Carnegie Mellon

Electrical & Computer ENGINEERING

Packing - the next BLIS Frontier?

Tze Meng Low

BLIS Retreat 2017

Carnegie Mellon

Electrical & Computer ENGINEERING

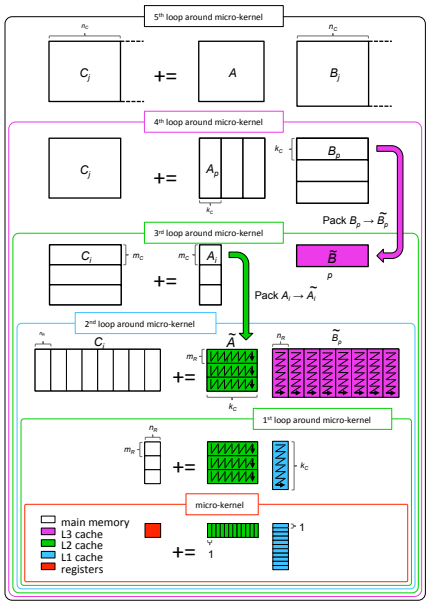
Looking back

- Original Goals

  The BLAS-like Library Instantiation Software (BLIS) framework is a new infrastructure for rapidly instantiating Basic Linear Algebra Subprograms (BLAS) functionality. ….

- Increase Productivity
- Extensible

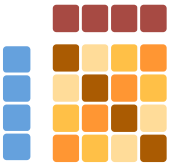
Etc…



F.G. Van Zee, R. van de Geijn, BLIS: A Framework for Rapidly Instantiating BLAS Functionality, ACM TOMS 2015

Ease of use

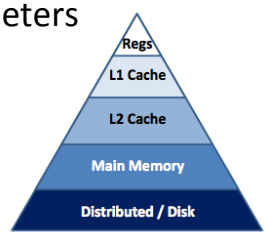
- What happens with a new architecture?
  - Write/Generate micro-kernel



  - Use analytical modeling for parameters

$$C_{A_r} \leq \left\lfloor \frac{m_r}{m_r + n_r}(W_{L1} - 1) \right\rfloor$$

$$k_c = \frac{C_{A_r} N_{L1} C_{L1}}{m_r S_{\text{data}}}$$



T.M. Low, F. Igual, T. M. Smith, E. Quintana-Ortis, Analytical Modeling is Sufficient for High Performance BLIS, ACM TOMS, 2017

Ease of use

- Actual work
  - Create directory (recursive copy and rename)
  - Edit `bli_kernel.h`
  - Drop your kernel(s) into `kernels/3`
  - `./configure new_kernel & make install`

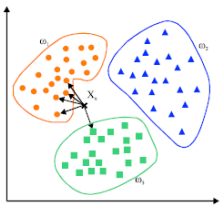
Carnegie Mellon

Electrical & Computer ENGINEERING

Unexpected benefits

- More than just BLAS

kNN



Linkage Disequilibrium

k

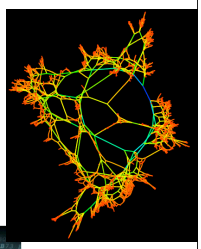
SNP A 010101…

SNP B 010011…

Comm 010001… → 2

N. Alachiotis, Popovici, T.M. Low, Efficient Computation of Linkage Disequilibria as Dense Linear Algebra Operations, HiCOMB 2016

APSP



DNA Fingerprinting

Carnegie Mellon

Electrical & Computer ENGINEERING

What everyone is saying…

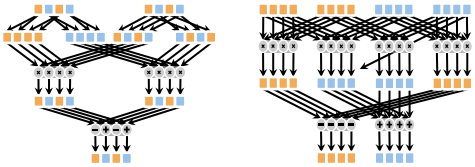
PACKING

Carnegie Mellon

Electrical & Computer ENGINEERING

Reasons for packing

- Work with different data layouts



- Hide additional operations

$$C = f \odot ((A \otimes B) \oplus C)$$

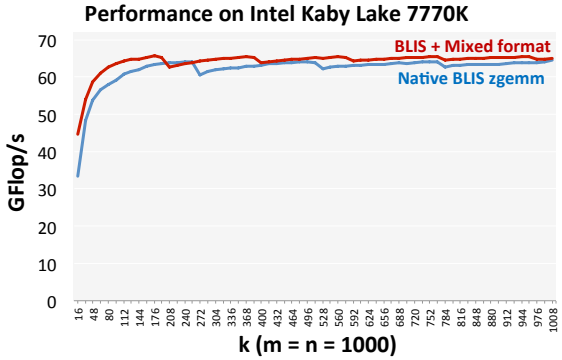
- Generalize the framework

Carnegie Mellon

Electrical & Computer ENGINEERING

Work with Different Data Layout

- Limited SIMD support for complex multiplication
- Interleaved vs Non-interleaved
- Switch format during packing



