PERFORMANCE

BATCHED LEVEL 3 BLAS DGEMM EXAMPLE

\[
\text{DGEMM (N)N) batch\_count = 500}
\]

CPU: Intel Xeon Gold 6140, 18 cores, 2 sockets
GPU: Tesla Titan V, 80 SMs

BATCHED LEVEL 2 BLAS DGEMV EXAMPLE

\[
\text{DGEMM (N)N) batch\_count = 100000}
\]

CPU: Intel Xeon Gold 6140, 18 cores, 2 sockets
GPU: Tesla Titan V, 80 SMs

ADVANTAGES

More efficient and portable implementations
HPC numerical library for modern architectures
Better hardware utilization and energy efficiency
Encourages and simplifies community efforts to build higher-level algorithms on top of Batched BLAS
Multiple precisions: 16, 32, and 64 bits in real and complex domains

TECHNOLOGIES

- Multicore
- Accelerators
- Fused Kernels
- Multiple Streams
- OneAPI
- OpenMP
- NVIDIA CUDA
- ROCm

IN COLLABORATION WITH

WITH SUPPORT FROM

SPONSORED BY
Batched BLAS:
multiple independent BLAS operations on small matrices grouped together as a single routine

Numerous applications require Batched BLAS:
- Structural mechanics
- Astrophysics
- Direct sparse solvers
- High-order FEM simulations

WORKSHOPS

Workshop on Batched, Reproducible, and Reduced Precision BLAS 2017
Atlanta, GA

Workshop on Batched, Reproducible, and Reduced Precision BLAS 2016
Knoxville, TN

PAPERS AND RELATED MATERIAL


ReproBLAS
http://bebop.cs.berkeley.edu/reproblas/


Compact Batched API Document
Intel MKL Team
https://www.dropbox.com/s/gplop3sxhg8le3r/MKL_COMPACT_v4.docx?dl=0