SAND2017-13012C

Applications of Compact Batched Kernels

Siva Rajamanickam, Andrew Bradley, Mehmet Deveci,

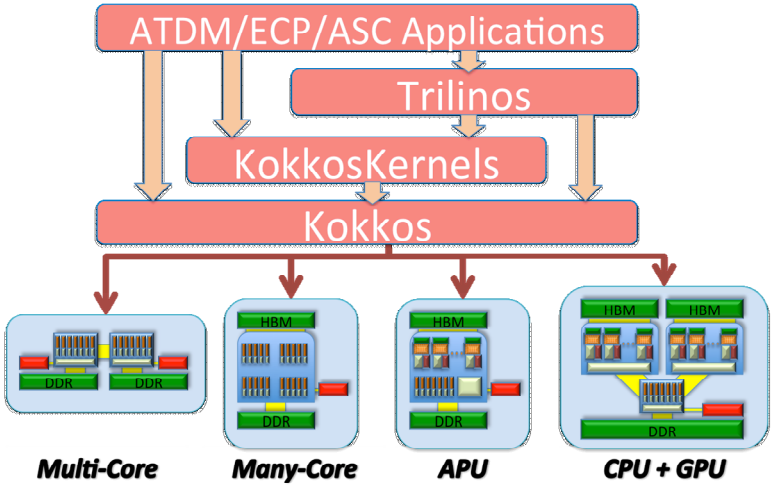
Kyungjoo Kim, Christian Trott

Supercomputing 2017

Sandia National Laboratories is a multimission laboratory managed and operated by National Technology and Engineering Solutions of Sandia, LLC, a wholly owned subsidiary of Honeywell International, Inc., for the U.S. Department of Energy's National Nuclear Security Administration under contract DE-NA-0003525. SAND NO. 2011-XXXXP

KokkosKernels

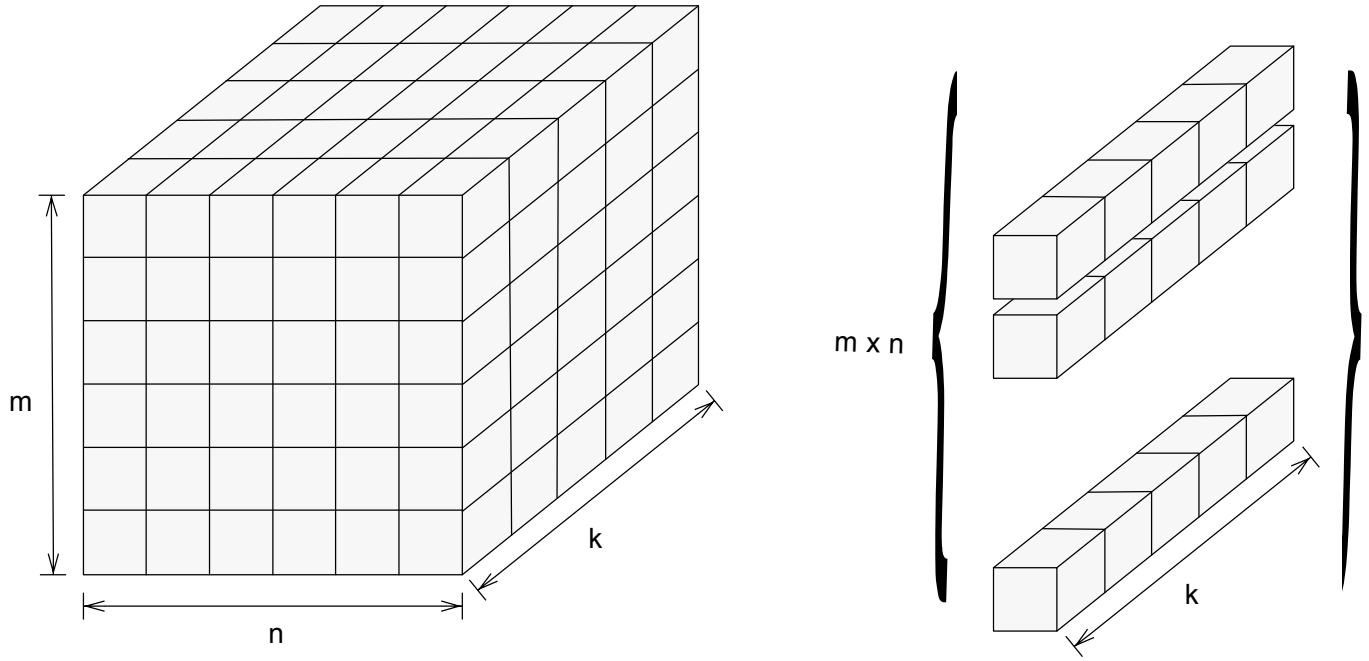
- Kokkos:
  - Layered collection of template C++ libraries
  - Threaded parallel programming model that manages data access patterns (intra-node)
  - Execution spaces, Memory spaces
- Kokkos provides tools for portability
  - Performance portability does not come for free.
  - Not trivial for sparse matrix, and graph algorithms



