BLIS: A Framework for Generating BLAS-like Libraries

FLAME Working Note #66

(We recommend reading the updated version of this paper titled "BLIS: A Framework for Rapidly Instantiating BLAS Functionality," which was submitted to ACM TOMS.)

Field G. Van Zee
Robert A. van de Geijn
Department of Computer Science
The University of Texas at Austin
Austin, TX 78712

November 15, 2012

Abstract

We propose the portable BLAS-like Interface Software (BLIS) framework which addresses a number of shortcomings in both the original BLAS interface and present-day BLAS implementations. The framework allows developers to rapidly instantiate high-performance BLAS-like libraries on existing and new architectures with relatively little effort. The key to this achievement is the observation that virtually all computation within level-2 and level-3 BLAS operations may be expressed in terms of very simple kernels. Higher-level framework code is generalized so that it can be reused and/or re-parameterized for different operations (as well as different architectures) with little to no modification. Inserting high-performance kernels into the framework facilitates the immediate optimization of any and all BLAS-like operations which are cast in terms of these kernels, and thus the framework acts as a productivity multiplier. Users of BLAS-dependent applications are supported through a straightforward compatibility layer, though calling sequences must be updated for those who wish to access new functionality. Experimental performance of level-2 and level-3 operations is observed to be competitive with two mature open source libraries (OpenBLAS and ATLAS) as well as an established commercial product (Intel MKL).

1 Introduction

The introduction of the Basic Linear Algebra Subprograms (BLAS) in the 1970s started a tradition of portable high performance for numerical software [22, 11, 10]. The level-1 BLAS were proposed in the 1970s and targeted the vector-vector operations that supported high performance on the vector supercomputers of that era. When cache-based architectures arrived in the 1980s, performance required better amortization of data movement between memory, caches, and registers. The BLAS interface was extended to include matrix-vector (level-2) and matrix-matrix (level-3) operations. These last operations can achieve high performance by amortizing $\mathcal{O}(n^2)$ memory operations over $\mathcal{O}(n^3)$ floating-point operations. By casting computation in terms of these routines, high-performance could be attained on a wide range of architectures on which high-performance implementations were made available.

While the BLAS were a huge step forward for the scientific computing community, the standard is, in our opinion, not ideal for today's needs. The original level-1 BLAS interface targeted Fortran-77 and was first

1

published in 1979, at a time when computer architectures were quite different. Since then, the community has gained a wealth of experience and encountered many new applications. Thus, it would not be unreasonable to ask whether an update is in order.

We are not the first to propose a major overhaul of the BLAS. In the 1990s, the BLAST forum convened multiple times in an attempt to develop various extensions to the BLAS, including an interface for the C programming language [8]. Unfortunately, the fruits of this forum never achieved broad acceptance.

Our effort is another such attempt. However, both our aim and approach are unique. Our goal is not to merely extend the BLAS, as with the BLAST Forum. Nor do we wish to simply create yet another high-performance BLAS implementation in the image of GotoBLAS [12, 13], OpenBLAS [25], or ATLAS [38]. Rather, we propose a software *framework* which will allow experts and vendors to rapidly instantiate high-performance BLAS-like libraries that contain functionality (and APIs) that are sufficiently generalized to avoid the shortcomings of the original BLAS altogether. In other words, the **B**LAS-**L**ike **I**nterface **S**oftware (BLIS) framework we propose constitutes a major redesign of the BLAS. It should be emphasized, though, that BLIS interfaces remain similar enough to BLAS that experienced users will likely find them intuitive. Furthermore, the BLIS framework will include an optional BLAS compatibility layer which will allow existing applications to link to BLIS libraries with no modification.

This paper makes the following contributions:

- It discusses and summarizes our evaluation of the shortcomings of the original BLAS and explains why each shortcoming is worth addressing.

- It proposes a new dense linear algebra framework which is designed to avoid the aforementioned shortcomings of the BLAS, while at the same time allowing a developer to (1) rapidly instantiate an entire BLAS-like library on a new architecture with relatively little effort, and (2) implement new BLAS-like operations that, whenever possible, leverage existing components of the framework.

- It observes that level-2 and level-3 operations can be implemented in such a way that virtually *all* key differences can be factored out to generic, architecture-agnostic code, which in turn helps us isolate essential kernels that are easy to understand, implement, and reuse. This observation is key because it allows us to drastically reduce the scope of code that must be manually optimized for a given architecture, allowing the BLIS framework to act as a productivity multiplier.

- It demonstrates level-2 and level-3 performance that is highly competitive with (and, in some cases, superior to) existing BLAS implementations such as OpenBLAS, ATLAS, and Intel's Math Kernel Library (MKL) [19].

- It makes the BLIS framework (including an optional BLAS compatibility layer) available to the community under an open source software license.[1]

These contributions, captured within the BLIS framework, lay the foundation for building modern dense linear algebra software that is flexible, portable, backwards compatible, and capable of high performance.

While we believe this paper makes significant contributions to the field, we readily admit that the proposed framework, as presented, does not address certain application needs such as multithreaded parallelism and environments that rely upon Graphics Processing Units (GPUs). We intend to investigate and add these features as part of future work.

2   Motivation

Motivation for BLIS was found primarily while developing the `libflame` library [33]. Since its earliest versions, `libflame` has relied upon the BLAS to provide high performance building blocks for both sequential and multithreaded algorithms [34, 29]. While developing and maintaining `libflame`'s external wrapper interfaces (i.e., object-based functions that directly call BLAS routines), we observed several shortcomings of the BLAS interface and its underlying functionality. We will now discuss each of these shortcomings.

---

[1] We make the BLIS framework available under the so-called "new" or "modified" BSD license.

Level	Name	Operations		
1	AXPY	$y := y + \alpha\overline{x}$		
	COPY	$y := \overline{x}$		
2	GEMV	$y := \beta y + \alpha\overline{A}x$	$y := \beta y + \alpha A\overline{x}$	$y := \beta y + \alpha\overline{Ax}$
	GER	$A := A + \alpha\overline{x}y^T$	$A := A + \alpha\overline{xy^T}$	
	HER	$A := A + \alpha\overline{xx^H}$		
	TRMV	$x := \overline{A}x$		
	TRSV	$x := \overline{A^{-1}}x$		
3	GEMM	$C := \beta C + \alpha\overline{A}B$	$C := \beta C + \alpha A\overline{B}$	$C := \beta C + \alpha\overline{AB}$
	HERK	$C := \beta C + \alpha\overline{AA^H}$	$C := \beta C + \alpha\overline{A^H A}$	
	TRMM	$B := \alpha\overline{A}B$		
	TRSM	$B := \alpha\overline{A^{-1}}B$		

Figure 1: A representative sample of variations of complex domain BLAS operations which are not implementable via the BLAS interface.

- **Inflexible storage.** BLAS requires that matrices be stored in column-major order [11, 10]. A C interface to the BLAS, known as CBLAS, supports both row- and column-major storage [8, 7]. However, when row-major storage is used, CBLAS requires that all matrix operands be stored in row-major order; thus, mixing row- and column-stored matrices within the same operation invocation is not allowed. To be fair, *some* combinations of row- and column-stored operands can be induced with BLAS that only support column (or row) storage by adjusting parameters, particularly the `trans`, `uplo`, and `side` parameters. However, this technique only results in partial support. For some storage combinations, the operation is possible only by making temporary copies of certain matrices with explicit transpositions (and/or conjugations). Furthermore, no implementation of the BLAS allows general striding, whereby neither rows nor columns are contiguous in memory. Support for this advanced storage scheme is needed for dense matrix computation on tensors [31].

- **Incomplete parameter support for the complex domain.** BLAS supports many operations on single- and double-precision complex data. However, the BLAS omits the "conjugate without transposition" option from all instances of the `trans` argument. Thus, many operations in the complex domain are not supported. Figure 1 contains a partial list of complex operations that are not supported by BLAS. One may argue that this is not a critical omission. After all, the programmer can simply conjugate an operand in-place, execute the desired BLAS operation, and then undo the conjugation. The problem with this is technique—besides the obvious conjugation overhead incurred—is that it may not be thread-safe. In a multithreaded environment, another thread could attempt to read an operand in a temporarily conjugated state, which would result in a race condition. Alternatively, one could create a temporary conjugated copy of the operand. However, this still comes with potentially significant allocation, workspace, and memory copy costs.

- **Opaque API.** BLAS exports a "single-layer" API. That is, the application developer can access routines such as `dgemm`, but he or she cannot access the lower-level building blocks (the so-called "kernels") that facilitate the routine's high-performance implementation. Access to these building blocks is crucial for experts who wish to efficiently construct specialized routines that cannot be found in the BLAS. Without access to these lower-level APIs, an application developer who needs specialized functionality is forced to compose a solution strictly in terms of the exposed BLAS interfaces. This often results in sub-optimal implementations because intermediate data products cannot be re-used across successive BLAS calls and thus must be redundantly packed and/or computed. Figure 2 illustrates
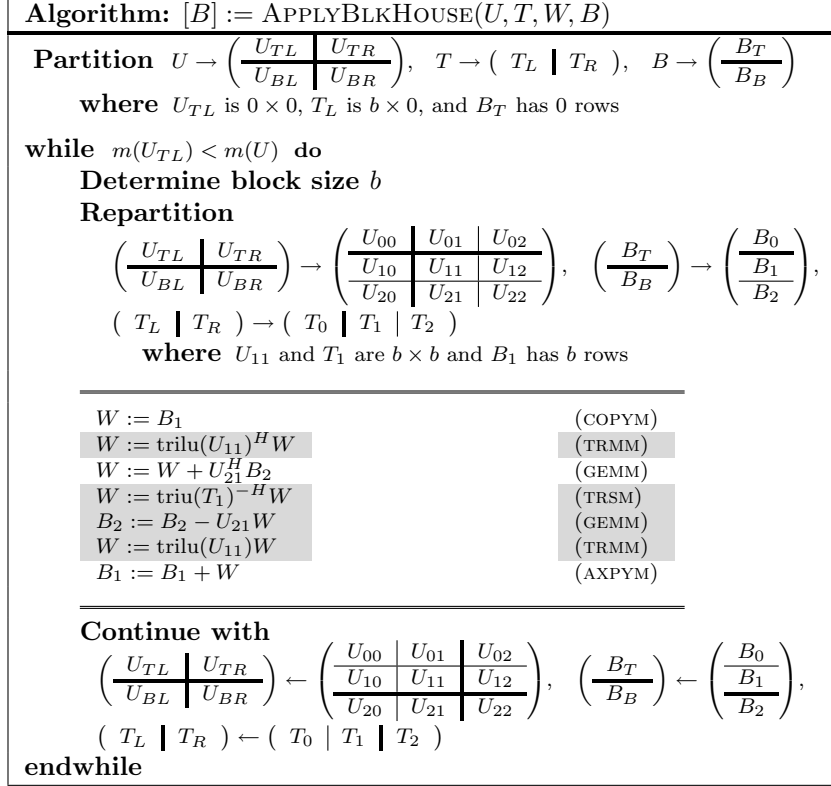
> **Algorithm:** $[B] := \textsc{ApplyBlkHouse}(U, T, W, B)$
>
> **Partition** $U \rightarrow \left(\begin{array}{c|c} U_{TL} & U_{TR} \\ \hline U_{BL} & U_{BR} \end{array} \right)$, $\quad T \rightarrow \left(\begin{array}{c|c} T_L & T_R \end{array} \right)$, $\quad B \rightarrow \left(\begin{array}{c} B_T \\ \hline B_B \end{array} \right)$
>
> $\quad$ **where** U_{TL} is 0×0, T_L is $b \times 0$, and B_T has 0 rows
>
> **while** $m(U_{TL}) < m(U)$ **do**
> $\quad$ **Determine block size** b
> $\quad$ **Repartition**
> $\quad\quad \left(\begin{array}{c|c} U_{TL} & U_{TR} \\ \hline U_{BL} & U_{BR} \end{array} \right) \rightarrow \left(\begin{array}{c|c|c} U_{00} & U_{01} & U_{02} \\ \hline U_{10} & U_{11} & U_{12} \\ \hline U_{20} & U_{21} & U_{22} \end{array} \right)$, $\quad \left(\begin{array}{c} B_T \\ \hline B_B \end{array} \right) \rightarrow \left(\begin{array}{c} B_0 \\ \hline B_1 \\ \hline B_2 \end{array} \right)$,
> $\quad\quad \left(\begin{array}{c|c} T_L & T_R \end{array} \right) \rightarrow \left(\begin{array}{c|c|c} T_0 & T_1 & T_2 \end{array} \right)$
> $\quad\quad$ **where** U_{11} and T_1 are $b \times b$ and B_1 has b rows
>
> ---
>
> $\quad$ $W := B_1$ $\hfill$ (COPYM)
> $\quad$ $W := \mathrm{trilu}(U_{11})^H W$ $\hfill$ (TRMM)
> $\quad$ $W := W + U_{21}^H B_2$ $\hfill$ (GEMM)
> $\quad$ $W := \mathrm{triu}(T_1)^{-H} W$ $\hfill$ (TRSM)
> $\quad$ $B_2 := B_2 - U_{21} W$ $\hfill$ (GEMM)
> $\quad$ $W := \mathrm{trilu}(U_{11}) W$ $\hfill$ (TRMM)
> $\quad$ $B_1 := B_1 + W$ $\hfill$ (AXPYM)
>
> ---
>
> $\quad$ **Continue with**
> $\quad\quad \left(\begin{array}{c|c} U_{TL} & U_{TR} \\ \hline U_{BL} & U_{BR} \end{array} \right) \leftarrow \left(\begin{array}{c|c|c} U_{00} & U_{01} & U_{02} \\ \hline U_{10} & U_{11} & U_{12} \\ \hline U_{20} & U_{21} & U_{22} \end{array} \right)$, $\quad \left(\begin{array}{c} B_T \\ \hline B_B \end{array} \right) \leftarrow \left(\begin{array}{c} B_0 \\ \hline B_1 \\ \hline B_2 \end{array} \right)$,
> $\quad\quad \left(\begin{array}{c|c} T_L & T_R \end{array} \right) \leftarrow \left(\begin{array}{c|c|c} T_0 & T_1 & T_2 \end{array} \right)$
> **endwhile**

Figure 2: An algorithm for applying block UT Householder transforms (from the left) to a matrix B [20], expressed in terms of BLAS-like operations (named in parentheses). This algorithm is very similar to the one used by LAPACK's `?larfb`, which is a building block to important operations such as the QR factorization. The BLAS routines associated with the highlighted subproblems are typically implemented such that the intermediate workspace matrix W (which is conformal to B_1) is copied and packed to contiguous memory [12, 13]. Since there is no way for the block Householder algorithm to access and reuse these packed matrices via the BLAS API, W ends up being redundantly packed three times during the course of the operation, incurring significant memory copy overhead.

this situation in the context of applying block Householder transformations [20], a crucial component of several important operations found in LAPACK, including the QR factorization.

