normal optimization pipeline. Using the generated detection function, it finds instances of the computation and replaces them with calls to the specified harness interface.

The harness, on the other hand, is compiled into a shared library. The LiLAC-compiler dynamically links applications to this shared library whenever it inserts harness calls. When multiple LiLAC-How programs are provided, the generated harnesses are compatible and linking the user program to a different harness library at runtime is sufficient.

3 What and How

This section describes in more detail the two components of the LiLAC language. LiLAC-What specifies the computations that a library performs; LiLAC-How describes how exactly the library should be invoked to perform these computations.

3.1 LiLAC-What: Functional Description

At the heart of our approach is a simple language to specify sparse and dense linear algebra operations. This serves two purposes in our LiLAC system: Firstly, it is used to generate a detection program for finding the computation in user code. Secondly, it identifies the variables that are arguments to the library, thus defining the harness interface.

$$
\begin{aligned}
\mathit{program} &::= \textbf{COMPUTATION}\ \langle \mathit{name}\rangle\ \langle \mathit{body}\rangle \\
\mathit{body} &::= \langle \mathit{forall}\rangle \mid \langle \mathit{dotop}\rangle \\
\mathit{range} &::= (\ \langle \mathit{exp}\rangle <= \langle \mathit{name}\rangle < \langle \mathit{exp}\rangle\ ) \\
\mathit{forall} &::= \textbf{forall}\ \langle \mathit{range}\rangle\ \{\ \langle \mathit{body}\rangle\ \} \\
\mathit{dotp} &::= \langle \mathit{addr}\rangle = \textbf{dot}\ \langle \mathit{range}\rangle\ \langle \mathit{addr}\rangle\ *\ \langle \mathit{addr}\rangle\ ; \\
\mathit{addr} &::= \langle \mathit{name}\rangle\ \{\ [\ \langle \mathit{exp}\rangle\ ]\ \} \\
\mathit{add} &::= \langle \mathit{exp}\rangle + \langle \mathit{exp}\rangle \\
\mathit{mul} &::= \langle \mathit{exp}\rangle\ *\ \langle \mathit{exp}\rangle \\
\mathit{exp} &::= \langle \mathit{name}\rangle \mid \langle \mathit{cnst}\rangle \mid \langle \mathit{addr}\rangle \mid \langle \mathit{add}\rangle \mid \langle \mathit{mul}\rangle
\end{aligned}
$$

Figure 3. Grammar of the LiLAC-What language

$$
\begin{bmatrix} 1 & 0 & 1 & 0 & 0 \\ 0 & 2 & 0 & 2 & 0 \\ 0 & -1 & 3 & 2 & 0 \\ 0 & 0 & 0 & 2 & 0 \\ 0 & 0 & -1 & 0 & 1 \end{bmatrix}
$$

$$
\begin{aligned}
\textbf{val} &= \begin{bmatrix} 1 & 1 & 2 & 2 & -1 & 3 & 2 & 2 & -1 & 1 \end{bmatrix} \\
\textbf{col\_ind} &= \begin{bmatrix} 0 & 2 & 1 & 3 & 1 & 2 & 3 & 3 & 2 & 4 \end{bmatrix} \\
\textbf{row\_ptr} &= \begin{bmatrix} 0 & 2 & 4 & 7 & 8 & 10 \end{bmatrix}
\end{aligned}
$$

Figure 4. Compressed Sparse Row (CSR) representation as used by the LiLAC-What example in Figure 1 and Figure 2

$$
\begin{bmatrix} 0 & -1 & 3 & 2 & 0 \\ 1 & 0 & 1 & 0 & 0 \\ 0 & 2 & 0 & 2 & 0 \\ 0 & 0 & -1 & 0 & 1 \\ 0 & 0 & 0 & 2 & 0 \end{bmatrix}
$$

$$
\begin{aligned}
\textbf{perm} &= \begin{bmatrix} 1 & 2 & 0 & 4 & 3 \end{bmatrix} \\
\textbf{val} &= \begin{bmatrix} -1 & 1 & 2 & -1 & 2 & 3 & 1 & 2 & 1 & 2 \end{bmatrix} \\
\textbf{col\_ind} &= \begin{bmatrix} 1 & 0 & 1 & 2 & 3 & 2 & 2 & 3 & 4 & 3 \end{bmatrix} \\
\textbf{jd\_ptr} &= \begin{bmatrix} 0 & 5 & 9 & 10 \end{bmatrix} \\
\textbf{nzcnt} &= \begin{bmatrix} 3 & 2 & 2 & 2 & 1 \end{bmatrix}
\end{aligned}
$$

```
COMPUTATION spmv_jds
  forall(0 <= i < rows) {
    output[perm[i]] = sum(0 <= j < nzcnt[i])
      val[jd_ptr[j]+i] * vector[col_ind[jd_ptr[j]+i]]; }
```

Figure 5. Jagged Diagonal Storage (JDS) in LiLAC-What

The key design challenge was to stay simple enough to automatically generate robust detection functionality, yet to be able to capture operations in all relevant data formats. Most importantly, this includes the CSR/CSC, JDS and COO formats. CSR and JDS are part of our evaluation. Across the different formats, the control flow is rigid and easy to express. This is reflected in the grammar as shown in Figure 3.

3.2 Sparse Matrix Variations in LiLAC-What

Sparse matrices can be stored in different formats. We introduce two of them explicitly, but others are supported in the same way by LiLAC-What.

Compressed Sparse Row (CSR) [44] All non-zero entries are stored in a flat array **val**. The **col_ind** array stores the column position for each value. Finally, the **row_ptr** array stores the beginning of each row of the matrix as an offset into the other two arrays. The number of rows in the matrix is given directly by the length of the **row_ptr** array minus one, however, the number of columns is not explicitly stored. In Figure 4, a 5x5 matrix is shown represented in this format, the LiLAC-What code is in the top left of Figure 2.

Jagged Diagonal Storage (JDS) [43] The matrix rows are reordered such that the number of non-zeros per row is decreasing. The permutation is stored in a vector **perm**, the number of nonzeros in **nzcnt**. The nonzero entries are then stored in an array **val** in the following order: The first nonzero entry in each row, then the second nonzero entry in each row etc. The array **col_ind** stores the column for each of the values and **jd_ptr** stores offsets into **val** and **col_idx**. The product of a sparse matrix in JDS format with a dense vector is specified in LiLAC-What at the bottom of Figure 5.

Dense Detecting dense is easier than sparse, and existing literature covers it well. We fully support dense but evaluate it only briefly for completeness.

3.3 LiLAC-How

Where LiLAC-What specifies the computations implemented by a library, LiLAC-How describes how precisely library calls can be used to perform them. The language was designed to support important existing libraries such as cuSPARSE, clBLAS, and Intel MKL. The idiosyncrasies of these libraries require LiLAC-How to capture some boilerplate C++ code that manages the construction of parameter structures, calling conventions etc. Aside from this aspect, we designed it as high-level as possible without compromising performance. In particular, LiLAC-How abstracts away memory transfers.

