THE UNIVERSITY OF TENNESSEE KNOXVILLE

ICL Innovative Computing Laboratory

BoF 156:

Batched, Reproducible, and Reduced Precision BLAS

Jack Dongarra, Piotr Luszczek

BoF Schedule

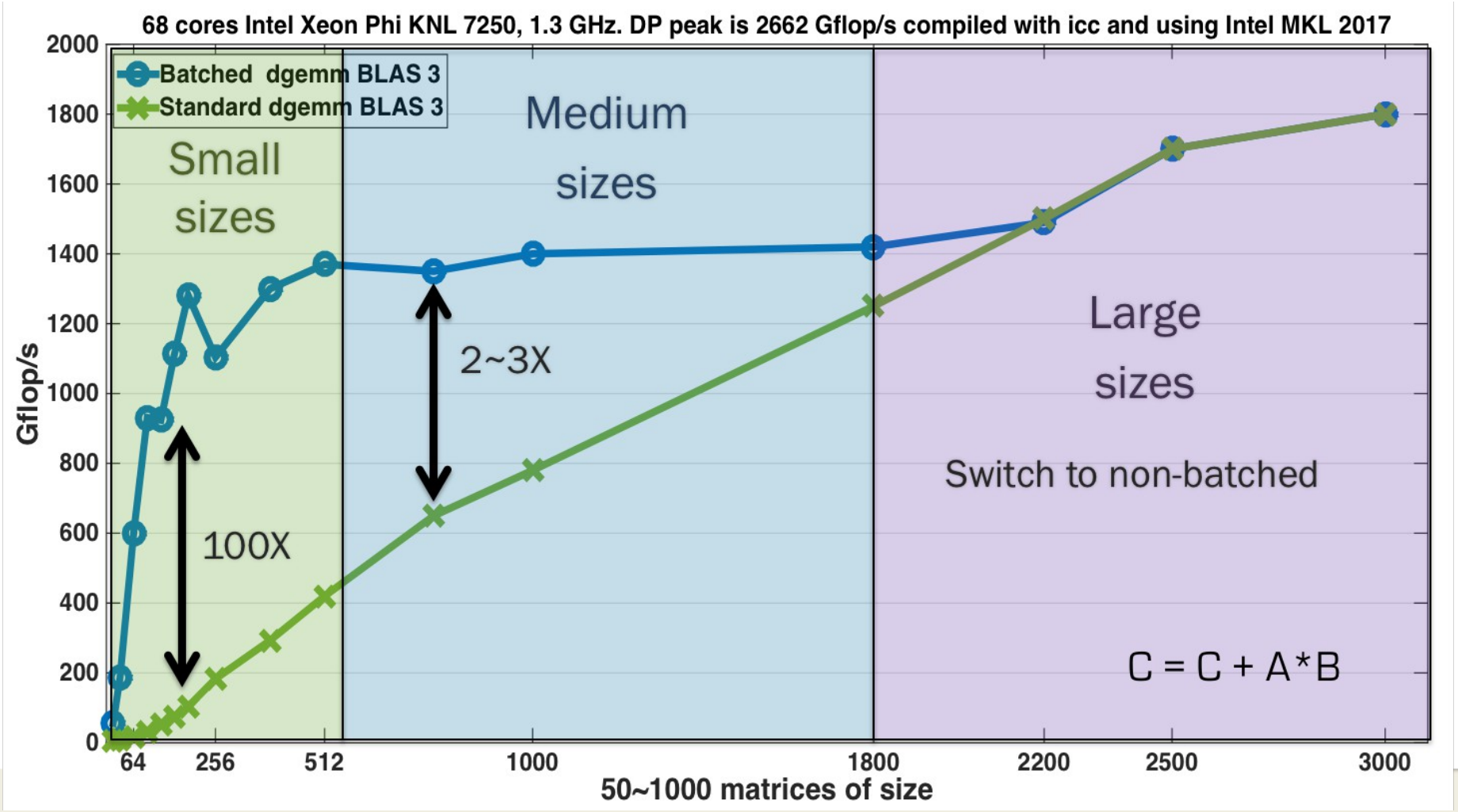
- Introductions
  - Piotr Luszczek, University of Tennessee
- Batched BLAS at NVIDIA and possibly a short overview of CUTLASS
  - Cris Cecka, NVIDIA
- Batched BLAS and related functionality inside Intel MKL
  - Alexander Kalinkin, Intel
- An update on the Next-Generation BLAS Proposal
  - Jason Riedy, Georgia Tech
- Update on Batched implementation and usage in Kokkos Kernels
  - Siva Rajamanickam, Sandia Labs

Batched BLAS Origin: Applications

- Many BLAS-like computations on small sized matrices and vectors
  - Machine learning / DNNs
  - Data mining / analytics
  - High-order FEM,
  - Graph analysis,
  - Neuroscience,
  - Astrophysics,
  - Quantum chemistry,
  - Signal processing
  - And more

