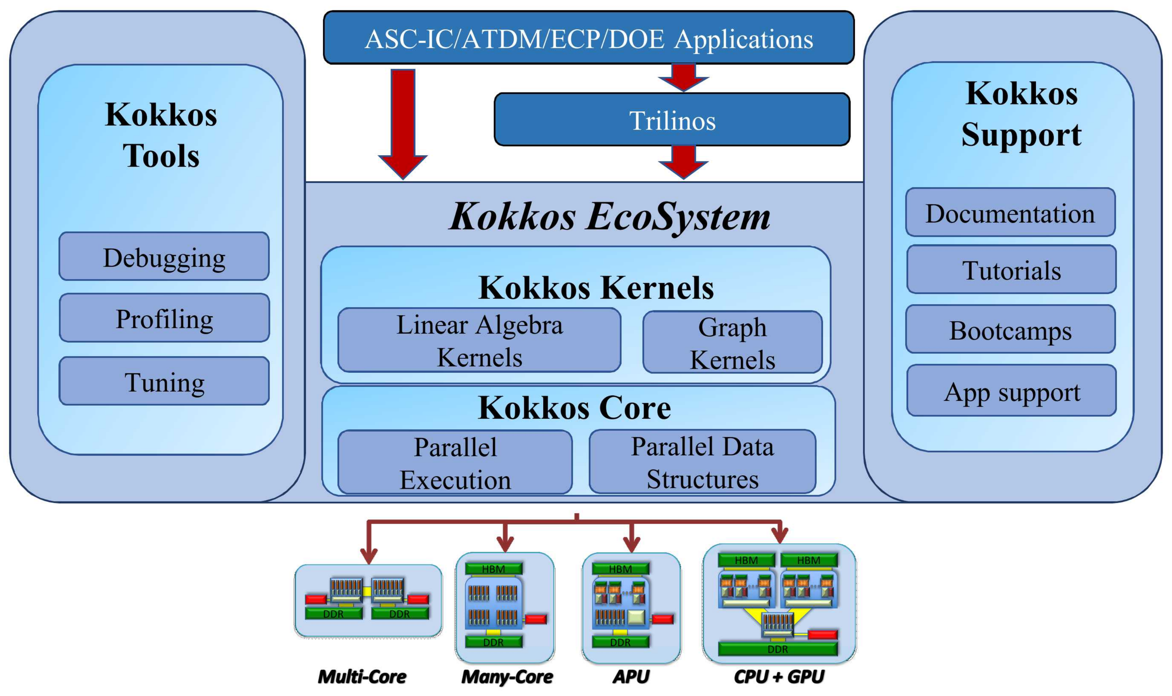
SAND2020-1391C

Kokkos Kernels: Performance Portable Kernels for Sparse/Dense Linear Algebra, Graph and Machine Learning Kernels

Sivasankaran Rajamanickam, Seher Acer, Luc Berger-Vergiat, Vinh Dang, Nathan Ellingwood, Brian Kelley, Kyungjoo Kim, Christian Trott, Jeremiah Wilke, Ichitaro Yamazaki
Center for Computing Research, Sandia National Laboratories, NM, USA

Kokkos Ecosystem provides performance portability for DOE applications



Kokkos ecosystem provides performance portability with **Kokkos Core** programming model, **Kokkos Kernels** for performance portable kernels and **Kokkos Tools** for profiling and debugging.