These considerations result in two interacting components. Firstly, a *harness* describes the boilerplate code for individual library invocations. Secondly, data *marshaling* between the core program and the library is specified, which is crucial for heterogeneous compute environments. Figure 6 shows the grammar specification of LiLAC-How.

$$
\begin{aligned}
\mathit{harness} ::= {} & \textbf{HARNESS}\ \langle \mathit{name} \rangle\ \textbf{IMPLEMENTS}\ \langle \mathit{name} \rangle \\
& \langle C++ \rangle \\
& [\ \langle \mathit{marshaling} \rangle\ ]\ [\ \langle \mathit{persistence} \rangle] \\
& [\ \textbf{CppHeaderFiles}\ \{\ \langle \mathit{name} \rangle\ \}\ ] \\
\mathit{persistence} ::= {} & \textbf{PersistentVariables}\ \{\ \langle \mathit{name} \rangle\ \langle \mathit{name} \rangle\ \} \\
& [\ \textbf{BeforeFirstExecution}\ \langle C++ \rangle\ ] \\
& [\ \textbf{AfterLastExecution}\ \langle C++ \rangle\ ] \\
\hline
\mathit{marshalling} ::= {} & \textbf{Marshaling} \\
& \{\ \langle \mathit{type} \rangle\ \langle \mathit{name} \rangle = \langle \mathit{name} \rangle\ \textbf{of} \\
& \langle \mathit{name} \rangle\ \ [\ \textbf{0}\ ..\ \langle \mathit{exp} \rangle\ ]\ \} \\
\mathit{input} ::= {} & \textbf{INPUT}\ \langle \mathit{name} \rangle\ \langle C++ \rangle \\
& [\ \textbf{BeforeFirstExecution}\ \langle C++ \rangle\ ] \\
& [\ \textbf{AfterLastExecution}\ \langle C++ \rangle\ ] \\
\mathit{output} ::= {} & \textbf{OUTPUT}\ \langle \mathit{name} \rangle\ \langle C++ \rangle \\
& [\ \textbf{BeforeFirstExecution}\ \langle C++ \rangle\ ] \\
& [\ \textbf{AfterLastExecution}\ \langle C++ \rangle\ ]
\end{aligned}
$$

Figure 6. Grammar of LiLAC-How

3.3.1 Individual Library Invocations

We need to encapsulate the boilerplate code that any given library requires, such as setup code, filling of parameter structures etc. This part of the language is straightforward.

Harness The harness construct is the central way of telling the LiLAC system how a library can be used to perform a computation that was specified in LiLAC-What. As we can see at the top of Figure 6, a harness refers to a LiLAC-What program by name and also has a name itself. It is built around some C++ code, which can use all the variables from the LiLAC-What program to connect with the surrounding program. It also needs to specify the relevant C++ header files that the underlying library requires. Lastly, the harness can incorporate persistent state and utilize data marshaling.

Persistence Many libraries need setup and cleanup code, which is specified with the keywords *BeforeFirstExecution* and *AfterLastExecution*. These are used in combination with *PersistentVariables*, allowing state to persist between harness invocations, e.g. to retain handlers to hardware accelerators.

Example In Figure 7, we see a trivial LiLAC-What program for implementing `spmv_csr` with the Intel MKL library. The actual call to the relevant library function is in line 16. To prepare for that call, there is boilerplate code in lines 7–14 to fill parameter structures.

Critically, there is an additional parameter required by the library that is data-dependent: the number of columns, *cols*, in the sparse matrix. It is determined at runtime, in lines 2–5, leading to reduced performance. We will avoid this with the data marshaling constructs in the next section.

```
1  HARNESS mkl IMPLEMENTS spmv_csr
2    int cols = 0;
3    for(int i = 1; i < rowstr[rows]; i++)
4      cols = colidx[i]>cols?colidx[i]:cols;
5    cols = cols+1;
6
7    sparse_matrix_t A;
8    mkl_sparse_d_create_csr(&A, SPARSE_INDEX_BASE_ZERO,
9                            rows, cols, rowstr,
10                           rowstr+1, colidx, a);
11   struct matrix_descr dscr;
12   dscr.type = SPARSE_MATRIX_TYPE_GENERAL;
13   dscr.mode = SPARSE_FILL_MODE_LOWER;
14   dscr.diag = SPARSE_DIAG_NON_UNIT;
15
16   mkl_sparse_d_mv(SPARSE_OPERATION_NON_TRANSPOSE,
17                   1.0, A, dscr, iv, 0.0, ov);
18 PersistentVariables
19   "mkl.h"
```

Figure 7. This LiLAC-What program implements spmv-csr naïvely with Intel MKL. Performance is degraded because of lines 2–5. Figure 9 will present a solution to this bottleneck.

```
1  INPUT CudaRead
2    cudaMemcpy(out, in, sizeof(type_in)*size,
3               cudaMemcpyHostToDevice);
4  BeforeFirstExecution
5    cudaMalloc(&out, sizeof(type_in)*size);
6  AfterLastExecution
7    cudaFree(out);
```

Figure 8. LiLAC-How code to provide efficient automatic data marshaling between the host and the CUDA accelerator.

```
1  INPUT Maximum
2    out = in[0];
3    for(int i = 1; i < size; i++)
4      out = in[i]>out?in[i]:out;
5    out = out+1;
```

Figure 9. *INPUT* can also be used to specify data-dependent computations that are only recalculated when necessary.

```
1  HARNESS cuda IMPLEMENTS spmv_csr
2    double alpha = 1.0;
3    double beta  = 0.0;
4    cusparseMatDescr_t descrA;
5    cusparseCreateMatDescr(&descrA);
6    cusparseDcsrmv(handle,
7                   CUSPARSE_OPERATION_NON_TRANSPOSE,
8                   rows, cols, ranges[rows], &alpha,
9                   descrA, d_mat, d_ranges, d_indir,
10                  d_vec, &beta, d_out);
11 Marshaling
12   int      cols = Maximum of indir[0..ranges[rows]]
13   double* d_mat = CudaRead of matrix[0..ranges[rows]]
14   double* d_vec = CudaRead of vector[0..cols]
15   int* d_ranges = CudaRead of ranges[0..rows+1]
16   Int*  d_indir = CudaRead of indir[0..rowstr[rows]]
17   double* d_out = CudaWrite of output[0..rows]
18 PersistentVariables
19   cusparseHandle_t handle
20 BeforeFirstExecution
21   cusparseCreate(&handle);
22 AfterLastExecution
23   cusparseDestroy(handle);
24 CppHeaderFiles
25   <cuda_runtime.h> "cusparse_v2.h"
```

Figure 10. This LiLAC-What specification implements an efficient SPMV harness using cuSPARSE in 25 lines of code.

3.3.2 Data Marshaling

Heterogeneous accelerators require data transfers to keep memory consistent between host device and accelerator. To achieve the best performance, these have to be minimized.

Importantly, unchanged data should never be copied again. This requires program-wide analysis that is not available statically. LiLAC-How uses memory protection to implement this at runtime with minimal overhead by capturing read and write accesses to memory ranges. The same mechanism is used to cache data-dependent invariants across several invocations, such as *cols* in Figure 7.

Data *marshaling* routines are bound to ranges of memory in the harness. In the specification, the underlying array is available using the identifiers `in`, `size`, and `out`.

3.3.3 Detailed Example

In Figure 8, the `cudaMemcpy` function from NVIDIA CUDA is integrated with LiLAC-How. It is used to copy data from the host to the accelerator. For this to work, it first needs to allocate memory of the device using `cudaMalloc`, which is later freed with `cudaFree`. Minimal memory transfers are obtained by executing `cudaMemcpy` only when a value in the array changes.