- Fixed-size batches
- Variable-size batches
- Dynamic batches
- Tensors in batches

Batched BLAS Origin: Performance



Batched BLAS Specification

Batched BLAS (Basic Linear Algebra Subprograms) 2018 Specification

Jack Dongarra[*] [†] [‡], Iain Duff [§], Mark Gates[*], Azzam Haidar[*], Sven Hammarling [¶], Nicholas J. Higham [¶], Jonathan Hogg [%] , Pedro Valero Lara [#] , Piotr Luszczek[*], Mawussi Zounon [¶], Samuel D. Relton [$] , Stanimire Tomov[*], Timothy Costa [|] , and Sarah Knepper [|]

*Innovative Computing Laboratory, University of Tennessee, USA
†Oak Ridge National Laboratory, Tennessee, USA
‡School of Computer Science and School of Mathematics, The University of Manchester, Manchester, UK
§STFC Rutherford Appleton Laboratory, Harwell Oxford, UK
¶School of Mathematics, The University of Manchester, Manchester, UK | Intel Corp., USA
Barcelona Supercomputing Center (BSC, Barcelona, Spain
$ LIHS, The University of Leeds, UK
% Apple Inc.,USA

July 23, 2018

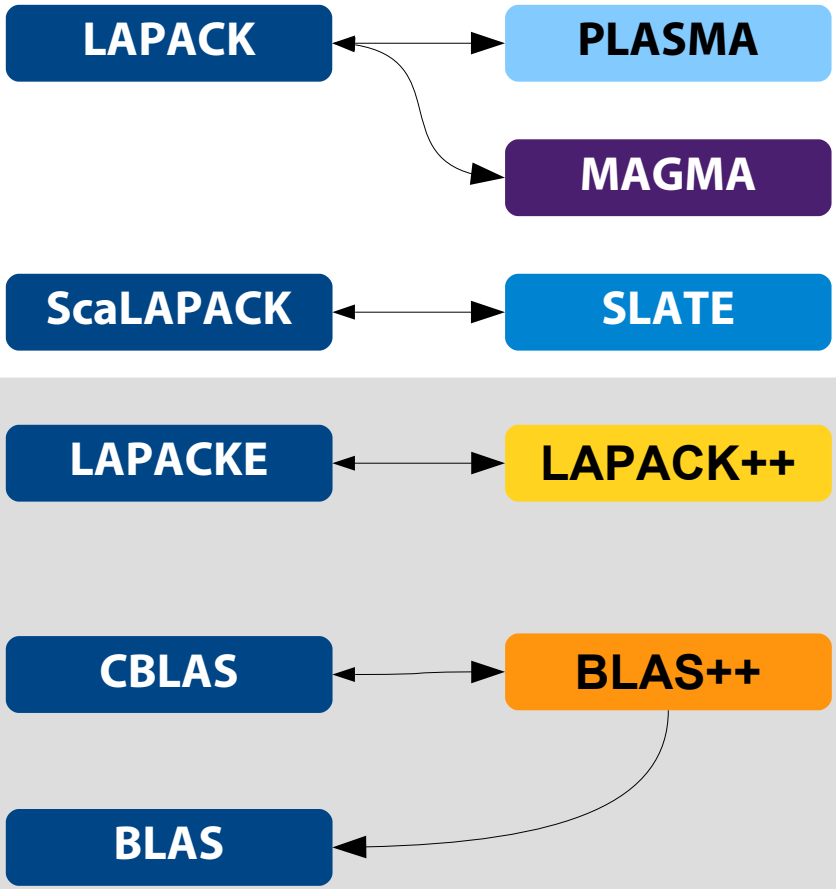
Abstract

Sample Routine from the Specification

```
BLAS <*>gemm batch(
group count     : int                          : In,
group sizes     : int[group_count]             : In,
layout          : enum Layout                  : In,
trans_A         : enum Transpose[count]        : In,
trans_B         : enum Transpose[count]        : In,
m               : int[count]                   : In,
n               : int[count]                   : In,
k               : int[count]                   : In,
alpha           : <float,double,...>[count]    : In,
A               : <float,double,...>*[count]   : In,
ld_A            : int[count]                   : In,
B               : <float,double,...>*[count]   : In,
```

On-Going Work

http://www.netlib.org http://icl.utk.edu/research



- Working with Kitware for CMake integration
- BLAS++ and LAPACK++ to be integrated with next LAPACK releasee

https://bitbucket.org/icl/lapackpp

https://bitbucket.org/icl/blaspp

MAGMA Batched Computations

- Team
  - Ahmad Abdelfattah
  - Azzam Haidar
  - Stan Tomov
- Team leader
  - Jack Dongarra

- Functionality (fixed, variable)
  - Batched BLAS
  - LAPACK
    - LU, QR, Cholesky, ...
  - Tensor contractions
  - Deep learning kernels