Kokkos Kernels functionality includes

- **Sparse linear algebra kernels** – matrix-matrix addition, matrix-matrix multiplication, matrix transpose
- **Graph Algorithms** – Distance-1 graph coloring, Distance-2 graph coloring, deterministic coloring, triangle counting
- **Batched BLAS and LAPACK** – LU factorization, matrix-matrix multiplication, triangular solves, and eigen solvers
- **BLAS interface** – BLAS kernels for non-standard data types and interface to vendor BLAS

Kokkos Ecosystem addresses complexity of supporting numerous many/multi-core architectures that are central to DOE HPC enterprise

New Features in Kokkos Kernels 3.0

Sparse Linear Algebra
- ✓ Cluster Gauss-Seidel
- ✓ Sparse ILU factorization
- ✓ Sparse triangular solves for sparse L and U
- ✓ Sparse triangular solves for supernodal L and U
- ✓ Structured sparse matrix vector multiply

Dense Linear Algebra
- ✓ Faster kernels for orthogonalization
- ✓ Complex support for dense LU factorization
- ✓ Interfaces to vendor libraries
- ✓ More BLAS and LAPACK support with Kokkos views

Graph Algorithms
- ✓ Distance-2 graph coloring
- ✓ Faster distance-1 graph coloring
- ✓ Balanced distance-1 coloring
- ✓ Balanced "well shaped" graph clustering
- ✓ RCM ordering for preconditioners

Portable Vectorization
- ✓ Support ARM platforms
- ✓ Improved application performance on CPU, KNL, GPU and ARM
- ✓ Portable SIMD primitive

Team Level Kernels
- ✓ Team level sorting utilities
- ✓ Team level DFS
- ✓ More team level BLAS and LAPACK support

Software
- ✓ CMake support
- ✓ ETI changes to allow ETI file generation at compile time
- ✓ Improved testing
- ✓ Increased robustness

Kokkos Kernels is rapidly growing to support the needs of computational science applications.

Recent Performance Highlights in Kokkos Kernels

New distance-2 coloring algorithm improves multigrid aggregation time by 8x on V100 GPUs. Faster than Zoltan, Colpack and older Kokkos Kernels version.

Cluster Gauss-Seidel preconditioner for smoothing on GPUs reduced CG iterations by 19% on af_shell7, compared to MT-SGS

Kokkos Kernels triangular solver faster than NVIDIA cuSparse

	DAG	Merge	Invert-offdiag	SpMV-DAG	Merge	Invert-offdiag
L-solve (CSC)	3.54x	4.24x	5.31x	3.03x	6.59x	14.11x
U-solve (CSC)	4.00x	4.98x	4.87x	4.39x	8.30x	13.17x

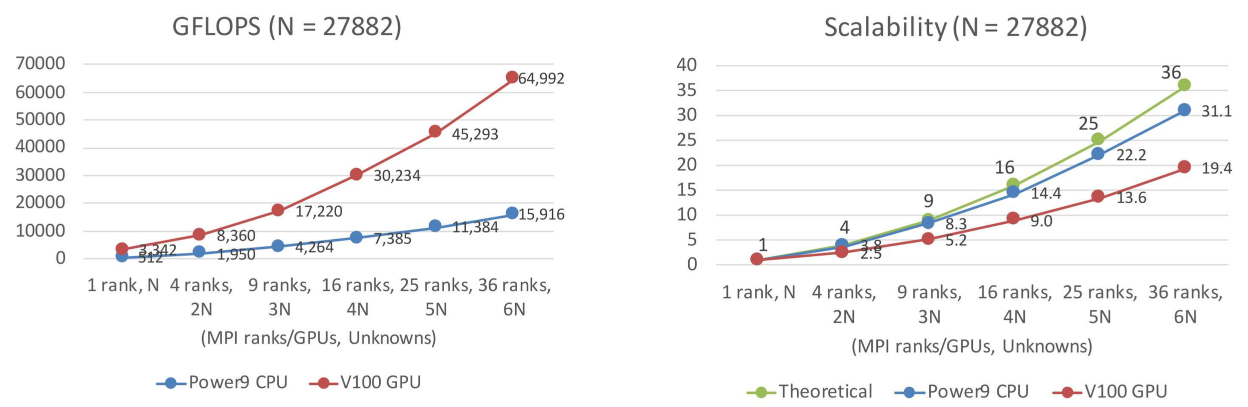
Speedups over CuSparse for A_20x20x20_electricity (n=27,783) on P100