We can use the same construct to efficiently compute values such as the `cols` variable in Figure 7, as shown in Figure 9. The optimized implementation is derived from Figure 7 lines 2-5. However, instead of the concrete variable names, the reserved identifiers `in`, `size`, and `out` are used.

Figure 10 shows an `spmv_csr` LiLAC-How program for the cuSPARSE library. A number of data marshaling variables are introduced in lines 12–17, that automatically optimize both memory transfers and the computation of the `cols` variable. The core of the harness in lines 2–10 is again nothing more than library-specific boilerplate C++ code.

4 Implementation

The LiLAC system, as shown in Figure 2 is entirely integrated into the LLVM build system. When LLVM is compiled, the LiLAC specification is parsed using a Python program. Based on the LiLAC-What and LiLAC-How sections, C++ code is generated that is automatically incorporated into LLVM in further stages of the build process.

The result is an LLVM optimization pass that is available when linking LLVM with the clang C/C++ compiler. This pass performs the discovery of linear algebra code and the insertion of harness calls. Furthermore, the harness libraries themselves are built at compile time of LLVM, using C++ code emitted from the LiLAC-How sections.

The two crucial implementation details are therefore the following: Firstly, how automatic detection functionality in C++ is generated from the LiLAC-What specifications. Secondly, how the LiLAC-How sections are used to generate fast C++ implementations of the specified library harnesses.

4.1 LiLAC-What

The parsed LiLAC-What sections are turned into C++ functions that recognize places for harness call insertions in an LLVM pass. This builds on previous work via a formulation in CAnDL [21]. Detection is done on optimized compiler intermediate representation. Standard `-O2` optimizations, excluding loop unrolling and vectorization, normalize the intermediate code. Optimizations minimize programming language-specific artifacts and the impact of syntax-level programmer decisions.

The effect is demonstrated in Figure 11, which shows three implementations of a dot product in different languages: C, C++, and FORTRAN. After translating to LLVM IR and performing optimizations, the dot product is recognized in the LiLAC system using the same LiLAC-What specification.

The detection comprises two steps, as demonstrated in Figure 13. Firstly, the control flow skeleton is recognized. This is simple, as LiLAC-What can only express control flow in the form of loop nests of a certain depth. After candidate loop nests have been identified, the index and loop range calculations from LiLAC-What are mapped onto the LLVM IR nodes. This is done via a backtracking search procedure and allows robust detection across many syntactically different input programs, as described in [21, 22].

4.1.1 Backtracking Search Algorithm

For detecting instances of LiLAC-What specifications in user programs, LLVM IR segments that match the control flow skeleton are identified. These control flow candidates are then processed with a backtracking search algorithm.

All ⟨*exp*⟩ expressions in the LiLAC-What program are identified. These have to be assigned instructions or other values from the LLVM IR segment. Those top-level ⟨*exp*⟩ expressions that are used as limits or iterators in ⟨*range*⟩ expressions are easily connected with the corresponding loop boundaries in the control flow candidates.

The remaining expressions are successively assigned by backtracking. Consider the example in Figure 12, which shows a candidate loop from the LLVM IR generated from the C++ dot product code in Figure 11. The iteration space is determined by loop analysis and this immediately allows us to assign the iterator and range in Figure 13 on the left. The LLVM IR values that correspond to `a[i]`, `a`, `b[i]`, `b`, `a[i]*b[i]` and `result` are then searched for. When a partial solution fails, the algorithm backtracks. This happens in the example once, when no suitable multiplication can be found in step 5. If no complete solution can be determined, the control flow candidate is discarded.

4.1.2 Code Replacement

Each loop nest that matches a LiLAC-What specification is replaced with a harness call. To minimize the invasiveness of our pass, this is performed as follows: Firstly, a harness call is inserted directly before the loop. The function call arguments are selected from the backtracking result and passed to the harness. Secondly, the LLVM instruction that stores the result of the computation or passes it out of the loop as a phi node is removed. The remainder of the loop nest is removed automatically by dead code elimination.

```
COMPUTATION dotproduct
    result = sum(0 <= i < length) a[i] * b[i];
```

```
int i = 0;
while(i < N) {
    x += (*(A+i))*(*(B+i));
    i+=1; }
```

```
for(int i = 0; i < vec_a.size(); i++)
    x += vec_a[i]*vec_b[i];
```

```
DO I = 1, N, 1
    X = X + A(i)*B(i)
END DO
```

Figure 11. Syntactically different computations in C, C++, or FORTRAN are captured by one LiLAC-What specification.

```
; <label>:17:
  %18 = phi i64 [ 0, %10 ], [ %26, %17 ]
  %19 = phi double [ 0.0, %10 ], [ %25, %17 ]
  %20 = getelementptr double, double* %9, i64 %18
  %21 = load double, double* %20
  %22 = getelementptr double, double* %12
  %23 = load double, double* %22
  %24 = fmul double %21, %23
  %25 = fadd double %19, %24
  %26 = add nuw i64 %18, 1
  %27 = icmp ugt i64 %14, %26
  br i1 %27, label %17, label %15
```

Figure 12. LiLAC intercepts LLVM IR after optimizations. This ensures normalized and language-independent features.

			a[i]	←	**1:** %21		
			a	←	**2:** %9		
i	←	%18	b[i]	←	**3:** %21	**6:** %23	
length	←	%14	b	←	**4:** %9	**7:** %12	
			a[i] * b[i]	←	**5:** fail!	**8:** %24	
			result	←		**9:** %25	

Figure 13. After finding a candidate loop and receiving some variables from loop analysis (left), the backtracking solver attempts to assign the remaining variables one by one (right).

4.2 LiLAC-How

LiLAC-How syntax elements that take C++ code generate generic functions, and template parameter deduction inserts concrete types during the compilation process.

In Figure 14, we see the correspondence between generated C++ template functions and the specification in Figure 8. The three function bodies are directly inserted. The functions are used to specialize the *ReadObject* class template, which guarantees the following properties via memory protection: ***construct*** is called before the first invocation and when `in` or `size` change for consecutive harness invocations.

```
template<typename type_in, typename type_out>
void CudaRead_update(type_in* in, int size,
                     type_out& out) {
    cudaMemcpy(out, in, sizeof(type_in)*size,
               cudaMemcpyHostToDevice);
}
template<typename type_in, typename type_out>
void CudaRead_construct(int size, type_out& out) {
    cudaMalloc(&out, sizeof(type_in)*size);
}
template<typename type_in, typename type_out>
void CudaRead_destruct(int size, type_out& out) {
    cudaFree(out);
}
template<typename type_in, typename type_out>
using CudaRead = ReadObject<type_in, type_out,
    CudaRead_update<type_in,type_out>,
    CudaRead_construct<type_in,type_out>,
    CudaRead_destruct<type_in,type_out>>;
```

Figure 14. LiLAC uses code from Figure 8 to define three functions that specialize the ReadObject template, which uses `mprotect` for capturing memory accesses internally.

update is called after *construct* and if any of the data in the array is changed between consecutive harness invocations. ***destruct*** is called in between consecutive *construct* calls and before the program terminates.