- **BLAS has not grown with community's needs.** The initial BLAS standard lacked many basic dense linear algebra operations, for example, element-wise operations on matrices. (Figure 3 lists some of these unsupported operations.) Furthermore, the set of operations supported today by the level-1, level-2, and level-3 BLAS has not changed since its inception in the 1980s [22, 11, 10]. Indeed, the BLAST Forum attempted to standardize extensions to the BLAS that address some of the same issues discussed in the present paper. However, most of these extensions appear to be unsupported by major commercial (MKL [19], ACML [2], ESSL [18]) and open source (OpenBLAS [25], ATLAS [38]) implementations. (While some extensions may be supported by some BLAS libraries, the extensions are not supported uniformly by all such libraries, and hence using them would tend to reduce portability.) Furthermore, the operations missing from the BLAS go beyond those identified during the BLAST forum. A recent paper showed that by providing an optimized routine for applying multiple sets of Givens rotations, high-performance eigenvalue (EVD) and singular value decomposition (SVD) can be achieved using a simple restructuring of the tridiagonal and bidiagonal QR algorithms [35], which, while numerically robust, were previously thought to be inherently low-performance algorithms. And prior to that, other efforts identified fused kernels that helped accelerate reduction to condensed form

Level		Name	Operation	Description
1	V	SCAL2V	$y := \alpha\text{CJ}(x)$	Non-destructive vector scale.
		SETV	$\psi_i := \alpha, \quad \forall\psi_i \in y$	Set all vector elements to α.
		INVERTV	$\psi_i := \psi_i^{-1}, \quad \forall\psi_i \in y$	Invert all vector elements.
	M	AXPYM	$B := B + \alpha\text{CT}(A)$	Element-wise AXPY on matrices.
		COPYM	$B := \text{CT}(A)$	Element-wise COPY on matrices.
		SCALM	$B := \alpha\text{CJ}(B)$	Element-wise SCAL on matrices.
		SCAL2M	$B := \alpha\text{CT}(A)$	Element-wise SCAL2V on matrices.
		SETM	$\beta_{ij} := \alpha, \quad \forall\beta_{ij} \in B$	Set all matrix elements to α.
2		TRMV3	$y := \beta y + \alpha A x$	Non-destructive TRMV.
		TRSV3	$y := \beta y + \alpha A^{-1} x$	Non-destructive TRSV.
3		TRMM3	$C := \beta C + \alpha A B$	Non-destructive TRMM.
		TRSM3	$C := \beta C + \alpha A^{-1} B$	Non-destructive TRSM.

Figure 3: A sample of BLAS-like operations which were omitted from the BLAS. The functions CJ() and CT() denote optional conjugation and optional conjugation and/or transposition, respectively.

Level	Name	Operations		
1M	AXPYM	$B_C := B_C + \alpha_C A_R$	$B_C := B_C + \alpha_R A_R$	
	COPYM	$B_C := A_R$		
	SCAL2M	$B_C := \alpha_C A_R$	$B_C := \alpha_R A_R$	
	SETM	$B_C := \alpha_R$		
2	GEMV	$y_C := y_C + A_C x_R$	$y_C := y_C + A_R x_C$	$y_C := y_C + A_R x_R$
	GER	$A_C := A_C + x_C y_R^T$	$A_C := A_C + x_R y_C^T$	$A_C := A_C + x_R y_R^T$
	HER	$A_C := A_C + x_R x_R^T$		
	TRMV	$x_C := A_R x_C$		
	TRSV	$x_C := A_R^{-1} x_C$		
3	GEMM	$C_C := C_C + A_C B_R$	$C_C := C_C + A_R B_C$	$C_C := C_C + A_R B_R$
	HERK	$C_C := C_C + A_R A_R^T$		
	TRMM	$B_C := A_R B_C$		
	TRSM	$B_C := A_R^{-1} B_C$		

Figure 4: A representative sample of mixed domain operations which are not available via the BLAS interface. Subscripts $_R$ and $_C$ denote real and complex operands, respectively. Note that scalars, such as α and β, were omitted from the level-2 and level-3 operation expressions for clarity and typesetting reasons only. For the same reason we also omit the CJ() and CT() functions, which denote optional conjugation and optional conjugation and/or transposition, respectively.

operations, which serve as preprocesses to some implementations of EVD and SVD [17, 36].

- **Inability to compute with mixed domain operands.** With the exception of just a few level-1 routines, BLAS does not support operations where some operands are complex and some are real.[2] For example, the update of a complex vector by applying a real triangular matrix, via the TRMV

---

[2]The BLAST Forum proposed mixed domain and mixed precision functionality [8], but these extensions were never adopted by the community, perhaps because they targeted a set of less frequently-encountered use cases.

operation, is not possible. One could work around this problem by copying the real triangular matrix to a temporary complex matrix of the same size (initializing the imaginary components to zero) and then computing with this complex matrix using `ztrmv`. However, this approach incurs considerable cost in workspace, memory operations, and code complexity, not to mention an unnecessary factor of two increase in floating-point operations. Figure 4 contains a partial list of potentially useful variations on existing complex BLAS routines where at least one of the operands resides entirely in the real domain.[3]

- **Very few portable high-performance frameworks.** There exist very few portable frameworks that allow a hardware vendor to easily port existing BLAS operations to new architectures and/or design new operations. (Indeed, we suspect one of the reasons the BLAST extensions were never widely adopted can be attributed to the lack of a complete reference implementation which vendors could use as a guide when optimizing their own versions.) This typically means that a vendor who wants to create a BLAS library for his or her new architecture must start "from scratch." Granted, some work would be needed regardless, as there would be small kernels which require optimization at a very low (non-portable) level. But in our experience, we have found that large parts of a BLAS implementation need not differ from architecture to architecture. Yet, existing open source solutions lack the flexibility and infrastructural leverage we envision.[4]

Any one of these deficiencies may not be fatal for a given application, but to some developers, including us, the original BLAS may be simply inadequate. Perhaps most importantly, when taken together and multiplied over the hundreds or thousands of application developers who must spend time and effort to fashion nearly identical workarounds (which subsequently perform sub-optimally), these flaws in the BLAS constitute a considerable loss in human productivity.

3 Feature Set

In an effort to address the shortcomings of the BLAS, we propose the BLIS framework to ideally meet the following criteria:

- **Generalized storage.** The BLIS framework should support column-major, row-major, or general stride storage, and should do so for individual matrix operands, and should do so with minimal (if any) impact on performance. This means, for example, that the BLIS version of GEMM should support the operation $C = \beta C + \alpha AB$, where C is row-stored, A is column-stored, and B is stored with general stride (i.e., non-unit in both dimensions). In addition to this, we want BLIS to support basic element-wise operations on lower- or upper-stored trapezoidal matrices with arbitrary diagonal offsets.

- **Full support for the complex domain.** Operations should be primarily designed and developed in their most general form, which is typically in the complex domain. These formulations then simplify down to real domain analogues with relative ease (not to mention mathematical elegance). For mathematical "symmetry" between the real and complex domains, conjugation on real vectors and matrices should be allowed by the interface and treated as a *no-op*.

- **Multi-layered API with exposed kernels.** The framework should expose its implementations in various API layers. Low-level kernels will be available to experts who wish to design new linear algebra

---

[3]It should be noted that not all combinations of real and complex operands make sense for all operations. However, there are enough combinations that *do* make sense to merit inclusion.

[4]Among the most prominent open source BLAS implementations are netlib BLAS, OpenBLAS, and ATLAS. The netlib BLAS constitutes the original reference implementation and is quite portable, but makes very little reuse of internal code [7] and does not take advantage of the memory cache hierarchies found on nearly all modern architectures. Thus, it tends to perform very poorly. OpenBLAS, a fork of the discontinued GotoBLAS, offers competitive performance on a range of architectures, but is implemented as a BLAS library rather than a framework for building such libraries. ATLAS provides a system similar to that of OpenBLAS, except that in certain instances block sizes are automatically fine-tuned for cache performance [38]. However, once again, while ATLAS is well-suited to generating BLAS libraries with a fixed set of functionality, it does not appear designed to facilitate the implementation of new BLAS-like operations. Thus, we are left without a true framework for building high performance BLAS-like libraries.

operations, or reconfigure existing ones to meet the needs of the application or hardware. Operation implementations would further consist of higher-level blocked algorithm codes as well as lower-level (but still portable) "macro-kernels," which interface directly with architecture-aware micro-kernels. Optimizations can occur at various levels, in part thanks to exposed packing and unpacking facilities, which by default are highly parameterized (and thus flexible).

- **Portability _and_ high performance.** Portability should be a top priority. To achieve this, non-portable, architecture-specific codes should be compartmentalized into very small kernels.[5] As such, once these codes are provided, virtually all BLIS operations would be enabled at very little additional effort because all BLIS operations are expressed and implemented in terms of these kernels. Performance would be attained via careful implementation of the kernels, and thus would be inherited into multiple families of operations. For example, once an efficient matrix-matrix multiplication kernel is developed, nearly all level-3 operations should be enabled and attain high performance.

- **Ease of use.** The BLIS framework, and the library of routines it generates, should be easy to use for end users, experts, and vendors alike. A BLAS compatibility layer would be available to allow effortless linkage with existing BLAS-depending codes. Of course, this layer will only export the smaller subset of BLIS functionality that directly corresponds to the BLAS interface. Users who are willing to use BLIS's new BLAS-like interfaces will have access to the full set of BLIS operations and features. Motivated users and vendors would also have access to easy-to-use, object-based APIs.

- **Functionality that grows with the community's needs.** The BLIS framework should be, as its name suggests, a framework with which one can rapidly instantiate BLAS-like libraries rather than a library or static API alone. Furthermore, the framework should be extensible, allowing its developers (and others) to leverage existing components within the framework to support new operations as they are identified. As time goes on, it may be the case that new kernels are required in order to most efficiently support certain operations, such as the case with the application of multiple Givens rotations [35].

- **Support for mixed domain (and/or mixed precision) operations.** The BLIS framework should allow an expert to install additional kernels that enable various mixed domain operations. Similarly, operations that mix precision (and those that mix both domain and precision) should also be available to those who need them. However, neither should be mandatory. By making these features optional, we avoid unnecessarily burdening vendors with operations that their users do not need. But, importantly, the mixed domain/precision infrastructure should already be present, even when the corresponding kernels are left unoptimized. This way, a user can, at the very least, conveniently solve his or her mixed domain/precision problem, even if it requires the use of an unoptimized reference code.

- **Prioritize code re-use to minimize binary footprint.** A brute force or auto-generation approach to achieving the aforementioned goals tends to quickly lead to code bloat due to the multiple dimensions of variation that must be supported, namely: operation (i.e., GEMM, HEMM, TRMM, etc.); parameter (i.e., side, transposition, upper/lower structure, unit/non-unit diagonal); datatype (i.e., single-/double-precision real/complex); storage (i.e., row, column, general); and algorithm (i.e., partitioning path and kernel shape). The reason is simple: auto-generation techniques inherently tend to consider and optimize one specific case of one operation at a time. The BLIS framework should feature a holistic design that takes reasonable steps to limit the incursion of code bloat while still providing the targeted functionality. This is primarily achieved by careful abstraction, layering, and re-use of generic codes, subject to the constraint that performance-penalizing design decisions and coding conventions should be minimized.

Realizing all of these goals simultaneously in one software framework constitutes a significant improvement over the BLAS. While this list may seem ambitious, our preliminary findings suggest that attaining these goals is entirely within reach on modern architectures.

---

[5]By "small," we mean that these kernels are typically implemented as a single loop around highly-optimized code, typically expressed in an assembly language.

Name	Operation	Description
AXPYV	$y := y + \alpha \mathrm{CJ}(x)$	Accumulate scaled vector.
COPYV	$y := \mathrm{CJ}(x)$	Copy vector.
DOTV	$\rho := \mathrm{CJ}(x)^T \mathrm{CJ}(y)$	Dot product.
DOTXV	$\rho := \beta\rho + \alpha \mathrm{CJ}(x)^T \mathrm{CJ}(y)$	Generalized dot product.
INVERTV	$\psi_i := \psi_i^{-1}, \quad \forall \psi_i \in y$	Invert all vector elements.
SETV	$\psi_i := \alpha, \quad \forall \psi_i \in y$	Set all vector elements to α.
SCALV	$y := \alpha y$	In-place vector scale.
SCAL2V	$y := \alpha \mathrm{CJ}(x)$	Non-destructive vector scale.

Figure 5: A list of level-1v operations supported by the BLIS framework. The function CJ() denotes optional conjugation.

4 Supported Operations

Before discussing *how* the previously mentioned features are realized within BLIS, and other implementation details, we first wish to provide high-level descriptions of the core operations supported by the framework. The curious reader can find proposed APIs for these BLIS operations listed in Appendix A.

The BLIS framework organizes its operations into five classes of operations, where each class contains related operations. As with the BLAS, we refer to these classes as "levels."

4.1 Level-1v: Vector-vector Operations

The level-1 BLAS specification contains operations on vectors. The BLIS framework supports similar operations, which we collectively refer to as level-1v BLIS.

Figure 5 contains a list of level-1v operations supported by BLIS. Note that in the mathematical expressions shown here, an instance of the CJ() function denotes the opportunity to optionally conjugate an operand.[6] We emphasize that this list is incomplete and is only meant to provide a representative sample of operations which can (and should) be supported by BLIS.

Experienced users will immediately notice a few differences between level-1 BLAS and level-1v BLIS, most notably with regards to expanded functionality. We briefly discuss the main differences:

- Level-1v BLIS operations are presented with a "v" appended to their names. This is done to differentiate them from similar operations that take matrix (instead of vector) arguments, which will be discussed in Section 4.2

- All operations allow the user to optionally conjugate vectors that are strictly input operands. For example, a developer could use AXPYV to compute with a vector x as if it were conjugated. Note that this optional conjugation does *not* affect the state of vector x during the course of the computation, thus maintaining thread safety.

- New operations are now available, such as DOTXV and SCAL2V. The former is useful when one wishes to scale and/or accumulate a dot product into an existing scalar, while the latter can be used to simultaneously copy and scale a vector. Other routines are provided for similar convenience, such as SETV and INVERTV.

The net effect of the changes above will be to simplify certain unblocked algorithms, including those for operations typically found within LAPACK [3].

Name	Operation	Description
AXPYM	$B := B + \alpha \text{CT}(A)$	Accumulate scaled matrix.
COPYM	$B := \text{CT}(A)$	Copy matrix.
SETM	$\beta_{ij} := \alpha, \quad \forall \beta_{ij} \in B$	Set all matrix elements to α.
SCALM	$B := \alpha B$	In-place matrix scale.
SCAL2M	$B := \alpha \text{CT}(A)$	Non-destructive matrix scale.

Figure 6: A list of level-1m operations supported by the BLIS framework. The functions CJ() and CT() denote optional conjugation and optional conjugation and/or transposition, respectively. Each of the operations listed above is supported on dense as well as lower or upper triangular/trapezoidal matrices with arbitrary diagonal offsets, such that only the specified region(s) are referenced and/or updated.

4.2 Level-1m: Vector-Vector Operations on Matrices

One of the more obvious omissions from the BLAS can be found in its lack of element-wise operations on matrix operands. Many of the operations one might want to perform on vectors, such as AXPYV, COPYV, or SCALV, can also be meaningfully performed on matrices. One needs look no further than computational tools such as MATLAB to find this sort of functionality implemented and easily accessible [24].