Kokkos Kernels faster than cuBLAS for Tall Skinny matrices

Execution times of GEMM implementations in seconds
(on an NVIDIA Volta V100 GPU, A and B are of size n x s)

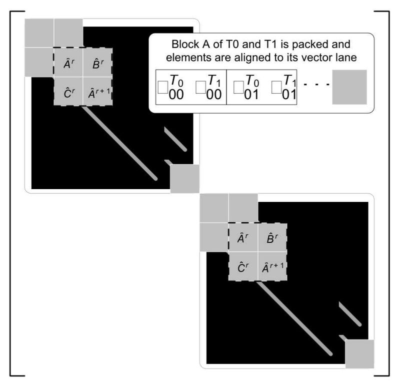
n	s = 3 cuBLAS	s = 3 Kokkos	s = 5 cuBLAS	s = 5 Kokkos	s = 7 cuBLAS	s = 7 Kokkos
1,000	0.02	0.03	0.02	0.03	0.03	0.04
10,000	0.07	0.04	0.07	0.03	0.11	0.05
100,000	0.17	0.04	0.18	0.07	0.21	0.11
1,000,000	2.58	0.19	2.59	0.45	3.00	0.91
10,000,000	45.68	3.08	45.62	6.81	46.40	12.03

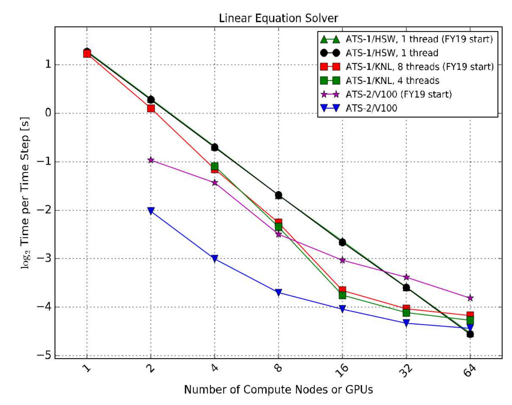
Kokkos Kernels based dense LU solver Adelus scales well to multiple GPUs



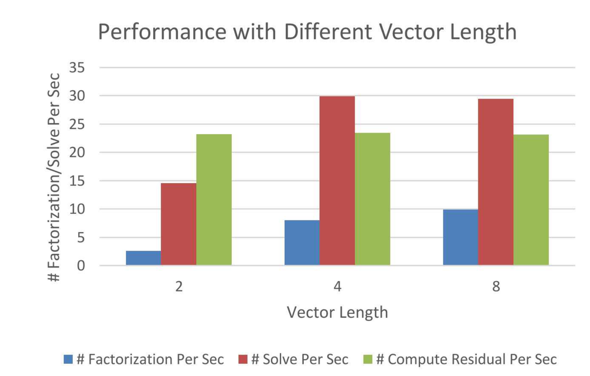
POC: Seher Acer, Vinh Dang, Brian Kelley, Siva Rajamanickam, Ichi Yamazaki

Impact of Kokkos Kernels in Hypersonic Simulations


Compact Data Layout for Block Tridiagonals



Strong Scaling of block tridiagonal solver on Intel HSW, KNL, and NVIDIA V100



Strong Scale of block tridiagonal solver on Intel HSW, KNL, and NVIDIA V100

- Kokkos Kernels and Ifpack2 demonstrated scalable performance for Intel Xeon and NVIDIA GPU architectures
- **Performance Test:** 2x28 Core ARM Thunder X2, 2GHz , **two 128 bit vector units**. A cube domain 256x224x100 with block size 7 (550k block tridiagonal matrices and total 40 million unknowns).
- Factorization and solve phases use compact data layout; compute residual phase uses block CRS matrix format.
- A wider vector length than hardware vector units (128bits) is necessary to hide latency and improve throughput.
- **Factorization and solve achieve 3x and 2x speedup** with 256 bit vectors instead of using h/w vector length (128bit).
- Demonstrated a portable (**no code change but the vector length**), vectorized and thread-scalable block line solver for ARM architectures.