4.3 FORTRAN

The LLVM frontend for FORTRAN under active development, flang, is in an unfinished state and produces unconventional LLVM IR code. Significant additional work was required to normalize the IR code. We developed normalization passes in LLVM to overcome the specific shortcomings, enabling FORTRAN programs to be managed as easily as C/C++.

The problems that we encountered included: differing indexing conventions requiring offsetting pointer variables on a byte granularity with untyped pointers; incompatible intermediate representation types where all parameters are passed in as `i64` pointers, frequently necessitating a pointer type conversion followed by a load from memory; obfuscated loops with additional induction variable that counts down instead of up such that the standard LLVM **indvars** pass is unable to merge the loop iterators.

5 Experimental Setup

We wrote short LiLAC programs for a collection of linear algebra libraries and applied our approach to a chemical simulation application, two graph analytics applications and a collection of standard benchmark suites.

Libraries We selected four different libraries for sparse linear algebra functions. These were: Intel MKL [3], Nvidia cuSPARSE [5], clSPARSE [2] and SparseX [19]. MKL is a general-purpose mathematical library, while clSPARSE and cuSPARSE are OpenCL and CUDA implementations of sparse linear algebra designed to be executed on the GPU, and SparseX uses an auto-tuning model and code generation to optimize sparse operations on particular matrices.

Table 1. Evaluated platforms and library harnesses; AMD-0 supports clSPARSE on both its internal and its external GPU.

Name	Hardware	Libraries
Intel-0	2× Intel Xeon E5-2620 Nvidia Tesla K20 GPU	MKL cuSPARSE
Intel-1	Intel Core i7-8700K Nvidia GTX 1080 GPU	clSPARSE SparseX
AMD	AMD A10-7850K AMD Radeon R7 iGPU Nvidia Titan X GPU	cuSPARSE clSPARSE ×2 SparseX

Applications To evaluate the impact of LiLAC in a real-world context, we used the PATHSAMPLE physical chemistry simulation suite, a large FORTRAN legacy application [56] consisting of over 40,000 lines of code. Recent work shows that applications in this area are amenable to acceleration using sparse linear algebra techniques [53], and PATHSAMPLE provides a useful example of this. We also evaluated two modern C++ graph analytics kernels (BFS and PageRank [11, 15]). PATHSAMPLE was run in two different modes and three different levels of pruning, in each case using a system of 38 atoms [18] commonly used to evaluate applications in this domain. The graph kernels were run against 10 matrices from the University of Florida's sparse matrix collection [14], with sizes between 300K and 80M non-zero elements.

For completeness and validation that our LiLAC-generated implementations were correct, we also applied our technique to sparse programs from standard benchmark suites: CG from the NAS parallel benchmarks [9], spmv from Parboil [48] and the Netlib sparse benchmark suites [17]. Each benchmark suite was run using their supplied inputs.

Platforms We evaluated our approach across 3 different machines with varying hardware performance and software availability. Each one was only compatible with a subset of our LiLAC-generated implementations—a summary of these machines is given in Table 1.

6 Results

We first present raw performance impact, then we analyze two intermediate metrics: reliability of linear algebra discovery and effectiveness of memory transfer optimizations.

6.1 Performance

LiLAC achieves significant speedups on real applications as well as benchmarks, as shown in Figure 15. Baselines were compiled with `-O2` using the same version of clang without LiLAC extensions. Higher optimization levels (`-O3`) had a negligible impact on performance. Different platforms and applications profit from different libraries (subsection 6.2). Speedup ranges from 1.1–3× on the scientific application codes to 12× on well-known sparse benchmark programs.

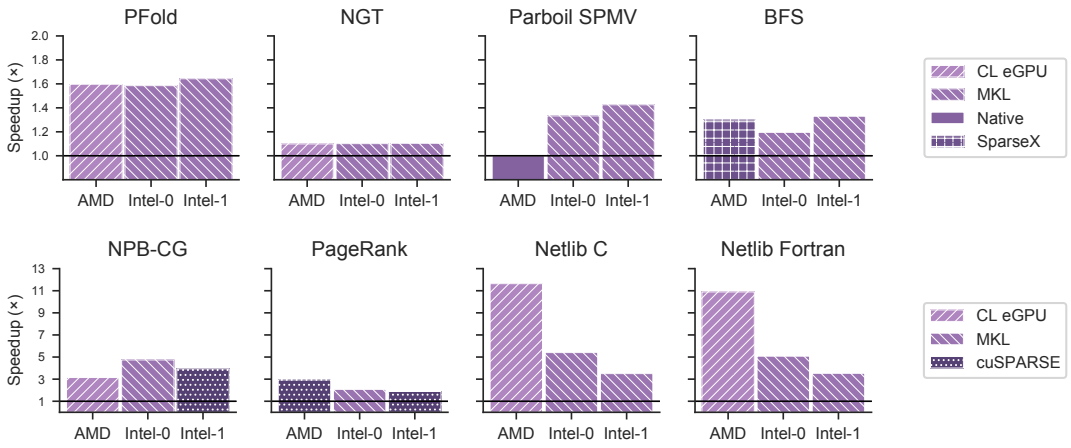


Figure 15. Evaluation on real-world applications and well-known benchmarks: Bars show the geomean speedup of the best-performing LiLAC harness across the set of input examples for each program and platform. Hatchings encode the selected implementations. The baseline is the identical source code compiled with clang `-O2`, yielding sequential CPU-only programs.

Applications On the PATHSAMPLE applications (PFold and NGT), we measured consistent speedups of approximately 50% and 10% respectively across all 3 platforms. For large applications, Amdahl's law is a severe limitation for approaches like ours – other parts of the applications dominate execution times when linear algebra is accelerated.

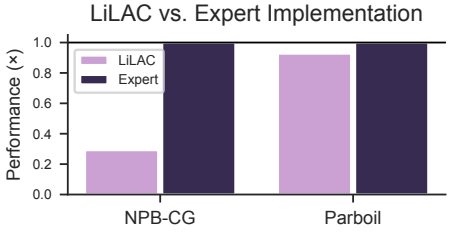
Graph kernels PageRank requires a large number of SPMV calls using the same input matrix to iterate until convergence. The GPU implementations running on AMD and Intel-1 take advantage of data remaining in memory. The larger number of CPU cores and slower GPU available on Intel-0 make MKL its best-performing implementation. CPU implementations perform best on BFS by avoiding memory copies entirely – on AMD, SparseX outperforms GPU implementations.

Benchmarks LiLAC achieves speedups of up to 12× on standard sparse linear algebra benchmarks. The impact is independent of the source language, as the C and FORTRAN versions of the Netlib benchmark demonstrate. LiLAC is able to achieve consistent, useful speedups across a variety of hardware configurations.

Dense We evaluated on some dense benchmarks as well. In line with the literature, dense is very amenable to heterogeneous acceleration. We achieve 20× speedup on Parboil sgemm by inserting LiLAC-harnessed calls into sequential baseline. However, impressive heterogeneous speedups on dense are well explored in the literature, we focus on sparse.

Comparison to Expert NPB and Parboil contain expert-written alternative versions with GPU acceleration. This allowed the evaluation of LiLAC against heterogeneous code reaching close to peak performance, shown in Figure 16.