As alluded to previously, there are workarounds for this situation that are not unreasonable for certain users. For example, if one wishes to perform $B := B + \alpha A$ (i.e., an AXPY operation on matrices), he or she can insert a loop over an AXPYV-like routine that scales and accumulates the matrices one column at a time. But what if the matrices are stored by rows? To achieve unit stride, the loop would need to accordingly be over rows rather than columns. Or what if the user wishes to perform a conjugation to the input matrix A, or a transposition (or both)? What if the matrices are triangular (or symmetric or Hermitian) in structure and thus only stored in the lower or upper triangle? Finally, what if the stored region is not strictly triangular, but rather dependent upon the precise location of the matrices' main diagonals? It is easy to see how implementing such functionality using only AXPYV could prove daunting for many BLAS users. Yet another problem with the aforementioned solution is that looping over a top-level AXPYV interface (such as via the BLAS) results in repeated error checking on matrix columns (or rows), which in this case one would preferably avoid in favor of a single set of checks up-front. Thus, leaving the application developer to implement a custom solution is not ideal.

The BLIS framework supports several level-1m operations on matrices. Of these, a representative list is shown in Figure 6. These operations are provided with enough parameterization and generality to handle any combination of the situations listed above, in addition to a few situations not mentioned such as general stride storage and implicit unit diagonals.

4.3 Level-1f: Fused Vector-vector Operations

It has been shown elsewhere that so-called "fused" kernels can be used to more efficiently implement certain operations that are rich in $\mathcal{O}(n^2)$ floating-point operations (flops) performed on $\mathcal{O}(n^2)$ data [17, 36]. The performance of such operations is greatly hindered by two facts. First, on modern architectures, memory load and store instructions tend to be much more expensive than floating-point instructions. Second, since the ratio of flops to memory operations (memops) is typically small, the computation often stalls while waiting for new data to be fetched from (or stored back to) memory.

While these level-2 operations will never perform as well as their level-3 brethren (which inherently call for $\mathcal{O}(n^3)$ flops to be performed on $\mathcal{O}(n^2)$ data), non-trivial performance gains can still be achieved by carefully reusing values once they have been loaded into the processor core's registers. We refer to this technique as *register-level fusing* [36].[7]

---

[6] As one might expect, this conjugation reduces to a *no-op* when the operand is stored in the real domain.

[7] One may also use cache-level fusing, which, as its name suggests, targets reuse of data while it resides in cache memory. However, in most situations, fusing at the register level will almost always yield superior performance because it allows one to avoid certain memory accesses altogether, rather than simply reduce their cost.

Name	Operation	Description
AXPY2V	$\begin{cases} y := y + \alpha_0 \text{CJ}(x_0) \\ y := y + \alpha_1 \text{CJ}(x_1) \end{cases}$	Fused AXPYV pair with shared y.
DOTAXPYV	$\begin{cases} \rho := \text{CJ}(x)^T y \\ z := z + \alpha \text{CJ}(x) \end{cases}$	Fused DOTV-AXPYV with shared x.
AXPYF	$\begin{cases} y := y + \alpha \text{CJ}(\chi_0)\text{CJ}(a_0) \\ \quad\vdots \\ y := y + \alpha \text{CJ}(\chi_{f-1})\text{CJ}(a_{f-1}) \end{cases}$	Fused sequence of f AXPYV with shared y, where f is implementation-dependent.
DOTXF	$\begin{cases} \psi_0 := \beta\psi_0 + \alpha \text{CJ}(a_0)^T \text{CJ}(x) \\ \quad\vdots \\ \psi_{f-1} := \beta\psi_{f-1} + \alpha \text{CJ}(a_{f-1})^T \text{CJ}(x) \end{cases}$	Fused sequence of f DOTXV with shared x, where f is implementation-dependent.
DOTXAXPYF	$\begin{cases} y_0 := \beta y_0 + \alpha \text{CJ}(A^T)\text{CJ}(x_0) \\ y_1 := y_1 + \alpha \text{CJ}(A)\text{CJ}(x_1) \end{cases}$	Fused DOTXF-AXPYF with shared $(a_0, \cdots, a_{f-1}) = A$, where f is implementation-dependent.

Figure 7: A list of level-1f operations supported by the BLIS framework. The function CJ() denotes optional conjugation. Note that above, f corresponds to the implementation-dependent fusing factor for the operation in question. These fusing factors may be queried from within the framework.

for $i = 0 : m - 1$
    LOAD $y_i \rightarrow \psi_1$
    $\psi_1 := \beta\psi_1$
    STORE $y_i \leftarrow \psi_1$
endfor
for $j = 0 : n - 1$
    LOAD $x_j \rightarrow \chi_1$
    $\chi_1 := \alpha\chi_1$
    for $i = 0 : m - 1$
        LOAD $A_{ij} \rightarrow \alpha_1$
        LOAD $y_i \rightarrow \psi_1$
        $\psi_1 := \psi_1 + \alpha_1\chi_1$
        STORE $y_i \leftarrow \psi_1$
    endfor
endfor

$2mn + m + n$ flops
$3mn + 2m + n$ memops

for $i = 0 : m - 1$
    LOAD $y_i \rightarrow \psi_1$
    $\psi_1 := \beta\psi_1$
    STORE $y_i \leftarrow \psi_1$
endfor
for $j = 0 : n - 1 : 4$
    LOAD $x_{j:j+3} \rightarrow \chi_{1:4}$
    $\chi_{1:4} := \alpha\chi_{1:4}$
    for $i = 0 : m - 1$
        LOAD $A_{i,j:j+3} \rightarrow \alpha_{1:4}$
        LOAD $y_i \rightarrow \psi_1$
        $\psi_1 := \psi_1 + \alpha_1\chi_1 + \alpha_2\chi_2$
        $\qquad\qquad + \alpha_3\chi_3 + \alpha_4\chi_4$
        STORE $y_i \leftarrow \psi_1$
    endfor
endfor

$\left. \vphantom{\begin{array}{c}a\\a\\a\\a\end{array}} \right\}$ AXPYF(α, A_1, x_1, y)
where
$A_1 = A_{j:j+3}$
$x_1 = x_{j:j+3}$

$2mn + m + n$ flops
$\frac{3}{2}mn + 2m + n$ memops

Figure 8: Left: A column-based (i.e., AXPYV-based) algorithm for computing the GEMV operation $y := \beta y + \alpha A x$. This algorithm requires $3mn + 2m + n$ memory operations (ignoring the loading of α and β). Center: By unrolling and fusing 4 iterations at a time, we are able to eliminate 3 out of every 4 loads and stores on vector y while still preforming the same number of flops. The affected code is highlighted in grey and corresponds to an AXPYF kernel operation with a fusing factor of $f = 4$. Thus, by implementing GEMV in terms of this fused kernel, we reduce the total number of memory operations incurred by nearly a factor of two.

Name	Operation(s)	Description
GEMV	$y := \beta y + \alpha \text{CT}(A)\text{CJ}(x)$	General matrix-vector multiply.
GER	$A := A + \alpha \text{CJ}(x)\text{CJ}(y)^T$	General rank-1 update.
HEMV	$y := \beta y + \alpha \text{CJ}(A)\text{CJ}(x)$	Hermitian matrix-vector multiply.
HER	$A := A + \alpha \text{CJ}(xx^H)$	Hermitian rank-1 update.
HER2	$A := A + \alpha \text{CJ}_x(x)\text{CJ}_y(y)^H$ $+ \bar{\alpha}\text{CJ}_y(y)\text{CJ}_x(x)^H$	Hermitian rank-2 update.
SYMV	$y := \beta y + \alpha \text{CJ}(A)\text{CJ}(x)$	Symmetric matrix-vector multiply.
SYR	$A := A + \alpha \text{CJ}(xx^T)$	Symmetric rank-1 update.
SYR2	$A := A + \alpha \text{CJ}_x(x)\text{CJ}_y(y)^T$ $+ \alpha \text{CJ}_y(y)\text{CJ}_x(x)^T$	Symmetric rank-2 update.
TRMV	$x := \alpha \text{CT}(A)x$	Triangular matrix-vector multiply.
TRSV	$x := \alpha \text{CT}(A)^{-1}x$	Triangular solve.

Figure 9: A list of level-2 operations supported by the BLIS framework. The functions CJ() and CT() denote optional conjugation and optional conjugation and/or transposition, respectively. A subscript is used in conjuction with CJ() to indicate two instances of the same conjugation.

Figure 7 shows a list of operations supported by the BLIS framework which, if implemented with register-level fusing, would facilitate higher levels of performance for key subproblems found within various level-2 algorithms. For example, the operation GEMV may be cast as a series AXPYV (or DOTV) operations. If groups of f AXPYV subproblems are fused together, then certain operand elements, once loaded into registers, may be reused $f - 1$ times. This reuse of data typically results in a higher performance, as less time is spent waiting for operand elements to arrive from memory. Naturally, the implementor would wish to maximize this reuse of data, and so he or she would aim to implement these level-1f operations with as large of a "fusing factor" f as possible, given the constraints of the floating-point datatype and the size of the architecture's register set. (While these fusing factors are implementation-dependent, they are made available to other framework components, as well as application code, via query macros.)

Figure 8 illustrates the previous example with pseudo-code that exposes individual load and store instructions. Here, it is easy to see that an AXPYF implemented with a fusing factor of $f = 4$ reduces the number of memops needed to compute GEMV by nearly a factor of two.

An alternative fusing may be achieved via the DOTXF kernel operation. Similarly, the other level-1f operations listed in Figure 7 may be used to optimize various other level-2 operations, which are discussed in the next section.

4.4 Level-2: Matrix-Vector Operations

The level-2 operations available in the BLIS framework are largely similar to the corresponding operations within the level-2 BLAS. Figure 9 lists these operations. The differences between the two are enumerated as follows:

- For all operations where a matrix or vector is strictly an input operand, that operand may be used as if it were conjugated. This option is in addition to the transposition or conjugate-transposition options already allowed by the BLAS.

- Unlike in the BLAS, the TRMV and TRSV operations scale their matrix-vector products by a scalar α before overwriting their results to vector x.

- BLIS supports non-destructive TRMV and TRSV operations via TRMV3 and TRSV3, respectively.

- While not obvious from Figure 9, BLIS implements complex SYMV, SYR, and SYR2, whereas the BLAS omits these operations.

Level-2 Operation	Matrix storage	By optimizing this kernel, level-2 performance is...		
		Improved by	Further improved by	Optimized by
GEMV, TRMV, TRSV	column-stored	AXPYV		AXPYF
	row-stored	DOTXV		DOTXF
HEMV, SYMV	column-stored	DOTXV + AXPYV	DOTAXPYV	DOTXAXPYF
	row-stored			
GER, HER, SYR	column-stored	AXPYV		
	row-stored			
HER2, SYR2	column-stored	AXPYV		AXPY2V
	row-stored			

Figure 10: Here we present optimization dependencies for each of the level-2 operations supported in the BLIS framework. The table shows which level-1v and/or level-1f operations must be implemented and optimized in order to achieve higher performance for a given level-2 operation. Note that for general stride matrix storage, unit stride through memory cannot be achieved, and thus when given the choice (i.e., for GEMV, TRMV, and TRSV) either DOT-based or AXPYV-based kernels may be used, though the former may be preferred due to the inherently lower ratio of memops to flops associated with DOT-based level-2 algorithms.

Like `libflame`, the BLIS framework provides multiple algorithmic variants[8] for many of the operations it implements [34]. In the case of level-2 operations, it is sometimes advantageous to choose a different variant depending on how the matrix operand is stored. To achieve unit stride[9] through memory, matrices that are column-stored should be computed upon with variants that traverse individual columns. Similarly, row-based variants tend to be well-suited to matrices stored by rows. (Note that in either case, a transposition can cause this affinity to flip.) Thus, for example, a GEMV variant based on AXPYV is appropriate for column-stored matrices (or transposed row-stored matrices), while GEMV based on DOTXV should be preferred for row-stored matrices (or transposed column-stored matrices). The BLIS framework allows us to support both of these situations. (A general strided matrix is probably best computed upon via a DOTXV-based kernel because it exhibits an inherently lower ratio of memops to flops than one based on AXPYV.)

Recall that in Section 4.3 we discussed various level-1f BLIS operations that were fused to avoid redundant memops within level-2 operations. Figure 10 shows which level-2 operations benefit from the optimization of various level-1v and level-1f operations. Importantly, our preliminary experience shows that optimizing these kernels is sufficient in order to achieve competitive level-2 performance on modern architectures. (Performance results for level-2 operations that employ these kernel operations may be found in Section 6.2.)

To put things concretely, Figure 10 reveals that optimizing *only* DOTXV and AXPYV is sufficient to enable accelerated implementations of *all* level-2 operations. These implementations would be efficient regardless of whether the matrix it operates upon was stored by rows or columns. If one were to go further and implement architecture-tuned versions of AXPY2V, DOTXF, AXPYF, and DOTXAXPYF, in addition to AXPYV, one could use architecture-agnostic components of the BLIS framework to quickly instantiate a set of level-2 BLIS (and BLAS) that was even more fully optimized.

4.5 Level-3: Matrix-Matrix Operations

Figure 11 lists the core set of level-3 operations supported by the BLIS framework. Indeed, these operations largely correspond to the level-3 operations provided by the BLAS. The most prominent differences mirror those found between the level-2 BLIS and BLAS: namely, the ability to conjugate individual input operands.

Just as the level-2 operations are built atop kernels which may be optimized to enable high performance, level-3 operations are likewise implemented in terms of smaller kernels. This idea is not new [12, 13, 15]. Section 5 discusses how these level-3 operations are implemented in the BLIS framework so that flexibility (i.e., generality), portability, and high performance are simultaneously achieved.

---

[8]Henceforth, we abbreviate the term "algorithmic variant" to simply "variant."

[9]Unit stride through memory is usually, though not universally, desired on modern cache-based systems because it typically increases spatial locality of data in the cache hierarchy.