- KokkosKernels:
  - Layer of performance-portable kernels
- We study design decisions for achieving portability for sparse/dense matrix and algorithms

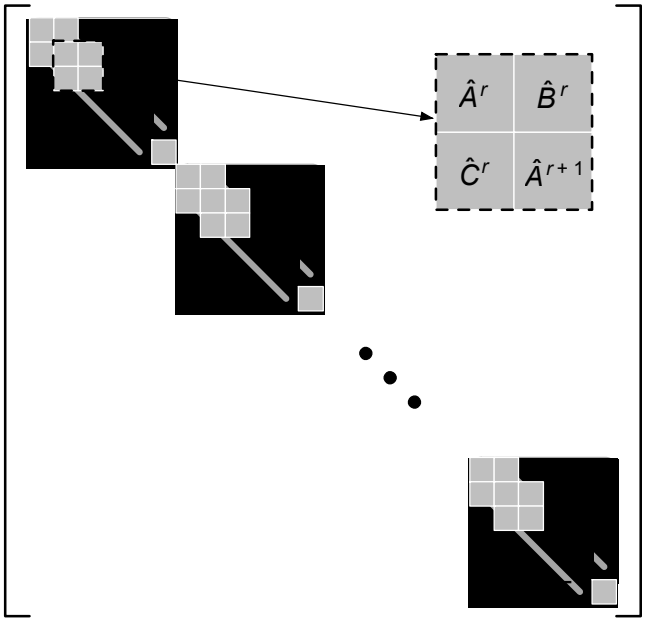
https://github.com/kokkos/kokkos-kernels

Line Preconditioner: Motivation for Batched BLAS



- Consider a block sparse system arising from coupled multi-physics problems.
- Line preconditioner is built by approximating the problem domain as a collection of lines of elements.
- A collection of lines of elements results in a set of block tridiagonal matrices. Block tridiagonal matrices are factorized once per solution (or every nonlinear iteration) and applied (triangular solve) multiple times.

Motivation for Batched BLAS

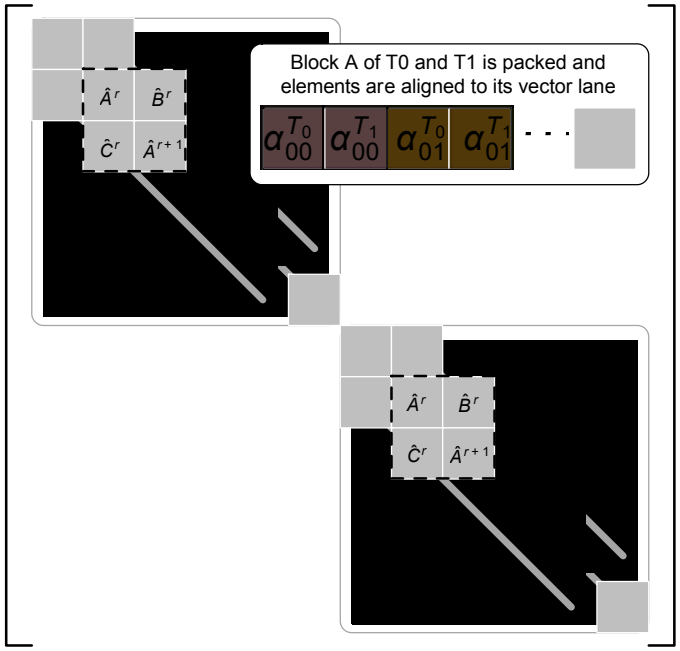


Algorithm 1: Reference impl. TriLU

```
1 for T in {T_0, T_1, ..., T_{m□n-1}} do in parallel
2     for r 0 to k − 2 do
3         Â^r := LU(Â^r);
4         B̂^r := L^{-1} B̂^r;
5         Ĉ^r := Ĉ^r U^{-1};
6         Â^{r+1} := Ĉ^{r+1} − Ĉ^r B̂^r;
7     end
8     Â^{k-1} := {L · U};
9 end
```

- Typical blocksize *b* is selected as 3, 5, 9 and 15, which are related to scientific applications e.g., elasticity, ideal gas and multi-physics fluid problems.
- Limit memory usage up to 16 GB i.e., MCDRAM on KNL and GPU device memory. With this constraint, typical local problems ($m \times n \times k$) are selected as 128 × 128 × 128 for b = 3,5 and 64×64×128 for b = 10,15.
- Batch parallelism is used running a sequential blocktridiagonal factorization consisting of GETRF, TRSM and GEMM within a parallel_for.

