Creating a Standardised Set of Batched BLAS Routines

Samuel D. Relton

 samuel.relton@manchester.ac.uk

 @sdrelton

 www.samrelton.com

 blog.samrelton.com

Joint work with Jack Dongarra, Sven Hammarling, Nick Higham
Pedro Valero-Lara, and Mawussi Zounon

NLAFET

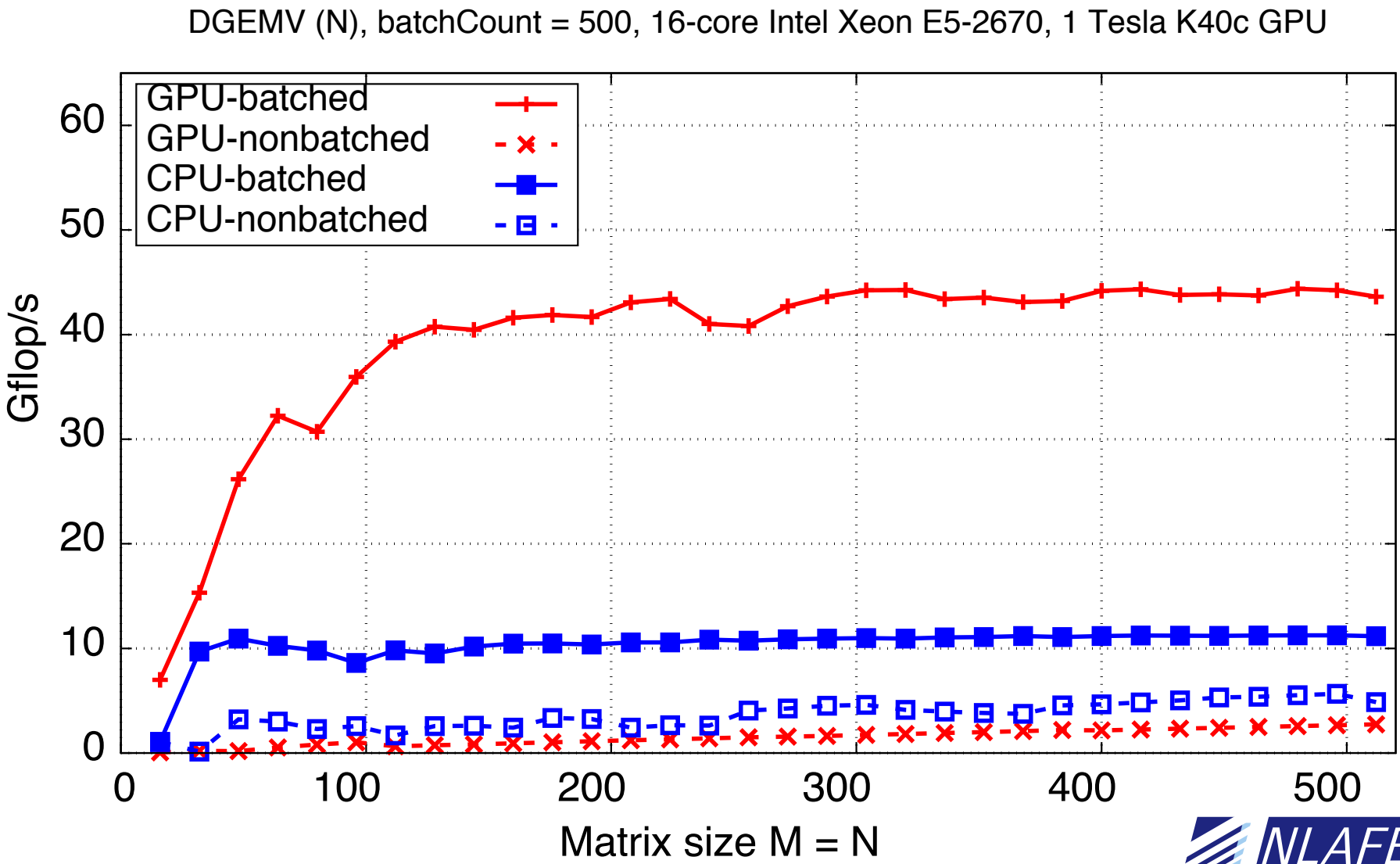
Batched BLAS

BBLAS is for computing multiple BLAS operations in parallel, e.g.

$$C_i \leftarrow \alpha_i A_i B_i + \beta_i C_i, \qquad i = 1 : N.$$

- Matrices are small (sizes 2–256 for example).
- Intel MKL: batched GEMM.
- cuBLAS: batched GEMM, TRSM, and LU, QR, least-squares solvers.
- MAGMA: batched GEMM, TRSM, HERK and RBT, QR, LU, Cholesky solvers.
- Applications in machine learning, multifrontal solvers, image processing, fluid dynamics, astrophysics etc.

BBLAS - Motivation



NLAFET

BBLAS - Issues to Face

- No standard API between vendors & academics.
- No research into impact of API and memory layout on performance.
- No full reference implementation of BBLAS.

Solutions:

- Run workshops to allow discussion between vendors & academics.
- Perform research into effect of API choices on performance.
- Obtain community feedback to shape the standard.

BBLAS - More Info

- Tech Reports:
  - *A Proposed API for Batched Basic Linear Algebra Subprograms.* MIMS EPrint 2016.25.
  - *Workshop on Batched, Reproducible, and Reduced Precision BLAS.* MIMS EPrint 2016.41
  - *A Comparison of Potential Interfaces for Batched BLAS Computations.* MIMS EPrint 2016.42
- Websites:
  - `https://software.intel.com/en-us/articles/introducing-batch-gemm-operations`
  - `http://docs.nvidia.com/cuda/cublas/#batching-kernels`