Name	Operations(s)	Description
GEMM	$C := \beta C + \alpha \textsc{ct}(A)\textsc{ct}(B)$	General matrix-matrix multiply.
HEMM	$C := \beta C + \alpha \textsc{cj}(A)\textsc{ct}(B)$, or $C := \beta C + \alpha \textsc{ct}(B)\textsc{cj}(A)$	Hermitian matrix-matrix multiply.
HERK	$C := \beta C + \alpha \textsc{ct}_A(A)\textsc{ct}_A(A)^H$	Hermitian rank-k update.
HER2K	$C := \beta C + \alpha \textsc{ct}_A(A)\textsc{ct}_B(B)^H$ $\quad + \bar{\alpha}\textsc{ct}_B(B)\textsc{ct}_A(A)^H$	Hermitian rank-$2k$ update.
SYMM	$C := \beta C + \alpha \textsc{cj}(A)\textsc{ct}(B)$, or $C := \beta C + \alpha \textsc{ct}(B)\textsc{cj}(A)$	Symmetric matrix-matrix multiply.
SYRK	$C := \beta C + \alpha \textsc{ct}_A(A)\textsc{ct}_A(A)^T$	Symmetric rank-k update.
SYR2K	$C := \beta C + \alpha \textsc{ct}_A(A)\textsc{ct}_B(B)^T$ $\quad + \alpha \textsc{ct}_B(B)\textsc{ct}_A(A)^T$	Symmetric rank-$2k$ update.
TRMM	$B := \alpha \textsc{ct}(A)B$, or $B := \alpha B \textsc{ct}(A)$	Triangular matrix-matrix multiply.
TRMM3	$C := \beta C + \alpha \textsc{ct}(A)\textsc{ct}(B)$, or $C := \beta C + \alpha \textsc{ct}(B)\textsc{ct}(A)$	Non-destructive triangular matrix-matrix multiply.
TRSM	$B := \alpha \textsc{ct}(A)^{-1}B$, or $B := \alpha B \textsc{ct}(A)^{-1}$	Triangular solve with multiple right-hand sides.

Figure 11: A list of level-3 operations supported by the BLIS framework. The functions CJ() and CT() denote optional conjugation and optional conjugation and/or transposition, respectively. A subscript is used in conjuction with CT() to indicate two instances of the same conjugation and/or transposition.

5   Implementation techniques

Implementing the BLIS framework is fraught with challenges, especially if one is to avoid explosive code bloat. The last bullet at the end of Section 3 touches upon several "dimensions" of generality to be captured by BLIS. If the framework itself was to be manageable and maintainable going forward, we had to find ways of collapsing each of these dimensions as much as possible while maintaining the desired functionality. Here, we provide a detailed discussion of some of the implementation techniques and design decisions that allow us to achieve most of our functionality goals for BLIS without incurring significant increases in source (or object) code. We consider these software engineering details to be important because they facilitate the primary contributions of this paper.

For reasons that will seem clearer later on, we discuss our techniques for implementing the level-3 BLIS separately from that of the other levels.

5.1   Level-3 techniques

Level-3 operations make up perhaps the most-used portion of the BLAS.[10] These operations are also typically the most difficult to implement, in part because the space of potential solutions is much larger than that of level-2 and lower operations. It may, therefore, seem counterintuitive that the portion of the BLIS framework associated with level-3 operations is actually *simpler* (and smaller) than that of the remaining operations. We now walk through the methods used to implement this functionality.

5.1.1   Operation

Perhaps the most obvious framework dimension we wish to manage is that of the operations supported. Despite seemingly significant differences, all level-3 operations are inherently very similar. It was observed

---

[10]This is likely no coincidence, as the level-3 operations are typically capable of achieving a large fraction of peak performance on modern architectures. Thus, application developers have an incentive to express their computation in terms of level-3 operations whenever possible.
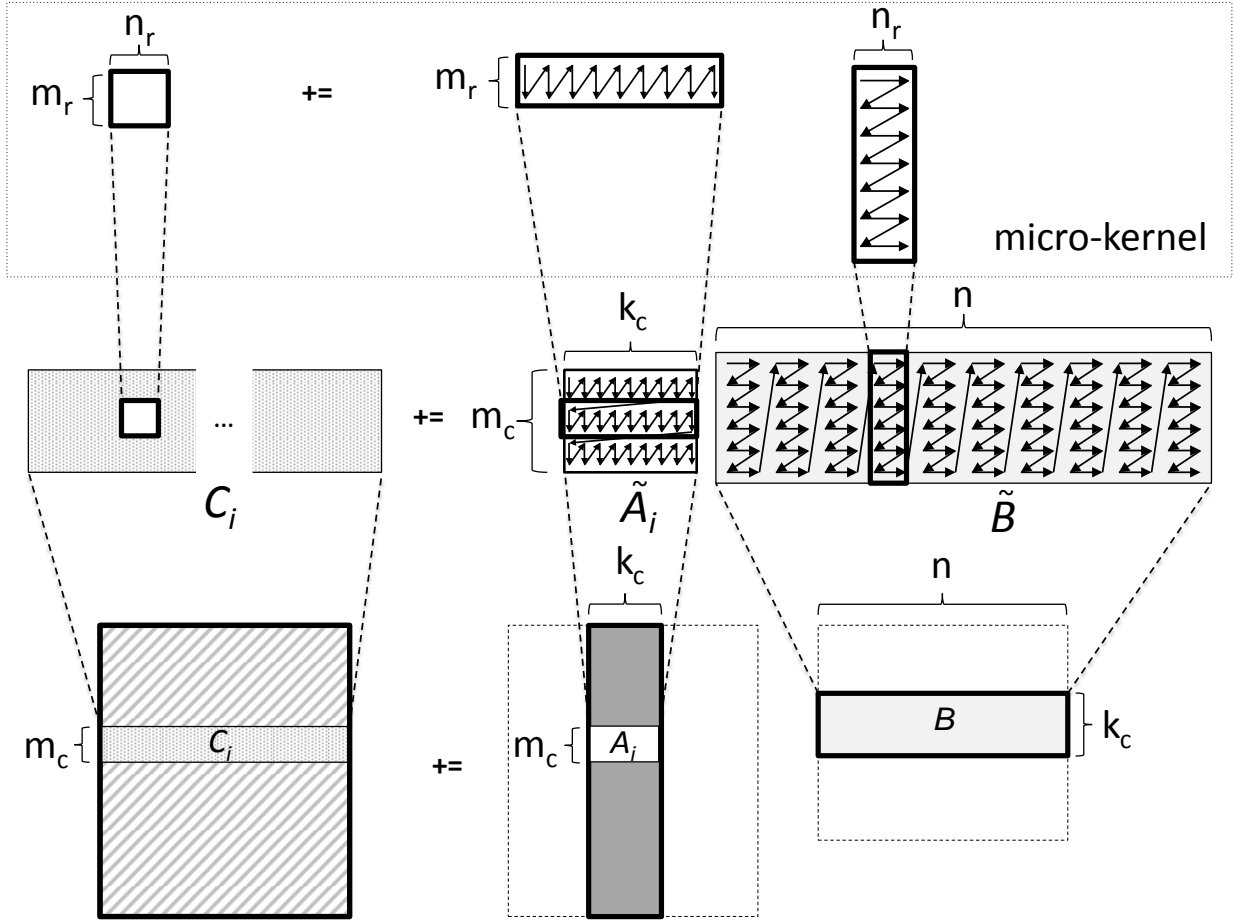
Figure 12: Illustration of the various levels of blocking and related packing when implementing GEMM in the style of [12]. Here, m_c and k_c serve as cache block sizes used by the higher-level blocked algorithms to partition the matrix problem down to a so-called "block-panel" subproblem (depicted in the middle of the diagram), implemented in BLIS as a portable macro-kernel. Similarly, m_r and n_r serve as register block sizes for the micro-kernel in the m and n dimensions, respectively, which also correspond to the length and width of the individual packed panels of matrices $\tilde{A}_i$ and $\tilde{B}$, respectively.

by [21] that level-3 operations can be cast in terms of general rank-k update (GEMM). The authors of [13] built on this idea but observed that packing facilities could be modified to induce all level-3 operations via the same low-level kernels that support the implementation of GEMM. Let us use Figure 12 to briefly describe this algorithm. A matrix multiplication with presumably large m, n, and k dimensions is first partitioned along k with a cache block size of k_c, creating rank-k subproblems (depicted at the bottom of Figure 12). At this stage, B is packed to contiguous memory (labeled as $\tilde{B}$) in a special format that facilitates unit stride at the lowest level of computation. Each rank-k subproblem is then partitioned along the m dimension with a cache block size of m_c, creating block-panel subproblems. The current $m_c \times k_c$ block A_i is then packed to $\tilde{A}_i$, which is once again stored in a special format. This block-panel subproblem that remains ($C_i := C_i + \tilde{A}_i\tilde{B}$) is then implemented as a highly optimized, assembly-coded kernel. This kernel will continue on to partition the matrices along the n, m_c, and finally k_c dimensions. Now, to extend this algorithm to other level-3 operations, the authors of [13] point out that sometimes the only change required is in the packing facility. For example, for HEMM and SYMM, if $\tilde{A}_i$ is packed in such a way that blocks that intersect the diagonal are made dense (since only one triangle is stored), then the block-panel kernel can be used unmodified. However,

HERK, SYRK, HER2K, and SYR2K require that the block-panel kernel be specialized because only the part of C_i that is stored is to be updated. Similar complications arise with TRMM and TRSM due to the potential triangular nature of A_i, which thus require additional specialization. To avoid this proliferation of kernels, the BLIS framework uses a somewhat different layering.

In BLIS, the fundamental kernel encompasses only the inner-most computation that iterates over the k_c dimension (depicted at the top of Figure 12).[11] Because this kernel is a much smaller, more basic operation, we refer to it as a "micro-kernel." Similarly, we use the term "macro-kernel" to refer to the two loops above the micro-kernel (i.e., those loops that partition along the n and m_c dimensions, which correspond to the outer and middle loops in a Goto-style kernel). This layering provides two key benefits:

- It allows virtually all of the architectural details to be confined to a much smaller, simpler kernel of computation; and

- It allows nearly all of the operation-specific nuances of the level-3 kernels for HERK, SYRK, HER2K, SYR2K, TRMM and TRSM to be factored out to different macro-kernels, each of which is a portable, pre-supplied component of the BLIS framework.[12]

An additional benefit of the macro-kernel/micro-kernel design used by BLIS is that it facilitates portable edge-case handling. That is, the edge case handling occurs entirely within the macro-kernel. When $m \bmod m_r \neq 0$ or $n \bmod n_r \neq 0$, the same micro-kernel is used, except the result is stored to a temporary $m_r \times n_r$ buffer. (This is always possible because packed matrix dimensions are inflated to a multiple of m_r or n_r, as necessary, so that the packing facility can perform zero-padding.) Then, only those elements of the local buffer corresponding to stored values are written to the output matrix. No additional micro-kernels are required to handle edge cases. This approach requires some extra macro-kernel logic, but results in only an additional $\mathcal{O}(n)$ memory operations, and thus the net impact on performance is, under modest assumptions, negligible.

Thus, the *only* code that must be optimized for a given architecture is that of the micro-kernel.[13]

Once the micro-kernel is written, the vendor need only choose appropriate cache and register block sizes to achieve a full implementation of GEMM. From there, virtually no additional work is needed (on the part of the vendor) to instantiate high-performance implementations of the remaining level-3 operations. This leveraging is immediate and automatic because the BLIS framework already encodes the differences in the various level-3 operations into (1) the portable macro-kernels, and (2) the framework's highly parameterized packing facility.

5.1.2 Parameter case

The dimension most vulnerable to code duplication is arguably that of operational parameters. Each BLAS operation requires that multiple parameter cases must be supported. For example, a full implementation of GEMM must allow nine parameter cases. This stems from the three accepted values for each of the `transA` and `transB` arguments that correspond to the two input matrices, A and B. In BLIS, GEMM must support 16 cases due to the extra conjugation-without-transposition option. The TRMM and TRSM operations must support 24 and 32 cases in BLAS and BLIS, respectively, due to the combinatorial impact of the `side`, `uplo`, `trans`, and `diag` parameters. It is easy to see how a BLAS-like library could quickly become unwieldy and difficult for vendors to manage.

---

[11] Our use of a micro-kernel as the basic kernel of computation, as well as its general implementation (as a series of rank-1 updates), is strikingly similar to the corresponding kernel of another effort, which investigates implementing a linear algebra co-processor in hardware [27, 26]. Interestingly, the two projects isolated their respective kernels independently of one another, despite the kernels' similarities and the projects' collaborative proximity to one another.

[12] To maximize TRSM performance, a vendor would typically need to supply an additional micro-kernel that fuses a GEMM subproblem and TRSM subproblem in order to avoid redundant memory operations. In practice, we have observed that 80–90% of potential performance is attainable even when this fused micro-kernel is not supplied. We illustrate this performance gap in Section 6.3.

[13] While BLIS sequesters almost all non-portable code within the micro-kernels, these codes must still be implemented with care. How to optimally implement these micro-kernels is heavily dependent on features of the memory hierarchy and instruction set, and thus is well beyond the scope of this document.

The BLIS framework allows level-3 operations to handle and implement most of their various parameter cases via the packing facility. But before discussing this further, we first broach the topic of how BLIS organizes its data. At higher levels of execution, the level-3 operands are expressed and tracked with "objects." This simple abstraction, implemented as C `struct`s, allows the framework to express its algorithms at a high level.[14] These objects encapsulate nearly all properties about a matrix and are "tagged" in various ways depending on the operation's parameterization. For example, if a particular instance of GEMM requires that matrix A be used as if it were conjugated, the implementation toggles the conjugation bit of the object associated with the matrix. As one might expect, the actual elements of matrix A are never modified, and so the conjugation of the matrix object is implicit. Similarly, the object determines whether the matrix requires transposition[15], the location (offset) of its diagonal, whether that diagonal is implicitly unit, its structure (symmetric, Hermitian, triangular), and its storage: lower-stored, upper-stored, dense, or zero.

Now, when an object is about to be packed, its properties are inspected to determine exactly how the packing should unfold. Transposition, conjugation, and unit diagonalization can all be applied at no cost during the packing (because the matrix data must be loaded from and/or stored to memory anyway). In general, upon completion, the packed matrix reflects *explicitly* what its corresponding unpacked object reflects *implicitly*. If the matrix is triangular and stored only in the upper or lower region, only that region will be packed. Similarly, if the matrix is Hermitian and "densification" is requested, the packed matrix will be packed so that the unstored region contains the conjugate-transpose of the stored region. (Packed objects also inherit diagonal offset information from their unpacked brethren, which greatly simplifies the handling of non-square blocks with structure.) The benefit of handling these parameter case combinations in the packing routine is that the lower-level codes, including the micro-kernel and, to some extent, the macro-kernels, are kept extraordinarily simple; that is, these lower-level codes need only support one or two very plain parameter cases (i.e., no transposition, no conjugation, non-unit diagonal, etc.).

We say "one or two" cases must be supported because, for some operations that assume operand structure, the BLIS framework explicitly implements both lower- and upper-stored cases (via separate variants and/or macro-kernels), either because a different region of the matrix is being updated (HERK, HER2K, SYRK, SYR2K) or because the algorithms promote movement in opposite directions (TRMM, TRSM). While this approach results in a modest increase in source code, these components of the framework are portable and would typically not require any modification when porting across architectures.

What remains? The `side` parameter, found in HEMM, SYMM, TRMM, and TRSM. The `side` parameter, since it describes the orientation of the matrices rather than the matrices themselves, is not embedded within the object, and so it remains a parameter of the operation throughout most layers of the BLIS framework. But, the `side` parameter need not be supported directly. Instead, in BLIS, one case is typically implemented (say, the left side case) and assumed by the mid- and lower-level algorithms. The right side case is handled early on at higher levels of the implementation; if the right side case of the operation is detected, it is induced by calling the left side case with transposed operands.[16]

5.1.3 Datatype

A prominent framework dimension we must manage is that of floating-point datatype. Specifically, we wish to minimize the amount of code that must reside within the framework (and the amount of code the vendor must optimize) to support the floating-point datatypes expected by the community. BLAS supports

---

[14]Unlike in `libflame`, the functions used to query and modify object properties are carefully implemented as C preprocessor macros. This technique avoids a great deal of potential function call overhead and also provides the compiler with more opportunities for optimization. Furthermore, objects are passed by address rather than by value, thereby minimizing the cost of actual function calls. In this way, the BLIS framework affords its developers (and expert users) an object-based API that is lightweight and thus performance-friendly.