While the expert version of NPB-CG is ∼ 3× faster, this is not due to an improved sparse linear algebra operation, but a complete parallelization and rewrite of the program for the GPU. In Parboil SPMV, the expert version focuses on improved sparse linear algebra. Here the difference between an expert and LiLAC is only 1.07×.



Benchmark	Modified LoC	
	LiLAC	Expert
NPB-CG	0 (44)	1948
Parboil SPMV	0 (44)	261

Figure 16. LiLAC performance as fraction of expert version performance. We achieve good speedup with no application programmer effort (measured as required LoC change). The LiLAC required code – identical across programs – is in parentheses. Amdahl's Law limits our impact on NPB-CG.

Productivity The bottom of Figure 16 shows the amount of code modified in order to add heterogeneous acceleration manually vs with LiLAC. This demonstrates the productivity improvements for application programmers. No lines of user code need to be modified using LiLAC, while both expert versions require significant application rewrites. Only 44 lines of application-independent LiLAC code is required.

6.2 Necessity of Flexible Backends

The relative performance of different accelerator libraries is highly dependent on the application, problem size, and platform, as Figure 17 shows.

Table 2. LiLAC speedups on each platform, across different applications and problem sizes. SparseX demonstrated promising performance on some applications, but we were unable to evaluate on every relevant instance due to instability. Implementation with best geomean speedup per benchmark and platform is bold.

Platform	Implementation	PFold			NGT			PageRank			BFS		
		L0	L1	L2	L0	L1	L2	Erdos	LJ-2008	Road	Erdos	LJ-2008	Road
AMD	cuSPARSE	1.38	1.18	0.67	0.69	0.69	0.70	**3.44**	**1.18**	**9.97**	1.62	6.55	1.96
	clSPARSE (eGPU)	**2.17**	**1.82**	**1.22**	**1.16**	**1.16**	**1.16**	3.08	1.24	6.06	0.50	11.03	0.24
	clSPARSE (iGPU)	2.03	1.78	1.03	0.90	0.90	0.90	3.26	1.31	4.05	0.14	4.17	0.05
	SparseX	-	-	-	-	-	-	-	-	-	**1.93**	-	-
Intel-0	MKL	**2.88**	**2.46**	**1.00**	**1.18**	**1.18**	**1.18**	**1.25**	**2.93**	**1.72**	**2.50**	**1.06**	**1.05**
	cuSPARSE	0.75	0.60	0.45	0.66	0.66	0.66	1.39	1.00	3.32	0.87	1.74	1.28
	clSPARSE	0.90	0.75	0.46	0.81	0.79	0.78	1.24	0.95	2.24	0.13	1.45	0.07
	SparseX	-	-	-	-	-	-	-	-	-	1.19	-	-
Intel-1	MKL	**2.70**	**2.43**	**1.01**	**1.20**	**1.20**	**1.19**	1.63	1.03	2.26	**1.06**	**2.09**	**1.27**
	cuSPARSE	0.48	0.41	0.30	0.68	0.69	0.68	**1.59**	**0.87**	**4.44**	1.01	1.83	1.63
	clSPARSE	1.00	1.00	1.00	1.00	1.02	1.00	1.50	0.87	3.46	0.23	1.81	0.13

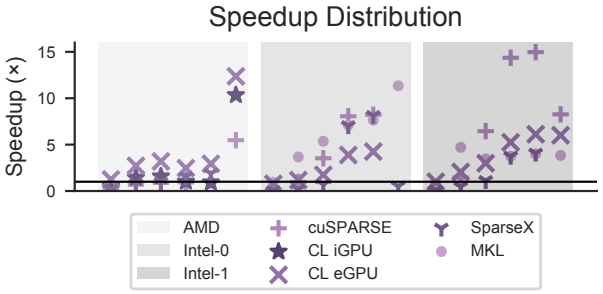


Figure 17. Distribution of speedups on NPB-CG. The stacks within each of the three columns are sorted by problem size, each point shows the speedup of a specific implementation.

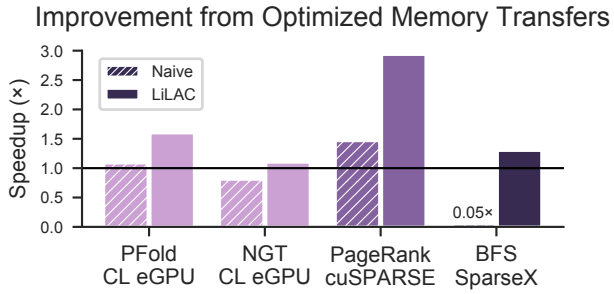


Figure 18. LiLAC vs. naïve library calls without marshaling optimizations, speedup over sequential baseline: Advanced marshaling features of LiLAC are critical for performance.

Table 2 has more detailed data. The best-performing implementation varies considerably, depending on characteristics of the problem in question. No accelerator library performs well reliably, each harness outperforms any other harness on some combination of data set and platform. For some small problem sizes, hardware acceleration is not profitable. Those slowdowns are due to inherent overheads, not LiLAC.

6.3 Effectiveness of Data Marshaling

Our implementation of LiLAC relies on a non-trivial data marshaling system that prevents redundant computations and memory transfers. We present performance results that show the importance and effectiveness of this system.

We repeated our experiments, using the best-performing implementations from Figure 15. Instead of using the data marshaling scheme, we recompute and transfer memory naively for each invocation. The results are in Figure 18. Across the best AMD versions of PFold, NGT, PageRank and BFS – where accelerators are profitable with marshaling – only PageRank achieves a significant speedup naively.

For BFS, the naive approach leads to drastic performance degradation, the marshaling version is 25× faster. This is because it performs an internal matrix tuning phase that is far more expensive than a memory copy. For the other three programs, there is a factor of 1.4–3.5× between the naive and the smart version.

6.4 Reliability of Discovery

For performance impact, LiLAC needs to first detect linear algebra computations. Previous results already implied that this works reliably, and Table 3 reiterates this. All relevant sparse matrix-vector multiplications were recognized.

Established approaches, like the polyhedral model, are unable to model sparse linear algebra, as verified with the Polly compiler. Similarly, the Intel C/C++ and FORTRAN compilers fail to auto-parallelize, as they cannot reason about sparsity and have to assume additional dependencies.

These results show the novelty of the abilities of LiLAC rather than implementation weaknesses of Polly and ICC, as neither were designed for accelerating sparse computations.

Table 3. Sparsity does not fit the polyhedral model; Polly is not available for FORTRAN; Intel compilers fail to parallelize sparse. Only LiLAC detects sparse linear algebra reliably.

Benchmark	LiLAC	Polly	Intel icc/ifort
PFold	CSR	-	parallel dependence
NGT	CSR	-	parallel dependence
Parboil-SPMV	JDS	no SCoP	parallel dependence
BFS	CSR	no SCoP	parallel dependence
NPB-CG	CSR	-	parallel dependence
PageRank	CSR	no SCoP	parallel dependence
Netlib C	CSR	no SCoP	parallel dependence
Netlib Fortran	CSR	-	parallel dependence

7 Related Work

Compiler centric linear algebra optimization Compiler management of indirect memory accesses was first examined using an inspector-executor model for distributed-memory machines [10]. The location of read data was discovered at runtime and appropriate communication inserted. Later work was focused on efficient runtime dependence analysis and the parallelization of more general programs [20, 38, 41, 49]. However, the performance achieved is modest due to runtime overhead and falls well short of library performance. More recent work developed equality constraints and subset relations that help reduce the runtime overhead [32].