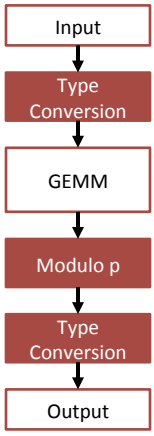
D. Popovici, F. Franchetti, T.M. Low, Mixed Data Format for Vectorized Complex Kernel, HPEC, 2017

Carnegie Mellon

Electrical & Computer ENGINEERING

Hide additional operations

- Finite Field over large primes
  - Reuse BLAS & LAPACK for finite fields



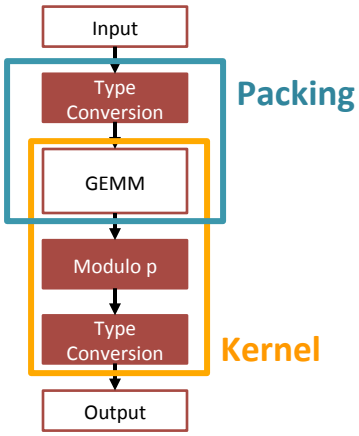
J. Johnson, T.M. Low, M. Lambert, P. Oostema, B. D. Saunders, High-Performance Kernels for Exact Linear Algebra, ACA 2017

Carnegie Mellon

Electrical & Computer ENGINEERING

Hide additional operations

- Finite Field over large primes
  - Reuse BLAS & LAPACK for finite fields



J. Johnson, T.M. Low, M. Lambert, P. Oostema, B. D. Saunders, High-Performance Kernels for Exact Linear Algebra, ACA 2017

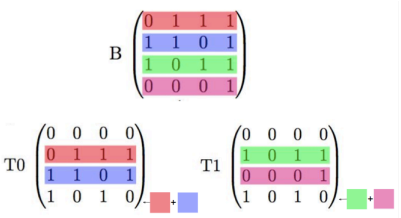
Carnegie Mellon

Electrical & Computer ENGINEERING

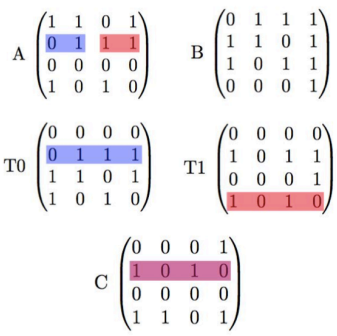
Generalize the framework

- Finite fields for small primes
  - 4 Russians Method

Table Creation



Compute with A as index

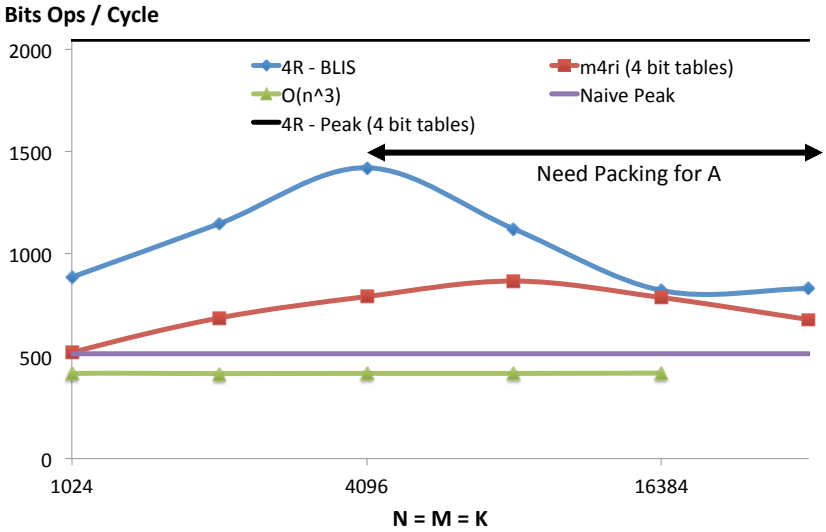


J. Johnson, T.M. Low, M. Lambert, P. Oostema, B. D. Saunders, High-Performance Kernels for Exact Linear Algebra, ACA 2017

Carnegie Mellon

Electrical & Computer ENGINEERING

Performance



J. Johnson, T.M. Low, M. Lambert, P. Oostema, B. D. Saunders, High-Performance Kernels for Exact Linear Algebra, ACA 2017