Usage of Kokkoskernels Batched BLAS

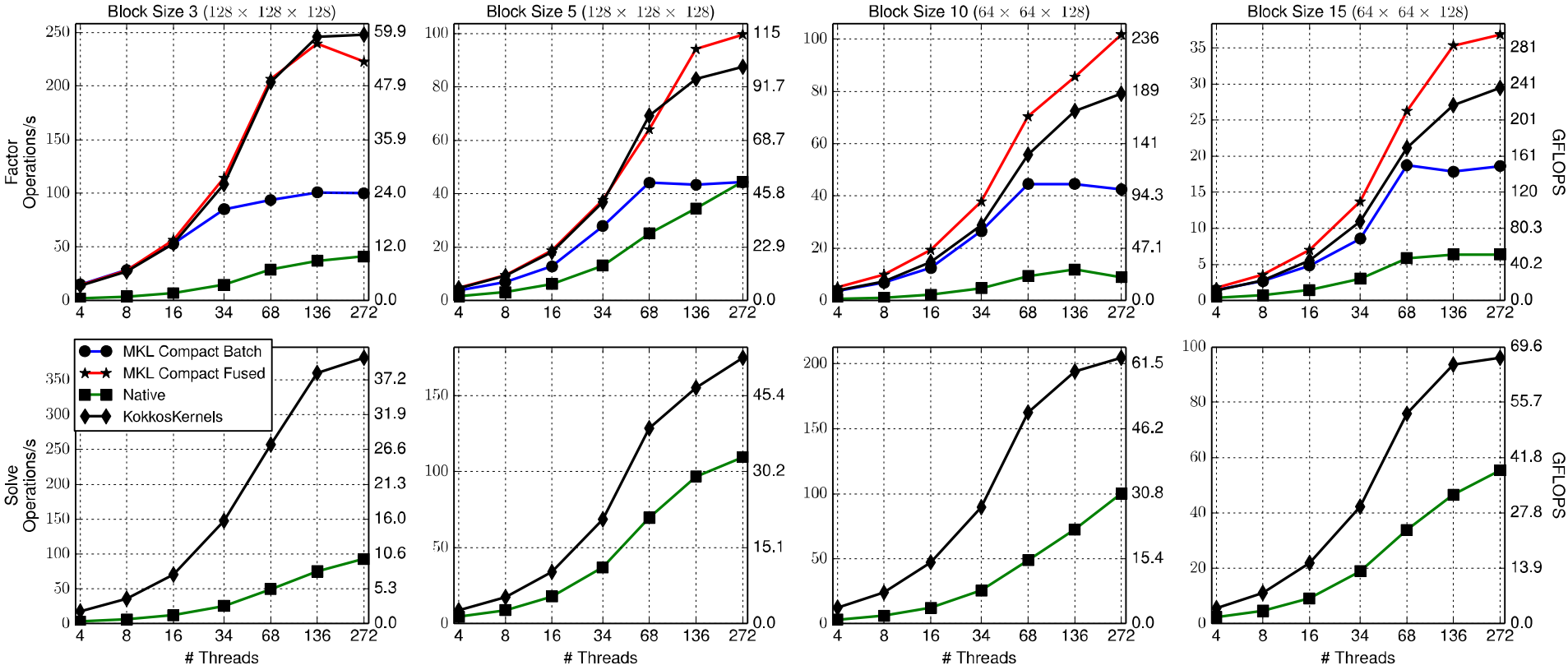


Algorithm 2: Batched impl. TriLU

1 for a pair $T(0,1)$ in $\{\{T_0, T_1\}, \{T_2, T_3\}, \cdots, \{T_{m\square n-2}, T_{m\square n-1}\}\}$ do in parallel
2 for r 0 to $k-2$ do
3 $\hat{A}^{r(0,1)} := LU(\hat{A}^{r(0,1)})$;
4 $\hat{B}^{r(0,1)} := L^{-1}\hat{B}^{r(0,1)}$;
5 $\hat{C}^{r(0,1)} := \hat{C}^{r(0,1)}U^{-1}$;
6 $\hat{A}^{r+1(0,1)} := \hat{C}^{r+1(0,1)} - \hat{C}^{r(0,1)}\hat{B}^{r(0,1)}$;
7 end
8 $\hat{A}^{k-1(0,1)} := \{L \cdot U\}$;
9 end

- On HSW and especially KNL, vectorization is important.
- Traditional BLAS implementations vectorize within a call- bad for small blocks.
- Physics problems typically have small blocks.
- Idea: Vectorize across blocks.
- KokkosKernels provides tools for this, and BLAS implementation using these tools.
- On KNL (HSW), process 8 (4) lines at a time with SIMD : KokkosKernels abstracts these details.

Impact of Compact Batch BLAS
Line Smoother Factor and Solve



- For more, see K. Kim, T.B. Costa, M. Deveci, A.M. Bradley, S.D. Hammond, M.E. Guney, S. Knepper, S. Story, S. Rajamanickam, "Designing Vector-Friendly Compact BLAS and LAPACK Kernels," SC17.
- Thursday Talk: Technical Papers Session - “Fast Multipole Methods and Linear Algebra”, 2.30-3.00 pm.