The polyhedral model is an established compiler approach for modeling data dependencies [12, 24, 26, 42, 47]. Such an approach has been implemented in optimizing compilers, such as the Polly extensions to LLVM [16]. Recent work has extended the polyhedral model beyond affine programs to some forms of sparsity with the PENCIL extensions [8]. These can be used to model important features of sparse linear algebra, such as counted loops [58], i.e. loops with dynamic, memory dependent bounds but statically known strides. Such loops are central to sparse linear algebra. The PPCG compiler [54] can detect relevant code regions, but it relies on well behaved C code with all arrays declared in variable-length C99 array syntax. This excludes most real-world programs; nothing in our evaluation fits this structure.

The Appollo system [51] integrates thread level speculation with the polyhedral model, allowing its application to sparse linear algebra. However, it requires sub-parts of the computation to perform dense accesses at runtime. Similar approaches [7] also require regular sub-computations.

Compiler detection Previous work has detected code structures in compilers using constraint programming. Early work was based on abstract computation graphs [37], but more recent approaches have used compiler intermediate code and made connections to the polyhedral model [21].

In [22] they implement a method that operates on SSA intermediate representation. It uses a general-purpose low-level constraint programming language aimed at compiler engineers. The paper focuses on code detection, with manual data marshaling. Recent work [13] uses type-guided program synthesis to model library routines, which are then detected by a solver. Again, data marshaling is not taken into account.

Other advanced approaches to extracting higher-level structures from assembly and well-structured FORTRAN code involve temporal logic [27, 31]. These approaches tend to focus on a more restricted set of computations (dense memory access). While this allows formal reasoning about correctness, is too restrictive to model sparse linear algebra.

Domain-Specific Languages There have been multiple domain-specific libraries proposed to formulate linear algebra computations. Many of these contain some degree of autotuning functionality to achieve good performance across different platforms [50]. Halide [39] was designed for image processing. [52]. Its core design decision is the scheduling model that allows the separation of the computation schedule and the actual computation. There has been work on automatically tuning the schedules [35] but in general, the computational burden is put on the application programmer.

The SLinGen [45] compiler takes a program expressed in the custom LA language, inspired by standard mathematical notation. It then implements custom code generation for the expressed calculations, with a focus on small, fixed-size operands. This is built on top of building blocks provided by previous work on LGen [46]. The approach outperforms libraries focused on large data sizes but is unable to utilize heterogeneous compute and requires program rewrites.

Libraries The most established way of encapsulating fast linear algebra routines is via numeric libraries, generally based on the BLAS interface [6]. These are generally very fast on specific hardware platforms, but require application programmer effort and offer little performance portability. Implementations of dense linear algebra are available for most suitable hardware platforms, such as cuBLAS [4] for NVIDIA GPUs, clBLAS [1] for AMD GPUs and the Intel MKL library [3] for Intel CPUs and accelerators.

Fast implementations of sparse linear algebra are fewer, but they exist for the most important platforms, including cuSPARSE [5] and clSPARSE [2]. There have been several BLAS implementations that attempt platform independent acceleration and heterogeneous compute [33, 34, 57].

CPU-GPU data transfer optimizations Data transfers between CPU and GPU have been studied extensively as an important bottleneck for parallelization efforts. Previous work [25, 30] established systems for automatic management of CPU-GPU communication. The authors of [29] implemented a system to move OpenMP code to GPUs, optimizing data transfers using data flow analysis. However, this approach performs a direct translation, not optimizing the code for the specific performance characteristics of GPUs.

8 Conclusion

This paper presented LiLAC, a language and compiler that enables existing codebases to exploit sparse (and dense) linear algebra accelerators. No effort is required from the application programmer. Instead, the library implementer provides a specification, which LiLAC uses to automatically and efficiently match user code to high-performance libraries.

We demonstrated this approach on C, C++, and FORTRAN benchmarks as well as legacy applications, and shown significant performance improvement across platforms and data sets. In future work, we will investigate how our framework can be adapted to other application domains, enabling effort-free access to an even larger set of accelerator libraries.

References

[1] [n.d.]. clMathLibraries clBLAS. https://github.com/clMathLibraries/clBLAS.

[2] [n.d.]. clMathLibraries clSPARSE. https://github.com/clMathLibraries/clSPARSE.

[3] [n.d.]. Intel® Math Kernel Library (MKL). https://software.intel.com/mkl.

[4] [n.d.]. NVIDIA cuBLAS. https://developer.nvidia.com/cublas.

[5] [n.d.]. NVIDIA CUDA Sparse Matrix library (cuSPARSE). https://developer.nvidia.com/cusparse.

[6] 2002. An Updated Set of Basic Linear Algebra Subprograms (BLAS). *ACM Trans. Math. Softw.* 28, 2 (June 2002), 135–151. https://doi.org/10.1145/567806.567807

[7] Travis Augustine, Janarthanan Sarma, Louis-Noël Pouchet, and Gabriel Rodríguez. 2019. Generating Piecewise-Regular Code from Irregular Structures. In *Proceedings of the 40th ACM SIGPLAN Conference on Programming Language Design and Implementation (PLDI 2019)*. Association for Computing Machinery, New York, NY, USA, 625–639. https://doi.org/10.1145/3314221.3314615

[8] R. Baghdadi, U. Beaugnon, A. Cohen, T. Grosser, M. Kruse, C. Reddy, S. Verdoolaege, A. Betts, A. F. Donaldson, J. Ketema, J. Absar, S. v. Haastregt, A. Kravets, A. Lokhmotov, R. David, and E. Hajiyev. 2015. PENCIL: A Platform-Neutral Compute Intermediate Language for Accelerator Programming. In *2015 International Conference on Parallel Architecture and Compilation (PACT)*. 138–149. https://doi.org/10.1109/PACT.2015.17

[9] D. H. Bailey, E. Barszcz, J. T. Barton, D. S. Browning, R. L. Carter, L. Dagum, R. A. Fatoohi, P. O. Frederickson, T. A. Lasinski, R. S. Schreiber, H. D. Simon, V. Venkatakrishnan, and S. K. Weeratunga. 1991. The NAS Parallel Benchmarks&Mdash;Summary and Preliminary Results. In *Proceedings of the 1991 ACM/IEEE Conference on Supercomputing (Supercomputing '91)*. ACM, New York, NY, USA, 158–165. https://doi.org/10.1145/125826.125925

[10] D. Baxter, R. Mirchandaney, and J. H. Saltz. 1989. Run-time Parallelization and Scheduling of Loops. In *Proceedings of the First Annual ACM Symposium on Parallel Algorithms and Architectures (SPAA '89)*. ACM, New York, NY, USA, 303–312. https://doi.org/10.1145/72935.72967

[11] Scott Beamer, Krste Asanović, and David Patterson. 2015. The GAP Benchmark Suite. (08 2015).

[12] Lam Chi-Chung, P Sadayappan, and Rephael Wenger. 1997. On optimizing a class of multi-dimensional loops with reduction for parallel execution. *Parallel Processing Letters* 7, 02 (1997), 157–168.