Carnegie Mellon

Electrical & Computer ENGINEERING

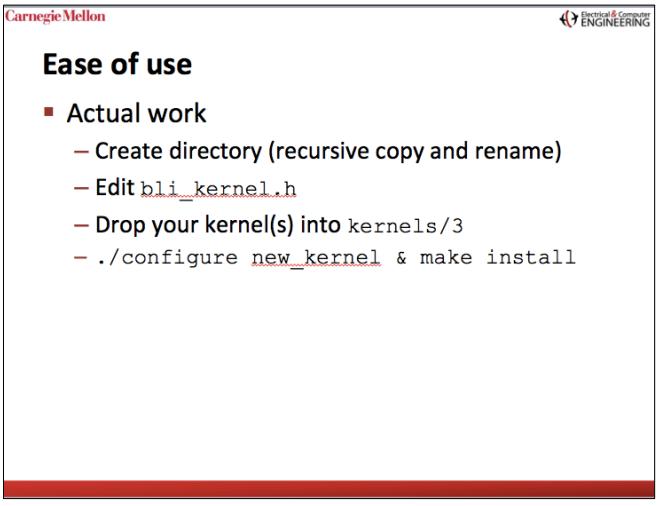
What everyone is NOT saying…

PACKING

Carnegie Mellon

Electrical & Computer ENGINEERING

Remember how easy to add kernels?



Carnegie Mellon

Electrical & Computer ENGINEERING

Introducing packing is not easy

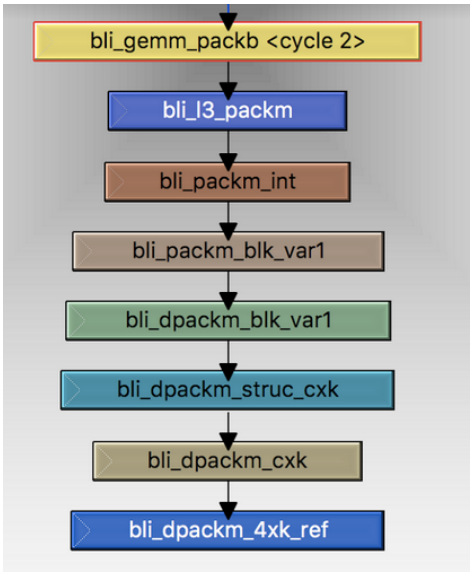
- Files edited or added

frame/1m/bli_l1m_ft.h
frame/1m/bli_l1m_ker.h
frame/1m/packm/bli_packm.h
frame/1m/packm/bli_packm_cxk.c
frame/include/bli_kernel_macro_defs.h
frame/include/bli_kernel_pre_macro_defs.h
frame/include/bli_kernel_prototypes.h
frame/ind/include/bli_kernel_ind_macro_defs.h
frame/ind/include/bli_packm_ind_pre_macro_defs.h
frame/3/gemm/bli_gemm_cntl.c
frame/1m/packm/bli_packm_blk_var2.c
frame/1m/packm/bli_packm_blk_var2.h
frame/1m/packm/bli_packm_cxk_cmu.c
frame/1m/packm/bli_packm_cxk_cmu.h
frame/1m/packm/bli_packm_struc_cxk_cmu.c
frame/1m/packm/bli_packm_struc_cxk_cmu.h
frame/1m/packm/ukernels/bli_packm_cxk_cmu_ref.c
frame/1m/packm/ukernels/bli_packm_cxk_cmu_ref.h
frame/ind/include/bli_packm_cmu_macro_defs.h

Carnegie Mellon

Electrical & Computer ENGINEERING

Calling sequence



Carnegie Mellon

Electrical & Computer ENGINEERING

My wish list

- As easy as adding kernels
- Files to edit or add

frame/1m/bli_l1m_ft.h
frame/1m/bli_l1m_ker.h
frame/1m/packm/bli_packm.h
frame/1m/packm/bli_packm_cxk.c
frame/include/bli_kernel_macro_defs.h
frame/include/bli_kernel_pre_macro_defs.h
frame/include/bli_kernel_prototypes.h
frame/ind/include/bli_kernel_ind_macro_defs.h
frame/ind/include/bli_packm_ind_pre_macro_defs.h
frame/3/gemm/bli_gemm_cntl.c
frame/1m/packm/bli_packm_blk_var2.c
frame/1m/packm/bli_packm_blk_var2.h
frame/1m/packm/bli_packm_cxk_cmu.c
frame/1m/packm/bli_packm_cxk_cmu.h
frame/1m/packm/bli_packm_struc_cxk_cmu.c
frame/1m/packm/bli_packm_struc_cxk_cmu.h
frame/1m/packm/ukernels/bli_packm_cxk_cmu_ref.c
frame/1m/packm/ukernels/bli_packm_cxk_cmu_ref.h
frame/ind/include/bli_packm_cmu_macro_defs.h

- Move from `frame` to `kernel`

Carnegie Mellon

Electrical & Computer ENGINEERING

Summary

- Lighter and more nimble framework
- Refactoring of the packing is proposed
- Inputs from BLIS user and developers wanted

Discussion