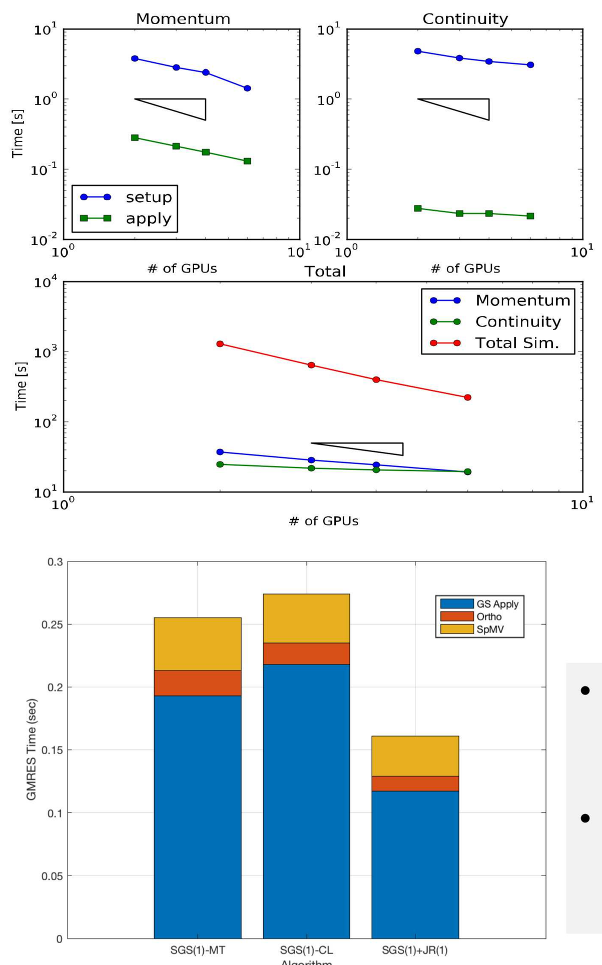
POC: Kyungjoo Kim, Siva Rajamanickam; Micah Howard

Improving linear solvers strong scaling on Summit for ExaWind

ExaWind
- Low Mach CFD wind turbine simulations
- Simulations employ two linear solves
  - Momentum: GMRES/SGS
  - Pressure: GMRES/AMG
- Several new smoothers in Kokkos Kernels

Test problem:
- 5kmx5kmx1km atmospheric boundary layer
- 20m resolution (i.e. ~3.2M nodes)

Observations:
- **All assembly and linear solvers are on GPU**
- **Momentum solver strong scales almost linearly**
- Continuity solver scaling still requires improvement but cost per iteration is low
- Overall simulation time is scaling very well on Summit



- Top: time for a single linear solver setup and apply for the momentum equation (left) and the continuity equation (right) Bottom: the total time associated with solving the momentum equation, the continuity equation and the total simulation time

- New Smoother in Kokkos Kernels faster than past approaches
- Integrated into the Exawind linear solver stack

New algorithms and kernels in Kokkos Kernels to support scalable linear solvers on GPUs

POC: Luc Berger-Vergiat, Jonathan Hu, Brian Kelley, Siva Rajamanickam, Steve Thomas, Ichi Yamazaki,

Ongoing Work and Future Directions

- **Kokkos Kernels**
  - Support complete set of BLAS and LAPACK kernels at the team or vector level
  - Productionize contraction kernels to support machine learning use cases
  - Matrix triple product for multigrid use cases
  - Fused kernels for linear solvers
  - Multi-precision kernels for ECP applications
  - JIT-based performance tuning of the kernels

- **References**

  - *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.*
  - *Howard, M., T. Fisher, M. Hoemmen, D. Dinzl, J. Overfelt, A. Bradley, K. Kim, and S. Rajamanickam. "Employing Multiple Levels of Parallelism for CFD at Large Scales on Next Generation High-Performance Computing Platforms", ICCFD 2018.*

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

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.
SAND2020-xxxx