[15]Since BLIS supports both row and column strides, we may, alternatively, induce transpositions at any time before packing by swapping the row and column strides, swapping the length and width dimensions, and if necessary, toggling the region stored and the diagonal offset. In some situations, we prefer this method because it allows us to further collapse the number of algorithms needed. TRMM serves as a good example of this, since some variants of this operation (those that partition matrix A) move in opposite directions for the lower- and upper-stored cases.

[16]In practice, a little more sophistication is sometimes employed than transposing every matrix operand. For example, with HEMM, the right side case is induced by toggling the conjugation of matrix A, which is mathematically equivalent to a transposition. (This holds in both the real and complex domains.)

operations on four standard datatypes: single-precision real, double-precision real, single-precision complex, and double-precision complex. This facet of the implementation is prominently encoded (usually) as the first character in the routine name: s, d, c, and z, respectively. For the level-3 BLAS, the implicit assumption is that the datatype of the operation corresponds to the datatype of all matrix operands. In BLIS, we instead associate an independent datatype property with each matrix object. So the question becomes, how do we handle arbitrary combinations of domains and precisions among matrices?

Let us first handle the special case that merely mimics the datatype support of the BLAS (i.e., the case where all matrices contain elements of the same datatype). The sole noteworthy expansion of code that is proportional to the number of datatypes supported (in our case, four) is that of the GEMM micro-kernel. For each datatype supported, one micro-kernel must be provided (and optimized, if high performance is desired). If a fused GEMM-TRSM micro-kernel is desired, this code must also be instantiated for each datatype. This is the *only* datatype-specific source code that is required of homogeneous-typed level-3 BLIS. All other level-3 codes, including high-level object-based algorithmic variants and mid-level macro-kernels, are datatype-agnostic.[17]

Now let us consider the cases where matrix datatypes are not all identical. We remind the reader that BLIS does not require the vendor to support these mixed-datatype cases. If this functionality is desired, however, very little effort is needed because mixed datatype support is largely handled by reusable framework code. When mixed-datatype operations require matrices to be typecast to a higher or lower precision, casting is performed as needed when the matrices are packed. When domains must be mixed, the framework activates logic that utilizes existing same-datatype micro-kernels. For example, the matrix product AB, where $A \in \mathbb{C}^{m \times k}$ and $B \in \mathbb{R}^{k \times n}$ and both are column-stored, can be thought of as a real matrix product $\hat{A}B$ where $\hat{A} \in \mathbb{R}^{2m \times k}$ since complex numbers are stored in such a way that they are, to a micro-kernel, indistinguishable from a pair of real numbers.

5.1.4 Storage format

Recall that a central goal of BLIS is to support matrices stored not only in column-major order and row-major order, but also general stride.[18] Furthermore, we wish to allow operations to mix storage types among matrix operands. While this may appear to be an unmanageable number of cases to support, especially for three-operand operations such as GEMM, the impact on the level-3 BLIS framework is quite minimal. (In fact, there are situations where separate row and column strides actually simplifies the implementation.) This expanded functionality is achieved by virtue of two aspects of the framework.

- Since input matrices are already universally packed to contiguous storage for performance reasons, we can easily design the packing facility to read these input operands according to their individual row and column strides.

- Generalized storage is supported for output operands because the BLIS micro-kernel interface is explicitly parameterized with row and column strides for the output matrix C (which need not be packed).

These two features combine to allow row storage, column storage, or general stride storage for any operand, as well as any combination across operands, for all supported operations.

5.1.5 Algorithm

As previously mentioned, BLIS provides multiple algorithmic variants for each operation supported. Each algorithmic variant is derived from a *loop invariant*, which, among other things, expresses the way matrix operands are partitioned and the direction(s) in which the variant moves through these matrices [14, 16, 5, 32]. The variant set provided by BLIS for each level-3 operation can be thought of as "complete" in the sense that there exists at least one variant that partitions each dimension of the problem. For example, the

---

[17]Lower-level macro-kernels are kept datatype-agnostic in a manner similar to that of unblocked variants for level-1v, -1m, and -2 operations. This topic is discussed in Section 5.2.3

[18]Note that the former two storage layouts are simply special cases of general stride where the row stride is unit and the column stride is unit, respectively.

framework provides three basic blocked variants for GEMM: one for partitioning along each of the m, n, and k dimensions. (Typically, we omit from BLIS variants that traverse operands backwards when a forward-moving variant is available and applicable.) In a more sophisticated level-3 (or LAPACK-like) algorithm, one may have an operation with two or more subproblems, where each subproblem belongs to a different family of operations. This recursion may involve several levels of partitioning and blocking, usually to promote cache reuse and locality, but also sometimes for parallelism or out-of-core execution (or both). Some algorithms are more advantageous for some circumstances, and thus the ability to jump between them during recursive subproblems becomes key to enabling high performance. But being able to select a given compound (overall) algorithm requires the set of compound algorithms to exist (at runtime) in the first place. It comes as no surprise that coding, maintaining, and storing the various compound algorithms for each operation can become a chore.

BLIS solves this problem by borrowing a relatively novel construct unique to `libflame` which we call *control trees*. Control trees are based on the following fundamental observation: *If we supply a basic (but "complete") set of algorithmic variants for a given operation, as well as basic sets of variants for all operations on which it might depend, then one can specify the execution path of an arbitrarily-deep compound algorithm by using a tree.* This tree can be represented with a recursive data structure. Each node in this structure is typed according to the operation it represents, and encodes information (such as block size, algorithmic variant, and packing format) that parameterizes the implementation to be used. Interior nodes in the tree correspond to blocked algorithms that further partition the problem into smaller subproblems while leaf nodes correspond to unblocked codes. Blocked algorithms, instead of calling a hard-coded implementation directly, call the operation's internal "back-end" (passing along the appropriate sub-tree). This back-end function acts as a decoder that inspects the contents of the current control tree node and dispatches the prescribed blocked or unblocked implementation.[19] And since control trees are stateless objects, they are thread-safe. The tree nodes themselves can even be used to encode information about when to partition the problem for parallelism [9].

The key consequence of the control tree infrastructure is that it allows one to eliminate virtually *all* redundant algorithm codes from the framework. This is possible because the trees, which may be specified statically prior to compilation or built dynamically at run-time, allow the algorithm developer to compose entire algorithms from the basic building blocks of its subproblems. Another consequence of control trees is that once the basic variants are developed and coded, one can change the overall algorithm, block size (at any level), or packing format, by simply changing the control tree. Thus, in BLIS, blocked algorithmic variants capture only the essence of the *algorithm*, while the *implementation* of subproblems are determined by the tree and not realized until it is decoded at run-time by the back-end. If new blocked variants (or macro-kernels) are added to the framework, their support is added by making the internal back-end code aware of these new routines. They are then available for use by simply specifying their usage via a control tree.

And so, by expressing the execution path (and key runtime parameters) of the overall level-3 algorithm with a tree structure, we allow arbitrary algorithms to be composed from a small set of variants, while simultaneously keeping the source code and object code footprints quite manageable.

5.2 Level-2, Level-1v, Level-1m, and Level-1f operations

We now discuss how each of the five dimensions of generality are achieved for the level-1v, level-1m, level-1f, and level-2 operations.

---

[19]It would reasonable for the reader to harbor some skepticism that control trees would not incur significant overhead during execution. Specifically, for every blocked algorithm subproblem, a control tree-based solution incurs an additional function call to the subproblem operation's internal back-end routine, whereas a conventional algorithm would encode the desired subroutine call directly into the body of the algorithm implementation. In our experience, this overhead tends to be negligible since it is amortized over a large amount of computation. The level-3 performance results in Section 6 confirm that control tree overhead can be quite minimal, if not indistinguishable.

5.2.1 Operation

Recall that with level-3 operations, we were able to greatly simplify the implementations by expressing the operations in terms of a GEMM micro-kernel, with the remaining differences expressed via either the packing facility or the operation's macro-kernel, or a combination of the two. However, level-2 (and lower) operations cannot afford to be implemented with a pack stage if performance is to be maximized. Indeed, performance is already hindered by the fact that the ratio of flops to memops for these operations is effectively $\mathcal{O}(1)$. Thus, there is not nearly as much opportunity for reuse of algorithms between level-2 operations.[20] However, there is a silver lining. As mentioned in Sections 4.3 and 4.4, all level-2 operations can be realized in optimized form by simply optimizing the level-1f kernels shown in Figure 7, and similarly, a level-1m implementation can be optimized by optimizing its corresponding level-1v operation. And so the amount of code that must be customized to achieve high performance for a given architecture, while greater than that of level-3 BLIS, remains relatively small.

It should be noted that, because BLIS exposes multiple layers, we can easily implement higher-level operations in terms of lower-level codes that forgo error checking, thereby avoiding redundant error checking. For example, level-1m BLIS operations are implemented in terms of level-1v kernels; however, these kernels are not accessed via the same APIs that an end-user would call. Instead, we use expert-level interfaces to access level-1v codes that skip error checking altogether. In this way, function parameters and other properties of the problem, such as operand dimensions, can be checked once up-front.

5.2.2 Parameter case

The number of parameter cases needed to fully support level-2 and lower operations is noticeably smaller than that of level-3 operations. For example, unlike their level-3 analogues, HEMV, TRMV, and TRSV have no `side` parameter. And while some operations still sometimes offer the option of transposing an operand or computing with only the lower- or upper-stored region of a matrix, both of these options can usually be handled quite naturally by manipulating the row and column strides of the matrix object. For example, a matrix transpose is typically induced by swapping the row and column strides (and sometimes making adjustments to dimensions, as with the case of GEMV). Furthermore, level-2 BLIS operations may be implemented to explicitly support only the lower-stored case, but still indirectly support the upper-stored case by internally inducing a matrix transposition prior to executing any underlying level-1v or -1f kernels. In the case of HEMV, HER, and HER2, switching the lower/upper storage parameter corresponds to toggling one or more conjugation parameters (since $\text{LowerTriangle}(H) = \text{UpperTriangle}(\bar{H})$, where H is Hermitian). Scalar-level conjugation is implemented directly within the underlying level-1v or level-1f kernels, which allows algorithms to be implemented without the need for temporarily conjugated (and thread-unsafe) operands. Support for arbitrary diagonal offsets for lower- and upper-stored matrices is implemented uniformly across all level-1m operations, and in such a way that factors out the logic that adjusts loop bounds and index offsets to account for the location of the diagonal.

5.2.3 Datatype

By default, the BLIS framework provides portable, C-based implementations of all level-1v, level-1m, and Level-2 operations, each with full datatype support. BLAS achieves this by providing a separate code for each datatype supported, despite these codes differing in only slight ways. BLIS avoids this redundancy by using the C preprocessor to employ macro-based templates whenever unblocked codes are written (including the macro-kernels used by level-3 operations). Thus, the framework stores only one datatype-agnostic copy of each variant; the datatype-specific codes are then instantiated by the preprocessor at compile-time. This technique drastically reduces the source code (though not the object code) footprint of the framework, which in turns makes the framework much easier to modify and maintain.

---

[20]Nevertheless, careful coding allows us to consolidate some operation implementations. For example, we implement real SYR and complex HER as the same algorithm. This is possible since the Hermitian transpose in the case of complex HER reduces to a transpose in the real domain. Furthermore, a simple switch that disables the conjugation component of the Hermitian transpose allows us implement complex SYR using the same algorithm codes. A similar consolidation may be performed for SYR2/HER2 and SYMV/HEMV.

These macro-based templates also allow one to capture mixed datatype algorithms by specifying the types of each operand separately. If and when the vendor configures BLIS to allow mixed domain and/or precision, these additional datatype combinations are instantiated automatically when the library is built. However, at first all implementations, whether of homogeneous or mixed datatypes, will be unoptimized. If the vendor chooses to support only operands of homogeneous datatypes, then he or she need only optimize the level-1v and/or level-1f kernels for homogeneous datatypes (i.e., four datatype instances of each operation, assuming all four floating-point datatypes are to be supported). If the vendor wishes to support datatype mixing, then he or she will need to provide additional level-1v and/or level-1f kernels that take operands of mixed types.[21]

5.2.4 Storage format

Generalized storage is supported for level-1m and -2 operations just as with level-3 operations. The mechanism, however, is different. Here, we rely upon underlying kernels to support general stride. Recall that the level-1m and -2 operations are implemented in terms of level-1v and -1f kernels; so, as long as the latter are implemented to support both contiguous (row- or column-stored) and general stride storage, the former will exhibit this support as well. Admittedly, *supporting* contiguous as well as general stride cases is distinct from *optimizing* those cases, as kernel developers often allow the assumption of contiguous storage to influence their kernel implementation (specifically, their choice of assembly instructions). Thus, the vendor will have to weigh whether including optimizations for both cases is important for his or her users.

5.2.5 Algorithm

The set of variants for level-2 and lower operations is much smaller due to the reduced number of possible matrix and vector partitionings. For a typical level-1m operation such as AXPYM, BLIS needs to provide only one variant, because the default runtime system can manipulate matrix objects (e.g. strides) to allow unit stride through memory whenever possible. For a typical level-2 operation, two or more variants are provided. This allows the runtime system to choose the best variant based on the exact nature of the problem, taking in account matrix storage as well as transposition. For example, GEMV invokes a DOTXV-based implementation if matrix A is row-stored and not in need of a transpose, or if A is column-stored and in need of a transpose, while an AXPYV-based variant is preferred for the two remaining cases. For general stride cases, unit stride cannot be achieved, and thus by default the BLIS framework chooses a DOTXV-based variant, if one is available, since it incurs fewer memory operations while performing the same number of flops.

5.3 Summary remarks

We have now discussed a wide breadth (though we withhold depth for space reasons) of implementation techniques that we employ in BLIS which help us keep the framework as small, manageable, and flexible as possible. It should be noted that *some* increase in code is sustained even when employing the methods described above. But our primary goal here was simply to avoid the situation whereby increasing functionality along one dimension would *proportionally* inflate the resulting library's code size. In our humble opinion, the only dimension along which we clearly fail at this goal is that of datatype instantiation for level-1v and level-1f kernels.

The BLIS framework employs other more mundane techniques that contribute generally to its usability and maintainability, which we choose not to document here for, again, space reasons.

