Sandia National Laboratories

CCR Center for Computing Research

Kokkos Kernels 4.2.0 Release highlights

Presented by

Luc Berger-Vergiat

U.S. DEPARTMENT OF ENERGY | NNSA National Nuclear Security Administration

Sandia National Laboratories is a multimission laboratory managed and operated by National Technology & 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-NA0003525.

Content

- Major new features
  1. BLAS 2 functions
  2. LAPACK
  3. Sparse on stream
  4. ODE: Newton solver
- Other news
  1. TPLs support improvement (oneMKL SpMV, GEMV, dot)
  2. Sparse BSR and MDF improvements
- Upcoming features in 4.3
  1. Algorithms development in LAPACK
  2. TPL support for LAPACK

New major features

BLAS 2 functions

New functions added in BLAS level 2

- SYR (symmetric rank update) fully supported with native and TPL implementations
- HER (hermitian rank update) fully supported with native and TPL implementations
- As all other BLAS functions, they are non-blocking, thread-safe and can be called on stream (for TPL implementation check vendor documentation)

LAPACK

New LAPACK component added

- Moved gesv and trtri from BLAS to Lapack
- No new algorithms or TPL support (LAPACK and MAGMA)
- Future work (already completed) will add support for cuSOLVER and rocSOLVER
- Expect more algorithms in the future (getrf, getri, geqrf, geqri, gesvd…)
- New algorithms may or may not have native implementation

Sparse on streams

Sparse algorithms enhancement for stream/queue execution

- Matrix-Vector product accepts execution space instance
  - Stream/Queue can be attached to individual instances to run on stream
  - Stream/Queue forwarded to TPL when possible
  - Change implemented for CrsMatrix SpMV and BsrMatrix SpMV

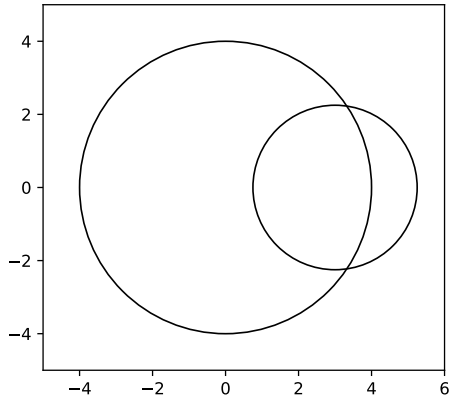
- Diagonal matrix extraction
  - Can divide a matrix in N row-blocks
  - The diagonal part of each row block can be extracted into individual sub-matrices
  - Stream/Queue preconditioners can operate on sub-matrices directly

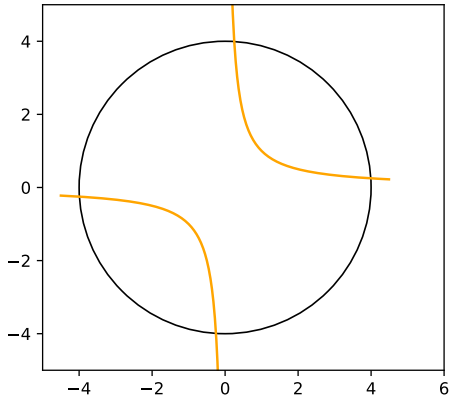
- Gauss-Seidel preconditioner can now run on streams/queues like ILU(K)

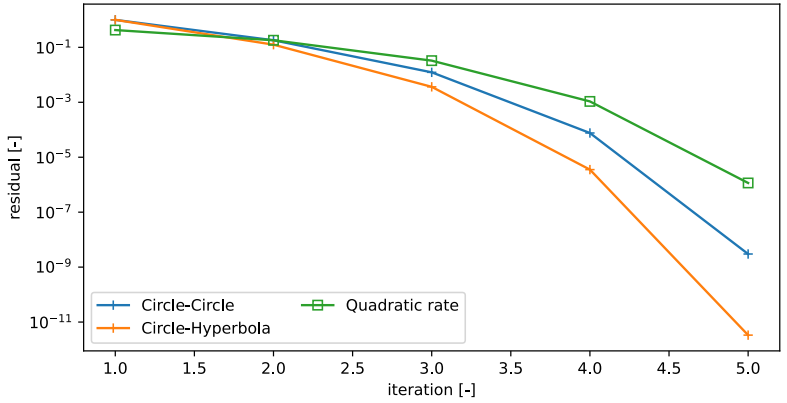
ODE: Newton Solver

ODE solvers are expanding toward implicit algorithms

- Newton solver implemented
- Only support analytical Jacobian
- Always recomputes Jacobian and calls gesv as solver
- No sparse Jacobian supported yet







Other updates

TPL Support: oneMKL

Adding support for matrix-vector multiplication

- GEMV supports oneMKL
- SpMV supports oneMKL

Upcoming support

- BLAS1 oneMKL support (nrm1 and dot already implemented)
- GEMM support for oneMKL
- SpGEMM support for oneMKL

After upcoming updates most used parts of the libraries should be more performant on Intel GPUs

Sparse: Bsr and MDF improvements

- Bsr improvements
  - SpMV TPL support improved with rocSPARSE and cuSPARSE
  - Removed redundant checks for performance
  - Check scalar type carefully before allowing tensor-core code path
- MDF improvements
  - More sub kernels are parallelized
    - count functor
    - Use unordered set to update non-factorized rows (avoid $O(n^2)$ check)

Upcoming work

Kokkos Kernels 4.3 forecast

More algorithms in LAPACK:

- Likely candidates
  - LU (getrf, getri)
  - QR (geqrf, geqri)
  - SVD (gesvd)
  - Let us know if you need something else?

Further TPL work

- Adding cuSOLVER and rocSOLVER for LAPACK vendor support
- Adding BLIS with standard BLAS interface

Thank you for your attention
Any Questions?