[13] B. Collie, P. Ginsbach, and M. F. P. O'Boyle. 2019. Type-Directed Program Synthesis and Constraint Generation for Library Portability. In *2019 28th International Conference on Parallel Architectures and Compilation Techniques (PACT)*. 55–67. https://doi.org/10.1109/PACT.2019.00013

[14] Timothy A. Davis and Yifan Hu. 2011. The University of Florida Sparse Matrix Collection. *ACM Trans. Math. Softw.* 38, 1, Article 1 (Dec. 2011), 25 pages. https://doi.org/10.1145/2049662.2049663

[15] Camil Demetrescu, Andrew V Goldberg, and David S Johnson. [n.d.]. *The Shortest Path Problem: Ninth DIMACS Implementation Challenge*. Vol. 74. American Mathematical Soc.

[16] Johannes Doerfert, Kevin Streit, Sebastian Hack, and Zino Benaissa. 2015. Polly's Polyhedral Scheduling in the Presence of Reductions. *CoRR* abs/1505.07716 (2015). http://arxiv.org/abs/1505.07716

[17] Jack Dongarra, Victor Eijkhout, and Henk van der Vorst. 2001. An Iterative Solver Benchmark. *Sci. Program.* 9, 4 (Dec. 2001), 223–231. https://doi.org/10.1155/2001/527931

[18] Jonathan P. K. Doye, Mark A. Miller, and David J. Wales. 1999. The double-funnel energy landscape of the 38-atom Lennard-Jones cluster. *The Journal of Chemical Physics* 110, 14 (1999), 6896–6906. https://doi.org/10.1063/1.478595 arXiv:https://doi.org/10.1063/1.478595

[19] Athena Elafrou, Vasileios Karakasis, Theodoros Gkountouvas, Kornilios Kourtis, Georgios Goumas, and Nectarios Koziris. 2018. SparseX: A Library for High-Performance Sparse Matrix-Vector Multiplication on Multicore Platforms. *ACM Trans. Math. Softw.* 44, 3 (Jan. 2018), 26:1–26:32. https://doi.org/10.1145/3134442

[20] Allan L Fisher and Anwar M Ghuloum. 1994. Parallelizing complex scans and reductions. In *ACM SIGPLAN Notices*, Vol. 29. ACM, 135–146.

[21] Philip Ginsbach, Lewis Crawford, and Michael F. P. O'Boyle. 2018. CAnDL: A Domain Specific Language for Compiler Analysis. In *Proceedings of the 27th International Conference on Compiler Construction (CC 2018)*. ACM, New York, NY, USA, 151–162. https://doi.org/10.1145/3178372.3179515

[22] Philip Ginsbach, Toomas Remmelg, Michel Steuwer, Bruno Bodin, Christophe Dubach, and Michael F. P. O'Boyle. 2018. Automatic Matching of Legacy Code to Heterogeneous APIs: An Idiomatic Approach. In *Proceedings of the Twenty-Third International Conference on Architectural Support for Programming Languages and Operating Systems (ASPLOS '18)*. ACM, New York, NY, USA, 139–153. https://doi.org/10.1145/3173162.3173182

[23] Tobias Grosser, Armin Größlinger, and Christian Lengauer. 2012. Polly - Performing Polyhedral Optimizations on a Low-Level Intermediate Representation. *Parallel Processing Letters* 22, 4 (2012). https://doi.org/10.1142/S0129626412500107

[24] Gautam Gupta and Sanjay V Rajopadhye. 2006. Simplifying reductions.. In *POPL*, Vol. 6. 30–41.

[25] Thomas B. Jablin, Prakash Prabhu, James A. Jablin, Nick P. Johnson, Stephen R. Beard, and David I. August. 2011. Automatic CPU-GPU Communication Management and Optimization. *SIGPLAN Not.* 46, 6 (June 2011), 142–151. https://doi.org/10.1145/1993316.1993516

[26] Pierre Jouvelot and Babak Dehbonei. 1989. A unified semantic approach for the vectorization and parallelization of generalized reductions. In *Proceedings of the 3rd international conference on Supercomputing*. ACM, 186–194.

[27] Shoaib Kamil, Alvin Cheung, Shachar Itzhaky, and Armando Solar-Lezama. 2016. Verified Lifting of Stencil Computations. In *Proceedings of the 37th ACM SIGPLAN Conference on Programming Language Design and Implementation (PLDI '16)*. ACM, New York, NY, USA, 711–726. https://doi.org/10.1145/2908080.2908117

[28] Jeremy Kepner, David A. Bader, Aydin Buluç, John R. Gilbert, Timothy G. Mattson, and Henning Meyerhenke. 2015. Graphs, Matrices, and the GraphBLAS: Seven Good Reasons. In *ICCS*.

[29] Seyong Lee, Seung-Jai Min, and Rudolf Eigenmann. 2009. OpenMP to GPGPU: A Compiler Framework for Automatic Translation and Optimization. *SIGPLAN Not.* 44, 4 (Feb. 2009), 101–110. https://doi.org/10.1145/1594835.1504194

[30] Christos Margiolas and Michael F. P. O'Boyle. 2014. Portable and Transparent Host-Device Communication Optimization for GPGPU Environments. In *Proceedings of Annual IEEE/ACM International Symposium on Code Generation and Optimization (CGO '14)*. Association for Computing Machinery, New York, NY, USA, 55–65. https://doi.org/10.1145/2544137.2544156

[31] Charith Mendis, Jeffrey Bosboom, Kevin Wu, Shoaib Kamil, Jonathan Ragan-Kelley, Sylvain Paris, Qin Zhao, and Saman Amarasinghe. 2015. Helium: Lifting High-performance Stencil Kernels from Stripped x86 Binaries to Halide DSL Code. In *Proceedings of the 36th ACM SIGPLAN Conference on Programming Language Design and Implementation (PLDI '15)*. ACM, New York, NY, USA, 391–402. https://doi.org/10.1145/2737924.2737974

[32] Mahdi Soltan Mohammadi, Tomofumi Yuki, Kazem Cheshmi, Eddie C. Davis, Mary Hall, Maryam Mehri Dehnavi, Payal Nandy, Catherine Olschanowsky, Anand Venkat, and Michelle Mills Strout. 2019. Sparse Computation Data Dependence Simplification for Efficient Compiler-generated Inspectors. In *Proceedings of the 40th ACM SIGPLAN Conference on Programming Language Design and Implementation (PLDI 2019)*. ACM, New York, NY, USA, 594–609. https://doi.org/10.1145/3314221.3314646

[33] Ana Moreton-Fernandez, Arturo Gonzalez-Escribano, and Diego Ferraris. 2017. Multi-device Controllers: A Library to Simplify Parallel Heterogeneous Programming. (12 2017).

[34] Ana Moreton-Fernandez, Eduardo Rodriguez-Gutiez, Arturo Gonzalez-Escribano, and Diego R. Llanos. 2017. Supporting the Xeon Phi Coprocessor in a Heterogeneous Programming Model. In *Euro-Par 2017: Parallel Processing*, Francisco F. Rivera, Tomás F. Pena, and José C. Cabaleiro (Eds.). Springer International Publishing, Cham, 457–469.