Figure 13: Performance of various implementations of double-precision real GEMM (top-left), SYMM (top-right), SYRK (center-left), SYR2K (center-right) TRMM (bottom-left), and TRSM (bottom-right), on a single core of a 2.66GHz Intel Xeon 7400 system.
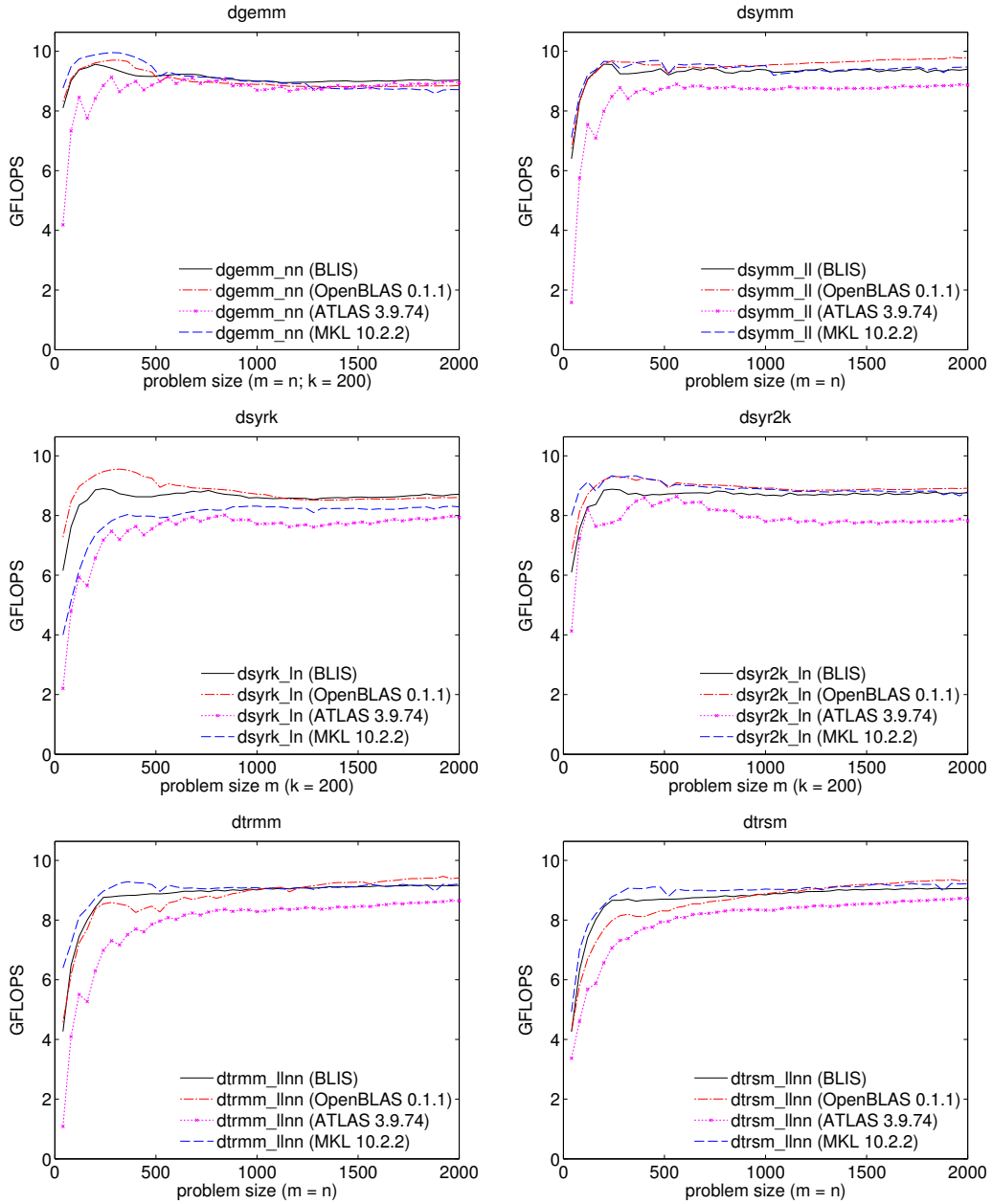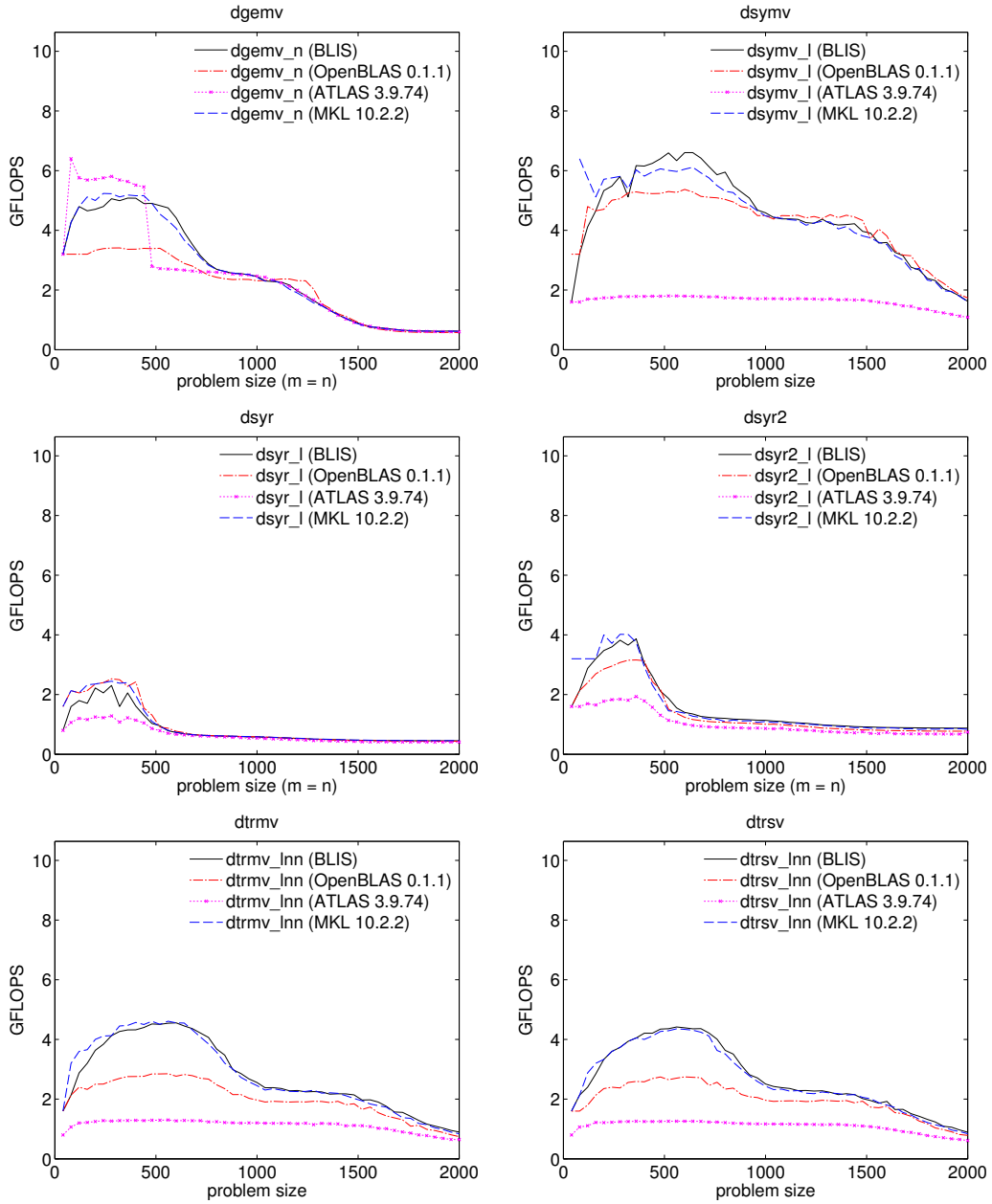
Figure 14: Performance of various implementations of double-precision real GEMV (top-left), SYMV (top-right), SYR (center-left), SYR2 (center-right) TRMV (bottom-left), and TRSV (bottom-right), on a single core of a 2.66GHz Intel Xeon 7400 system.

6 Performance

In this section we present performance of various level-2 and level-3 BLIS operations. Comparisons are made with corresponding BLAS implementations of various libraries.

6.1 Platform and implementation details

All experiments reported in this paper were performed on a single core of a Dell PowerEdge R900 server consisting of four six-core Intel Xeon 7400 processors. Each core provides a peak performance of 10.64 GFLOPS. Performance experiments were gathered under the GNU/Linux 2.6.18 operating system. Source code was compiled by the GNU C compiler, version 4.1.2. All experiments were performed in double-precision floating-point arithmetic on randomized real domain matrices stored in column-major order.

We compare BLIS performance to three other single-threaded implementations: OpenBLAS 0.1.1, ATLAS 3.9.74 and MKL 10.2.2. The BLIS implementations are based on an early development version of the BLIS framework combined with a hand-coded micro-kernel similar to the corresponding inner-most loop of one of the kernels found within OpenBLAS. This micro-kernel was expressed in GNU extended inline assembly syntax and coded to leverage vector (SSE2 and later) instructions. Level-1 and Level-1f BLIS kernels were coded at a somewhat higher level, using SSE vector intrinsics.

In the following subsections, we discuss the performance results of Figures 13 and Figure 14. These graphs show performance, measured in GFLOPS (10^9 floating-point operations per second), as a function of matrix operand (problem) size, where each data point shows the best of three trials. For operations with matrix operands that have both m and n dimensions, we fix $m = n$, and for those that have a k dimension, we fix $k = 200$. All graphs are scaled such that the maximum y-axis value is equal to the peak performance of a single core, 10.64 GFLOPS.

The underlying micro-kernel utilized by level-3 operations reported in Figure 13 used register block sizes $m_r = n_r = 4$, while the higher-level blocked algorithms used cache block sizes $m_c = 368$ and $k_c = 256$. Similarly, the level-1f kernels AXPYF, DOTXF, and DOTXAXPYF used by level-2 operations reported in Figure 14 were implemented with fusing factors of $f = 4$.

6.2 General results

Figure 13 shows results for the double-precision real instances of six level-3 operations: GEMM, SYMM, SYRK, SYR2K, TRMM, and TRSM. In all six graphs, the performance of the BLIS implementation can be seen to be highly competitive with existing implementations, including those provided by the Intel MKL library.

Similarly, Figure 14 shows performance for the double-precision real instances of six level-2 operations: GEMV, SYMV, SYR, SYR2, TRMV, and TRSV. (These are the level-2 analogues of the operations reported on in Figure 13.) Once again, observed performance of the implementations instantiated via the BLIS framework is comparable to that of codes available via OpenBLAS, ATLAS, and MKL.

These results confirm that the BLIS framework is capable of facilitating competitive performance for BLAS-like operations. Importantly, with the exception of TRSM, the BLIS performance results shown in Figure 13 were achieved by optimizing *only* the micro-kernel illustrated in Figure 12; all other aspects of the implementations employed portable component codes from the framework. Similarly, the optimizations that enabled the BLIS performance results in Figure 14 were limited to those of the level-1 and level-1f kernels shown in right-hand column of Figure 10.

6.3 Benefit of fused GEMM-TRSM micro-kernel

As mentioned in Footnote 12, near-optimal TRSM performance requires more than just an optimized GEMM micro-kernel. TRSM is inherently dissimilar from the other level-3 operations in that it cannot be entirely

---

[21]Unfortunately, this aspect of the framework—whereby the vendor must provide one kernel for every datatype combination desired—is probably the weakest part of the framework. But, absent the vendor (and his or her users) tolerating a potentially significant cost in terms of workspace and memory copy overhead incurred by a runtime typecast/packing facility, we see no other way of providing this functionality in optimized form.
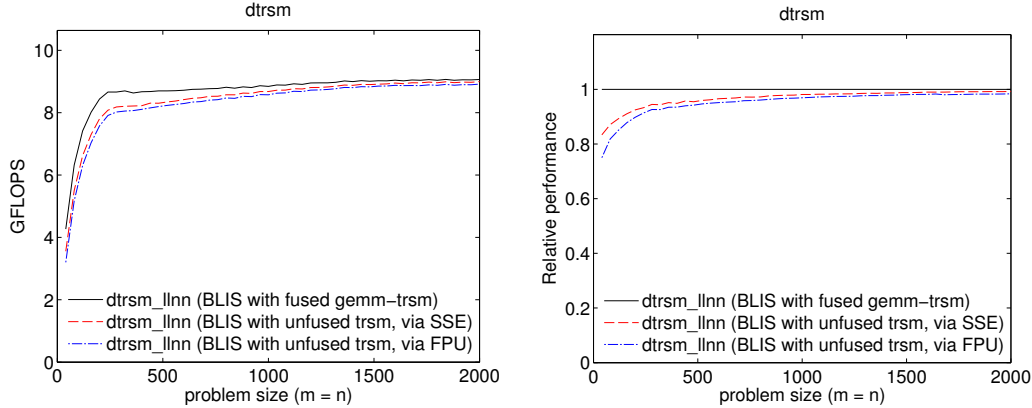
Figure 15: Left: Performance of BLIS TRSM with three levels of optimization: fully optimized via a fused GEMM-TRSM micro-kernel (as shown in the bottom-right graph of Figure 13); partially optimized via an unfused TRSM compiled with scalar SSE instructions (which are automatically emitted by the compiler via `-mfpmath=sse`); and unoptimized via an unfused TRSM compiled with traditional FPU-based instructions. Right: Performance of the latter two implementations relative to that of the former.

expressed in terms of a simple matrix-matrix multiplication. While the bulk of its computation can be cast in terms of GEMM, it also requires a small triangular solve with multiple right-hand sides (TRSM). While this TRSM subproblem constitutes a very small fraction of the total flops executed, it incurs extra memory operations that could otherwise be avoided altogether if the GEMM and TRSM subproblems were fused at the register level.

The BLIS implementation of TRSM reported on in Figure 13 uses a fused GEMM-TRSM micro-kernel, and thus is more fully-optimized than an implementation that leverages only the GEMM micro-kernel. How much TRSM performance is gained by crafting this specialized micro-kernel? To answer this question, we tested two additional implementations in addition to the one shown in Figure 13. These three implementations are described as follows:

- **BLIS with fused GEMM-TRSM.** This implementation fuses the GEMM micro-kernel with the small TRSM subproblem that follows it so that the current $m_r \times n_r$ submatrix of B is held in registers rather than being written out to memory and immediately read back in. This fused code is written with GNU extended inline assembly syntax, which utilizes vector SSE instructions.

- **BLIS with unfused TRSM via SSE.** This implementation of TRSM separates the GEMM micro-kernel call from the smallest TRSM subproblem. This small TRSM subproblem is coded portably in C and then compiled with the `-mfpmath=sse` compiler option, which causes the compiler to emit scalar SSE instructions.

- **BLIS with unfused TRSM via FPU.** This is similar to the above implementation, except the code is instead compiled with the `-mfpmath=387` compiler option, which causes the compiler to emit instructions that utilize only the standard floating-point unit present on Intel x86 processors.

We would expect the SSE-based unfused implementation to outperform the FPU-based code, and for both to underperform the code which employs a fused GEMM-TRSM micro-kernel. Figure 15 reports performance consistent with this prediction. Perhaps the key take-away from this graph is that while the fused implementation offers somewhat higher performance, the performance penalty incurred by forgoing this level of optimization is relatively small, especially when using compiler-generated scalar SSE instructions and/or when computing upon larger matrices.

