SAND2018-6466C

Sparse Matrix-Matrix Multiplication and Related Kernels on Knights Landing

Sandia National Laboratories
C. Siefert
Sandia National Laboratories, Albuquerque, 87185

Problem

- Sparse Matrix-Matrix Multiplication is a common kernel in libraries… but that isn't always the kernel you want!
- Smoothed Aggregation Algebraic Multigrid has unique requirements.
- Matrix-Matrix Damped Jacobi

$$P = (I - \omega D^{-1} A) P_{tent}$$

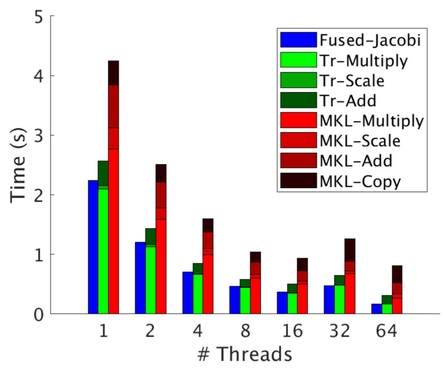
- Matrix-Matrix Multiply and Add

$$C = A\,(B_1 + B_2)$$

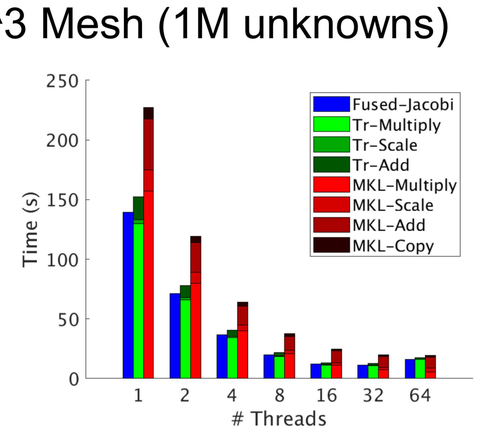
- Both of these are fairly minor modifications of an existing sparse matrix-matrix kernel.
- However, they're *not* provided by libraries.
- Tests: 1 Node (Xeon Phi 7250).
- Test Matrices from Trilinos/MueLu.
  - Jacobi: 3D 27pt stencil w/ tentative prolongator.
  - Multiply-Add: Same matrices as Jacobi, but last 10% of rows put into B2. Approximates "off-processor" rows in an MPI+X sim.

Matrix Jacobi

25^3 Mesh (15.6k unknowns)



100^3 Mesh (1M unknowns)



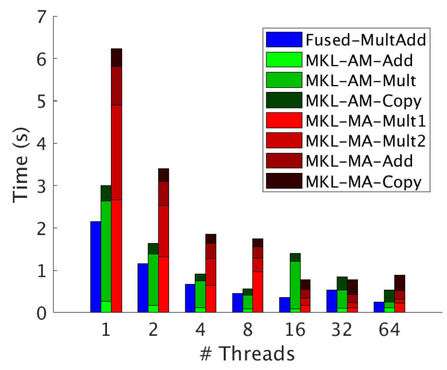
Even when MKL's multiply is faster, the add, scale and copy more than make up for it.

Matrix Multiply-Add

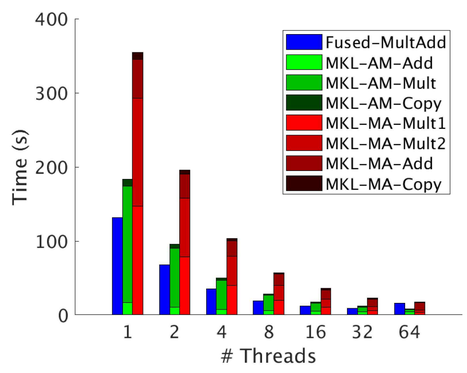
Compare against two MKL-based algorithms:

- AM = Add then Multiply
- MA = Multiply then Add

25^3 Mesh (15.6k unknowns)



100^3 Mesh (1M unknowns)



With rare exceptions, the fused multiply-add is faster than either MKL-based unfused algorithm.

Conclusions / Future Work

- The algorithm you want isn't always the algorithm you have.
- Fused kernels can make a non-trivial impact on performance.
- Looking Forward:
  - Memory allocation optimizations inside kernels.
  - Extension to multiply kernels that aren't dense accumulator-based for use w/ GPUs.

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-NA0003525.

SAND 2017-xxxP