[35] Ravi Teja Mullapudi, Andrew Adams, Dillon Sharlet, Jonathan Ragan-Kelley, and Kayvon Fatahalian. 2016. Automatically Scheduling Halide Image Processing Pipelines. *ACM Trans. Graph.* 35, 4, Article 83 (July 2016), 11 pages. https://doi.org/10.1145/2897824.2925952

[36] Dorit Nuzman, Sergei Dyshel, Erven Rohou, Ira Rosen, Kevin Williams, David Yuste, Albert Cohen, and Ayal Zaks. 2011. Vapor SIMD: Auto-vectorize Once, Run Everywhere. In *Proceedings of the 9th Annual IEEE/ACM International Symposium on Code Generation and Optimization (CGO '11)*. IEEE Computer Society, Washington, DC, USA, 151–160. http://dl.acm.org/citation.cfm?id=2190025.2190062

[37] Shlomit S Pinter and Ron Y Pinter. 1994. Program optimization and parallelization using idioms. *ACM Transactions on Programming Languages and Systems (TOPLAS)* 16, 3 (1994), 305–327.

[38] Bill Pottenger and Rudolf Eigenmann. 1995. Idiom recognition in the Polaris parallelizing compiler. In *Proceedings of the 9th international conference on Supercomputing*. ACM, 444–448.

[39] Jonathan Ragan-Kelley, Connelly Barnes, Andrew Adams, Sylvain Paris, Frédo Durand, and Saman Amarasinghe. 2013. Halide: A Language and Compiler for Optimizing Parallelism, Locality, and Recomputation in Image Processing Pipelines. *SIGPLAN Not.* 48, 6 (June 2013), 519–530. https://doi.org/10.1145/2499370.2462176

[40] Gil Rapaport, Ayal Zaks, and Yosi Ben-Asher. 2015. Streamlining Whole Function Vectorization in C Using Higher Order Vector Semantics. In *Proceedings of the 2015 IEEE International Parallel and Distributed Processing Symposium Workshop (IPDPSW '15)*. IEEE Computer Society, Washington, DC, USA, 718–727. https://doi.org/10.1109/IPDPSW.2015.37

[41] Lawrence Rauchwerger and David A Padua. 1999. The LRPD test: Speculative run-time parallelization of loops with privatization and reduction parallelization. *IEEE Transactions on Parallel and Distributed Systems* 10, 2 (1999), 160–180.

[42] Xavier Redon and Paul Feautrier. 1994. Scheduling reductions. In *Proceedings of the 8th international conference on Supercomputing*. ACM, 117–125.

[43] Y. Saad. 1989. Krylov Subspace Methods on Supercomputers. *SIAM J. Sci. Statist. Comput.* 10, 6 (1989), 1200–1232. https://doi.org/10.1137/0910073 arXiv:https://doi.org/10.1137/0910073

[44] Y. Saad. 2003. *Iterative Methods for Sparse Linear Systems* (second ed.). Society for Industrial and Applied Mathematics. https://doi.org/10.1137/1.9780898718003 arXiv:https://epubs.siam.org/doi/pdf/10.1137/1.9780898718003

[45] Daniele G. Spampinato, Diego Fabregat-Traver, Paolo Bientinesi, and Markus Püschel. 2018. Program Generation for Small-Scale Linear Algebra Applications. In *International Symposium on Code Generation and Optimization (CGO)*. 327–339.

[46] Daniele G. Spampinato and Markus Püschel. 2014. A Basic Linear Algebra Compiler. In *International Symposium on Code Generation and Optimization (CGO)*. 23–32.

[47] Kevin Stock, Martin Kong, Tobias Grosser, Louis-Noël Pouchet, Fabrice Rastello, Jagannathan Ramanujam, and Ponnuswamy Sadayappan. 2014. A framework for enhancing data reuse via associative reordering. In *ACM SIGPLAN Notices*, Vol. 49. ACM, 65–76.

[48] John Stratton, Christopher Rodrigues, I-Jui Sung, Nady Obeid, Li-Wen Chang, Nasser Anssari, Daniel Geng, Wen-Mei Liu, and Wen-mei Hwu. 2018. Parboil: A Revised Benchmark Suite for Scientific and Commercial Throughput Computing. (Nov. 2018).

[49] Toshio Suganuma, Hideaki Komatsu, and Toshio Nakatani. 1996. Detection and global optimization of reduction operations for distributed parallel machines. In *Proceedings of the 10th international conference on Supercomputing*. ACM, 18–25.

[50] Arvind K. Sujeeth, Kevin J. Brown, Hyoukjoong Lee, Tiark Rompf, Hassan Chafi, Martin Odersky, and Kunle Olukotun. 2014. Delite: A Compiler Architecture for Performance-Oriented Embedded Domain-Specific Languages. *ACM Trans. Embed. Comput. Syst.* 13, 4s, Article 134 (April 2014), 25 pages. https://doi.org/10.1145/2584665

[51] Aravind Sukumaran-Rajam and Philippe Clauss. 2015. The Polyhedral Model of Nonlinear Loops. *ACM Trans. Archit. Code Optim.* 12, 4, Article Article 48 (Dec. 2015), 27 pages. https://doi.org/10.1145/2838734

[52] Patricia Suriana, Andrew Adams, and Shoaib Kamil. 2017. Parallel Associative Reductions in Halide. In *Proceedings of the 2017 International Symposium on Code Generation and Optimization (CGO '17)*. IEEE Press, Piscataway, NJ, USA, 281–291. http://dl.acm.org/citation.cfm?id=3049832.3049863

[53] Kyle H. Sutherland-Cash, Rosemary G. Mantell, and David J. Wales. 2017. Exploiting sparsity in free energy basin-hopping. *Chemical Physics Letters* 685 (2017), 288 – 293. https://doi.org/10.1016/j.cplett.2017.07.081

[54] Sven Verdoolaege, Juan Carlos Juega, Albert Cohen, José Ignacio Gómez, Christian Tenllado, and Francky Catthoor. 2013. Polyhedral Parallel Code Generation for CUDA. *ACM Trans. Archit. Code Optim.* 9, 4, Article 54 (Jan. 2013), 23 pages. https://doi.org/10.1145/2400682.2400713

[55] Richard Vuduc, James W Demmel, and Katherine A Yelick. 2005. OSKI: A library of automatically tuned sparse matrix kernels. *Journal of Physics: Conference Series* 16 (jan 2005), 521–530. https://doi.org/10.1088/1742-6596/16/1/071

[56] David J. Wales. 2002. Discrete path sampling. *Molecular Physics* 100, 20 (2002), 3285–3305. https://doi.org/10.1080/00268970210162691 arXiv:https://doi.org/10.1080/00268970210162691

[57] Linnan Wang, Wei Wu, Zenglin Xu, Jianxiong Xiao, and Yi Yang. 2016. BLASX: A High Performance Level-3 BLAS Library for Heterogeneous Multi-GPU Computing. In *Proceedings of the 2016 International Conference on Supercomputing (ICS '16)*. ACM, New York, NY, USA, Article 20, 11 pages. https://doi.org/10.1145/2925426.2926256

[58] Jie Zhao, Michael Kruse, and Albert Cohen. 2018. A Polyhedral Compilation Framework for Loops with Dynamic Data-dependent Bounds. In *Proceedings of the 27th International Conference on Compiler Construction (CC 2018)*. ACM, New York, NY, USA, 14–24. https://doi.org/10.1145/3178372.3179509