6.4 Potential for automation

Auto-tuning of level-3 BLAS started with IBM's effort to provide BLAS for the Power2 architecture [1]. That effort showed that high-performance implementations could be achieved in a high-level language (Fortran). The PHiPAC project [6] picked up on this, but introduced the idea of generating highly efficient C code, including automatic tuning of block sizes and loop orderings. This was then pushed to the next level by the ATLAS project, which reduced the search space by focusing on automatically tuning inner kernels and block sizes. Inner kernels are now actually hand-written in assembly code and catalogued within the system so that the appropriate kernel can be chosen later, when the library is configured and built.

The reader may notice that we do not discuss auto-tuning. Yotov et al. showed that architecture parameters and characteristics can be used to analytically derive the tuning parameters that ATLAS determines through exhaustive search [39]. Further evidence that auto-tuning is not necessary comes from the fact that Goto did not auto-tune (and achieves better performance than ATLAS) for all architectures he targeted [12, 13], nor did Volkov auto-tune when optimizing for GPUs [37]. We believe that the BLIS framework structures the implementation effort in a way that exposes fundamental techniques and hence allows architectural parameters and insight to be used in lieu of auto-tuning. Nonetheless, one could at some point add auto-tuning to the framework.

A related topic of research is that of automatically generating linear algebra libraries from domain and hardware specifications. We believe that the BLIS framework isolates architecture specifics to the point where a tool like Spiral [28] could be used to generate micro-kernels. Indeed, one of the motivations for creating the BLIS framework was to distill the dense linear algebra software stack into basic facilities (i.e., computational and packing kernels) so that Design by Transformation [23] could encode expert knowledge about the library operations and then generate highly optimized higher-level functionality. We intend to investigate this further in future work. Other groups, such as the Build to Order BLAS project [30, 4], may also benefit from the kernel and/or layering that BLIS exposes.

7 Conclusions

In this paper, we have proposed a new framework for building a modern alternative to the BLAS. This framework addresses several shortcomings of the original BLAS interface design, expands functionality, and most importantly, isolates most architecture-sensitive details to a few relatively small kernels, which, when optimized, immediately enable high-performance across virtually all supported operations. Experimental level-2 and level-3 performance was observed to be competitive with two highly-regarded open source libraries, as well as a famously optimized commercial solution from Intel.

This work lays the foundation for future efforts:

- One of our highest priorites is is to add support for multithreading (via either POSIX threads or OpenMP). We envision a multithreaded BLIS framework that exports *simultaneous* access to sequential and parallel implementations, where the number of threads for the latter can be changed by the user at runtime. This would allow applications to control precisely how much parallelism is achieved within BLIS at any given time.

- While early versions of the BLIS framework have been ported to a handful of architectures, we wish to gain exposure to even more types of hardware, including GPUs. These experiences will help us further refine and generalize the framework, and streamline the process of instantiating BLAS-like libraries on other similarly exotic systems.

- As mentioned previously, given that BLIS isolates performance-sensitive code to a few simple kernels, the framework may aid those who wish to automate the generation of high-performance linear algebra libraries from domain and hardware specifications.

- We may also look toward adding support for extended precision arithmetic.

Thus, the contributions of this work provide several avenues for further research.

Acknowledgements

References

[1] R.C. Agarwal, F.G. Gustavson, and M. Zubair. Exploiting functional parallelism of POWER2 to design high-performance numerical algorithms. *IBM Journal of Research and Development*, 38(5), Sept. 1994.

[2] AMD. AMD Core Math Library. `http://developer.amd.com/tools/cpu/acml/pages/default.aspx`, 2012.

[3] E. Anderson, Z. Bai, J. Demmel, J. E. Dongarra, J. DuCroz, A. Greenbaum, S. Hammarling, A. E. McKenney, S. Ostrouchov, and D. Sorensen. *LAPACK Users' Guide*. SIAM, Philadelphia, 1992.

[4] Geoffrey Belter, E. R. Jessup, Ian Karlin, and Jeremy G. Siek. Automating the generation of composed linear algebra kernels. In *Proceedings of the Conference on High Performance Computing Networking, Storage and Analysis*, SC '09, pages 59:1–59:12, 2009.

[5] Paolo Bientinesi, John A. Gunnels, Margaret E. Myers, Enrique S. Quintana-Ortí, and Robert A. van de Geijn. The science of deriving dense linear algebra algorithms. *ACM Trans. Math. Soft.*, 31(1):1–26, March 2005.

[6] Jeff Bilmes, Krste Asanović, Chee whye Chin, and Jim Demmel. Optimizing matrix multiply using PHiPAC: a Portable, High-Performance, ANSI C coding methodology. In *Proceedings of International Conference on Supercomputing*, Vienna, Austria, July 1997.

[7] `http://www.netlib.org/blas/`, 2012.

[8] Basic linear algebra subprograms technical forum standard. *International Journal of High Performance Applications and Supercomputing*, 16(1), Spring 2002.

[9] Ernie Chan, Field G. Van Zee, Paolo Bientinesi, Enrique S. Quintana-Ortí, Gregorio Quintana-Ortí, and Robert van de Geijn. SuperMatrix: A multithreaded runtime scheduling system for algorithms-by-blocks. In *ACM SIGPLAN 2008 symposium on Principles and practices of parallel programming (PPoPP'08)*, pages 123–132, 2008.

[10] Jack J. Dongarra, Jeremy Du Croz, Sven Hammarling, and Iain Duff. A set of level 3 basic linear algebra subprograms. *ACM Trans. Math. Soft.*, 16(1):1–17, March 1990.

[11] Jack J. Dongarra, Jeremy Du Croz, Sven Hammarling, and Richard J. Hanson. An extended set of FORTRAN basic linear algebra subprograms. *ACM Trans. Math. Soft.*, 14(1):1–17, March 1988.

[12] Kazushige Goto and Robert van de Geijn. Anatomy of high-performance matrix multiplication. *ACM Trans. Math. Soft.*, 34(3):12:1–12:25, May 2008.

[13] Kazushige Goto and Robert van de Geijn. High-performance implementation of the level-3 BLAS. *ACM Trans. Math. Soft.*, 35(1):1–14, July 2008.

[14] John A. Gunnels, Fred G. Gustavson, Greg M. Henry, and Robert A. van de Geijn. FLAME: Formal linear algebra methods environment. *ACM Trans. Math. Soft.*, 27(4):422–455, December 2001.

[15] John A. Gunnels, Greg M. Henry, and Robert A. van de Geijn. A family of high-performance matrix multiplication algorithms. In Vassil N. Alexandrov, Jack J. Dongarra, Benjoe A. Juliano, René S. Renner, and C.J. Kenneth Tan, editors, *Computational Science - ICCS 2001, Part I*, Lecture Notes in Computer Science 2073, pages 51–60. Springer-Verlag, 2001.

[16] John A. Gunnels and Robert A. van de Geijn. Formal methods for high-performance linear algebra libraries. In Ronald F. Boisvert and Ping Tak Peter Tang, editors, *The Architecture of Scientific Software*, pages 193–210. Kluwer Academic Press, 2001.

[17] Gary W. Howell, James W. Demmel, Charles T. Fulton, Sven Hammarling, and Karen Marmol. Cache efficient bidiagonalization using BLAS 2.5 operators. *ACM Transactions on Mathematical Software*, 34(3):14:1–14:33, May 2008.

[18] IBM. Engineering and Scientific Subroutine Library. `http://www.ibm.com/systems/software/essl/`, 2012.

[19] Intel. Math Kernel Library. `http://developer.intel.com/software/products/mkl/`, 2012.

[20] Thierry Joffrain, Tze Meng Low, Enrique S. Quintana-Ortí, Robert van de Geijn, and Field Van Zee. Accumulating Householder transformations, revisited. *ACM Transactions on Mathematical Software*, 32(2):169–179, June 2006.

[21] B. Kågström, P. Ling, and C. Van Loan. GEMM-based level 3 BLAS: High performance model implementations and performance evaluation benchmark. *ACM Trans. Math. Soft.*, 24(3):268–302, 1998.

[22] C. L. Lawson, R. J. Hanson, D. R. Kincaid, and F. T. Krogh. Basic linear algebra subprograms for Fortran usage. *ACM Trans. Math. Soft.*, 5(3):308–323, Sept. 1979.

[23] Bryan Marker, Jack Poulson, Don Batory, and Robert van de Geijn. Designing linear algebra algorithms by transformation: Mechanizing the expert developer. In *International Workshop on Automatic Performance Tuning (iWAPT2012). Proceedings of VECPAR 2012 Conference*, July 2012.

[24] C. Moler, J. Little, and S. Bangert. *Pro-Matlab, User's Guide*. The Mathworks, Inc., 1987.

[25] `http://xianyi.github.com/OpenBLAS/`, 2012.

[26] Ardavan Pedram, Andreas Gerstlauer, and Robert A. van de Geijn. On the efficiency of register file versus broadcast interconnect for collective communications in data-parallel hardware accelerators. *Computer Architecture and High Performance Computing (SBAC-PAD), 2012 24th International Symposium on*, 2012. submitted.

[27] Ardavan Pedram, Robert A. van de Geijn, and Andreas Gerstlauer. Codesign tradeoffs for high-performance, low-power linear algebra architectures. *IEEE Transactions on Computers*, 61:1724–1736, 2012.

[28] Markus Püschel, José M. F. Moura, Jeremy Johnson, David Padua, Manuela Veloso, Bryan Singer, Jianxin Xiong, Franz Franchetti, Aca Gacic, Yevgen Voronenko, Kang Chen, Robert W. Johnson, and Nicholas Rizzolo. SPIRAL: Code generation for DSP transforms. *Proceedings of the IEEE, special issue on "Program Generation, Optimization, and Adaptation"*, 93(2):232–275, 2005.

[29] Gregorio Quintana-Ortí, Enrique S. Quintana-Ortí, Robert A. van de Geijn, Field G. Van Zee, and Ernie Chan. Programming matrix algorithms-by-blocks for thread-level parallelism. *ACM Trans. Math. Softw.*, 36(3):14:1–14:26, July 2009.

[30] Jeremy G. Siek, Ian Karlin, and E. R. Jessup. Build to order linear algebra kernels. In *International Symposium on Parallel and Distributed Processing 2008 (IPDPS 2008)*, pages 1–8, 2008.

[31] Edgar Solomonik, Jeff Hammond, and James Demmel. A preliminary analysis of Cyclops Tensor Framework. Technical Report UCB/EECS-2012-29, EECS Department, University of California, Berkeley, Mar 2012.

[32] Robert A. van de Geijn and Enrique S. Quintana-Ortí. *The Science of Programming Matrix Computations*. `www.lulu.com/contents/contents/1911788/`, 2008.

[33] Field G. Van Zee. `libflame`*: The Complete Reference*. `www.lulu.com`, 2012.

[34] Field G. Van Zee, Ernie Chan, Robert van de Geijn, Enrique S. Quintana-Ortí, and Gregorio Quintana-Ortí. The libflame library for dense matrix computations. *IEEE Computation in Science & Engineering*, 11(6):56–62, 2009.

[35] Field G. Van Zee, Robert A. van de Geijn, and Gregorio Quintana-Ortí. Restructuring the tridiagonal and bidiagonal QR algorithms for performance. *ACM Trans. Math. Soft.*, 2012. submitted.

[36] Field G. Van Zee, Robert A. van de Geijn, Gregorio Quintana-Ortí, and G. Joseph Elizondo. Families of algorithms for reducing a matrix to condensed form. *ACM Trans. Math. Soft.*, 39(1):xxx–yyy, 2012.

[37] Vasily Volkov and James Demmel. LU, QR and Cholesky factorizations using vector capabilities of GPUs. Technical Report UCB/EECS-2008-49, EECS Department, University of California, Berkeley, May 2008.

[38] R. Clint Whaley and Jack J. Dongarra. Automatically tuned linear algebra software. In *Proceedings of SC'98*, 1998.

[39] Kamen Yotov, Xiaoming Li, María Jesús Garzarán, David Padua, Keshav Pingali, and Paul Stodghill. Is search really necessary to generate high-performance BLAS? volume 93, 2005.

Level-1v operation	BLAS-like BLIS interface

$y := y + \alpha\mathrm{CJ}(x)$
 	
where $\dim(x) = \dim(y)$
	`void bli_?axpyv( conj_t conjx, dim_t m,`
	`                 void* alpha, void* x, inc_t incx,`
	`                                void* y, inc_t incy );`

$y := \mathrm{CJ}(x)$ where $\dim(x) = \dim(y)$	`void bli_?copyv( conj_t conjx, dim_t m,` `                 void* x, inc_t incx, void* y, inc_t incy );`

$\rho := \mathrm{CJ}(x)^T\mathrm{CJ}(y)$
 	
where $\dim(x) = \dim(y)$
	`void bli_?dotv( conj_t conjx, conj_t conjy, dim_t m,`
	`                void* x, inc_t incx,`
	`                void* y, inc_t incy, void* rho );`

$\rho := \beta\rho + \alpha\mathrm{CJ}(x)^T\mathrm{CJ}(y)$
 	
where $\dim(x) = \dim(y)$
	`void bli_?dotxv( conj_t conjx, conj_t conjy, dim_t m,`
	`                 void* alpha, void* x, inc_t incx,`
	`                                void* y, inc_t incy,`
	`                 void* beta,  void* rho );`

$\psi_i := \psi_i^{-1}, \quad \forall\psi_i \in y$	`void bli_?invertv( dim_t m, void* y, inc_t incy );`
$\psi_i := \alpha, \quad \forall\psi_i \in y$	`void bli_?setv( dim_t m, void* alpha, void* y, inc_t incy );`
$y := \alpha y$	`void bli_?scalv( dim_t m, void* alpha, void* y, inc_t incy );`

$y := \alpha\mathrm{CJ}(x)$
 	
where $\dim(x) = \dim(y)$
	`void bli_?scal2v( conj_t conjx, dim_t m,`
	`                  void* alpha, void* x, inc_t incx,`
	`                                 void* y, inc_t incy );`

Figure 16: A table of BLAS-like interfaces for Level-1v BLIS operations. Function names contain a ? to indicate a type specification character (i.e., s, d, c, z), or string (for accessing mixed-datatype functionality). The function CJ() denotes optional conjugation.

Level-1m operation	BLAS-like BLIS interface

$B := B + \alpha\mathrm{CT}(A)$
 	
where B and $\mathrm{CT}(A)$ are
$m \times n$
	`void bli_?axpym( doff_t diagoffa, diag_t diaga, uplo_t uploa,`
	`                 trans_t transa, dim_t m, dim_t n,`
	`                 void* alpha, void* a, inc_t rsa, inc_t csa,`
	`                                void* b, inc_t rsb, inc_t csb );`

$B := \mathrm{CT}(A)$
 	
where B and $\mathrm{CT}(A)$ are
$m \times n$
	`void bli_?copym( doff_t diagoffa, diag_t diaga, uplo_t uploa,`
	`                 trans_t transa, dim_t m, dim_t n,`
	`                              void* a, inc_t rsa, inc_t csa,`
	`                              void* b, inc_t rsb, inc_t csb );`

$\beta_{ij} := \alpha, \quad \forall\beta_{ij} \in B$	`void bli_?setm( doff_t diagoffb, diag_t diagb, uplo_t uplob,`
	`                dim_t m, dim_t n, void* alpha,`
	`                void* b, inc_t rsb, inc_t csb );`

$B := \alpha B$	`void bli_?scalm( doff_t diagoffb, diag_t diagb, uplo_t uplob,`
	`                 dim_t m, dim_t n,`
	`                 void* alpha, void* b, inc_t rsb, inc_t csb );`

$B := \alpha\mathrm{CT}(A)$
 	
where B and $\mathrm{CT}(A)$ are
$m \times n$
	`void bli_?scal2m( doff_t diagoffa, diag_t diaga, uplo_t uploa,`
	`                  trans_t transa, dim_t m, dim_t n,`
	`                  void* alpha, void* a, inc_t rsa, inc_t csa,`
	`                                 void* b, inc_t rsb, inc_t csb );`

Figure 17: A table of BLAS-like interfaces for Level-1m BLIS operations. Function names contain a ? to indicate a type specification character (i.e., s, d, c, z), or string (for accessing mixed-datatype functionality). The functions CJ() and CT() denote optional conjugation and optional conjugation and/or transposition, respectively.

A   Selected BLIS interfaces

Level-1f operation	BLAS-like BLIS interface

$\begin{cases} y := y + \alpha_0 \mathrm{CJ}(x_0) \\ y := y + \alpha_1 \mathrm{CJ}(x_1) \end{cases}$
 where $\dim(y) = \dim(x_0)$
 $= \dim(x_1) = m$
	```void bli_?axpy2v(conj_t conjx0,```
	``` conj_t conjx1,```
	``` dim_t m,```
	``` void* alpha0, void* alpha1,```
	``` void* x0, inc_t incx0,```
	``` void* x1, inc_t incx1,```
	``` void* y, inc_t incy);```

$\begin{cases} \rho := \mathrm{CJ}(x)^T \mathrm{CJ}(y) \\ z := z + \alpha \mathrm{CJ}(x) \end{cases}$
 where $\dim(x) = \dim(y)$
 $= \dim(z) = m$
	```void bli_?dotaxpyv(conj_t conjxt,```
	``` conj_t conjx,```
	``` conj_t conjy,```
	``` dim_t m,```
	``` void* alpha,```
	``` void* x, inc_t incx,```
	``` void* y, inc_t incy,```
	``` void* rho,```
	``` void* z, inc_t incz);```

$\begin{cases} y := y + \alpha \mathrm{CJ}(\chi_0)\mathrm{CJ}(a_0) \\ \quad\vdots \\ y := y + \alpha \mathrm{CJ}(\chi_{f-1})\mathrm{CJ}(a_{f-1}) \end{cases}$
 where $\dim(y) =$
 $\dim(a_0) = \cdots = \dim(a_{f-1}) = m$
	```void bli_?axpyf(conj_t conja,```
	``` conj_t conjx,```
	``` dim_t m, dim_t f,```
	``` void* alpha,```
	``` void* a, inc_t inca, inc_t lda,```
	``` void* x, inc_t incx,```
	``` void* y, inc_t incy);```

$\begin{cases} \psi_0 := \beta\psi_0 + \alpha \mathrm{CJ}(a_0)^T\mathrm{CJ}(x) \\ \quad\vdots \\ \psi_{f-1} := \beta\psi_{f-1} + \alpha \mathrm{CJ}(a_{f-1})^T\mathrm{CJ}(x) \end{cases}$
 where $\dim(x) =$
 $\dim(a_0) = \cdots = \dim(a_{f-1}) = n$
	```void bli_?dotxf(conj_t conja,```
	``` conj_t conjx,```
	``` dim_t f, dim_t n,```
	``` void* alpha,```
	``` void* a, inc_t inca, inc_t lda,```
	``` void* x, inc_t incx,```
	``` void* beta,```
	``` void* y, inc_t incy);```

$\begin{cases} y_0 := \beta y_0 + \alpha \mathrm{CJ}(A^T)\mathrm{CJ}(x_0) \\ y_1 := y_1 + \alpha \mathrm{CJ}(A)\mathrm{CJ}(x_1) \end{cases}$
 where $\dim(x_0) = \dim(y_1) = m$,
 $\dim(x_1) = \dim(y_0) = f$, and
 A is $m \times f$.
	```void bli_?dotxaxpyf(```
	``` conj_t conjat, conj_t conja,```
	``` conj_t conjx0, conj_t conjx1,```
	``` dim_t m, dim_t f,```
	``` void* alpha,```
	``` void* a, inc_t inca, inc_t lda,```
	``` void* x0, inc_t incx0,```
	``` void* x1, inc_t incx1,```
	``` void* beta,```
	``` void* y0, inc_t incy0,```
	``` void* y1, inc_t incy1);```

Figure 18: A table of BLAS-like interfaces for Level-1f BLIS operations. Function names contain a ? to indicate a type specification character (i.e., s, d, c, z), or string (for accessing mixed-datatype functionality). The function CJ() denotes optional conjugation.

Level-2 operation	BLAS-like BLIS interface

$y := \beta y + \alpha \mathrm{CT}(A)\mathrm{CJ}(x)$
where $\mathrm{CT}(A)$ is $m \times n$
	```void bli_?gemv(trans_t transa, conj_t conjx,```
	`              dim_t m, dim_t n,`
	`              void* alpha, void* a, inc_t rsa, inc_t csa,`
	`                           void* x, inc_t incx,`
	`              void* beta,  void* y, inc_t incy );`

$A := A + \alpha \mathrm{CJ}(x)\mathrm{CJ}(y)^T$
where A is $m \times n$
	```void bli_?ger(conj_t conjx, conj_t conjy,```
	`             dim_t m, dim_t n,`
	`             void* alpha, void* x, inc_t incx,`
	`                          void* y, inc_t incy,`
	`                          void* a, inc_t rsa, inc_t csa );`

$y := \beta y + \alpha \mathrm{CJ}(A)\mathrm{CJ}(x)$
where A is $m \times m$ and Hermitian
	```void bli_?hemv(uplo_t uploa, conj_t conja, conj_t conjx,```
	`              dim_t m,`
	`              void* alpha, void* a, inc_t rsa, inc_t csa,`
	`                           void* x, inc_t incx,`
	`              void* beta,  void* y, inc_t incy );`

$A := A + \alpha \mathrm{CJ}_x(x)\mathrm{CJ}_x(x)^H$
where A is $m \times m$ and Hermitian
	```void bli_?her(uplo_t uploa, conj_t conjx,```
	`             dim_t m,`
	`             void* alpha, void* x, inc_t incx,`
	`                          void* a, inc_t rsa, inc_t csa );`

$A := A + \alpha \mathrm{CJ}_x(x)\mathrm{CJ}_y(y)^H$
$\quad + \bar{\alpha}\mathrm{CJ}_y(y)\mathrm{CJ}_x(x)^H$
where C is $m \times m$ and Hermitian
	```void bli_?her2(uplo_t uploa, conj_t conjx, conj_t conjy,```
	`              dim_t m,`
	`              void* alpha, void* x, inc_t incx,`
	`                           void* y, inc_t incy,`
	`                           void* a, inc_t rsa, inc_t csa );`

$y := \beta y + \alpha \mathrm{CJ}(A)\mathrm{CJ}(x)$
where A is $m \times m$ and symmetric
	```void bli_?symv(uplo_t uploa, conj_t conja, conj_t conjx,```
	`              dim_t m,`
	`              void* alpha, void* a, inc_t rsa, inc_t csa,`
	`                           void* x, inc_t incx,`
	`              void* beta,  void* y, inc_t incy );`

$A := A + \alpha \mathrm{CJ}_x(x)\mathrm{CJ}_x(x)^T$
where A is $m \times m$ and symmetric
	```void bli_?syr(uplo_t uploa, conj_t conjx,```
	`             dim_t m,`
	`             void* alpha, void* x, inc_t incx,`
	`                          void* a, inc_t rsa, inc_t csa );`

$A := A + \alpha \mathrm{CJ}_x(x)\mathrm{CJ}_y(y)^T$
$\quad + \alpha\mathrm{CJ}_y(y)\mathrm{CJ}_x(x)^T$
where C is $m \times m$ and symmetric
	```void bli_?syr2(uplo_t uploa, conj_t conjx, conj_t conjy,```
	`              dim_t m,`
	`              void* alpha, void* x, inc_t incx,`
	`                           void* y, inc_t incy,`
	`                           void* a, inc_t rsa, inc_t csa );`

$x := \alpha \mathrm{CT}(A)x$
where A is $m \times m$ and triangular
	```void bli_?trmv(uplo_t uploa, trans_t transa, diag_t diaga,```
	`              dim_t m,`
	`              void* alpha, void* a, inc_t rsa, inc_t csa,`
	`                           void* x, inc_t incx );`

$x := \alpha \mathrm{CT}(A)^{-1}x$
where A is $m \times m$ and triangular
	```void bli_?trsv(uplo_t uploa, trans_t transa, diag_t diaga,```
	`              dim_t m,`
	`              void* alpha, void* a, inc_t rsa, inc_t csa,`
	`                           void* x, inc_t incx );`

Figure 19: A table of BLAS-like interfaces for Level-2 BLIS operations. Function names contain a ? to indicate a type specification character (i.e., s, d, c, z), or string (for accessing mixed-datatype functionality). The functions CJ() and CT() denote optional conjugation and optional conjugation and/or transposition, respectively. A subscript is used in conjuction with CJ() to indicate two instances of the same conjugation and/or transposition.

Level-3 operation	BLAS-like BLIS interface
$C := \beta C + \alpha\mathrm{CT}(A)\mathrm{CT}(B)$ where C is $m \times n$	```void bli_?gemm(trans_t transa, trans_t transb,` ` dim_t m, dim_t n, dim_t k,` ` void* alpha, void* a, inc_t rsa, inc_t csa,` ` void* b, inc_t rsb, inc_t csb,` ` void* beta, void* c, inc_t rsc, inc_t csc);```
$C := \beta C + \alpha\mathrm{CJ}(A)\mathrm{CT}(B)$, or $C := \beta C + \alpha\mathrm{CT}(B)\mathrm{CJ}(A)$ where C and $\mathrm{CT}(B)$ are $m \times n$, A is Hermitian	```void bli_?hemm(side_t side, uplo_t uploa, conj_t conja,` ` trans_t transb,` ` dim_t m, dim_t n,` ` void* alpha, void* a, inc_t rsa, inc_t csa,` ` void* b, inc_t rsb, inc_t csb,` ` void* beta, void* c, inc_t rsc, inc_t csc);```
$C := \beta C + \alpha\mathrm{CT}_A(A)\mathrm{CT}_A(A)^H$ where C is $m \times m$ and Hermitian	```void bli_?herk(uplo_t uploc, trans_t transa,` ` dim_t m, dim_t k,` ` void* alpha, void* a, inc_t rsa, inc_t csa,` ` void* beta, void* c, inc_t rsc, inc_t csc);```
$C := \beta C + \alpha\mathrm{CT}_A(A)\mathrm{CT}_B(B)^H$ $\quad + \bar{\alpha}\mathrm{CT}_B(B)\mathrm{CT}_A(A)^H$ where C is $m \times m$ and Hermitian	```void bli_?her2k(uplo_t uploc, trans_t transa, trans_t transb,` ` dim_t m, dim_t k,` ` void* alpha, void* a, inc_t rsa, inc_t csa,` ` void* b, inc_t rsb, inc_t csb,` ` void* beta, void* c, inc_t rsc, inc_t csc);```
$C := \beta C + \alpha\mathrm{CJ}(A)\mathrm{CT}(B)$, or $C := \beta C + \alpha\mathrm{CT}(B)\mathrm{CJ}(A)$ where C and $\mathrm{CT}(B)$ are $m \times n$, A is symmetric	```void bli_?symm(side_t side, uplo_t uploa, conj_t conja,` ` trans_t transb,` ` dim_t m, dim_t n,` ` void* alpha, void* a, inc_t rsa, inc_t csa,` ` void* b, inc_t rsb, inc_t csb,` ` void* beta, void* c, inc_t rsc, inc_t csc);```
$C := \beta C + \alpha\mathrm{CT}_A(A)\mathrm{CT}_A(A)^T$ where C is $m \times m$ and symmetric	```void bli_?syrk(uplo_t uploc, trans_t transa,` ` dim_t m, dim_t k,` ` void* alpha, void* a, inc_t rsa, inc_t csa,` ` void* beta, void* c, inc_t rsc, inc_t csc);```
$C := \beta C + \alpha\mathrm{CT}_A(A)\mathrm{CT}_B(B)^T$ $\quad + \bar{\alpha}\mathrm{CT}_B(B)\mathrm{CT}_A(A)^T$ where C is $m \times m$ and symmetric	```void bli_?syr2k(uplo_t uploc, trans_t transa, trans_t transb,` ` dim_t m, dim_t k,` ` void* alpha, void* a, inc_t rsa, inc_t csa,` ` void* b, inc_t rsb, inc_t csb,` ` void* beta, void* c, inc_t rsc, inc_t csc);```
$B := \alpha\mathrm{CT}(A)B$, or $B := \alpha B\mathrm{CT}(A)$ where B is $m \times n$, A is triangular	```void bli_?trmm(side_t side, uplo_t uploa, trans_t transa,` ` diag_t diaga,` ` dim_t m, dim_t n,` ` void* alpha, void* a, inc_t rsa, inc_t csa,` ` void* b, inc_t rsb, inc_t csb);```
$C := \beta C + \alpha\mathrm{CT}(A)\mathrm{CT}(B)$, or $C := \beta C + \alpha\mathrm{CT}(B)\mathrm{CT}(A)$ where C and $\mathrm{CT}(B)$ are $m \times n$, A is triangular	```void bli_?trmm3(side_t side, uplo_t uploa, trans_t transa,` ` diag_t diaga, trans_t transb,` ` dim_t m, dim_t n,` ` void* alpha, void* a, inc_t rsa, inc_t csa,` ` void* b, inc_t rsb, inc_t csb,` ` void* beta, void* c, inc_t rsc, inc_t csc);```
$B := \alpha\mathrm{CT}(A)^{-1}B$, or $B := \alpha B\mathrm{CT}(A)^{-1}$ where B is $m \times n$, A is triangular	```void bli_?trsm(side_t side, uplo_t uploa, trans_t transa,` ` diag_t diaga,` ` dim_t m, dim_t n,` ` void* alpha, void* a, inc_t rsa, inc_t csa,` ` void* b, inc_t rsb, inc_t csb);```

Figure 20: A table of BLAS-like interfaces for Level-3 BLIS operations. Function names contain a ? to indicate a type specification character (i.e., s, d, c, z), or string (for accessing mixed-datatype functionality). The functions CJ() and CT() denote optional conjugation and optional conjugation and/or transposition, respectively. A subscript is used in conjuction with CT() to indicate two instances of the same conjugation and/or transposition.