INNOVATIVE COMPUTING LABORATORY

SI2-SSE: MATEDOR

MAtrix, TEnsor, and Deep-learning Optimized Routines

Stanimire Tomov (PI)
Ahmad Abdelfattah
Jack Dongarra
UNIVERSITY OF TENNESSEE

Azzam Haidar
NVIDIA

MATEDOR SCOPE

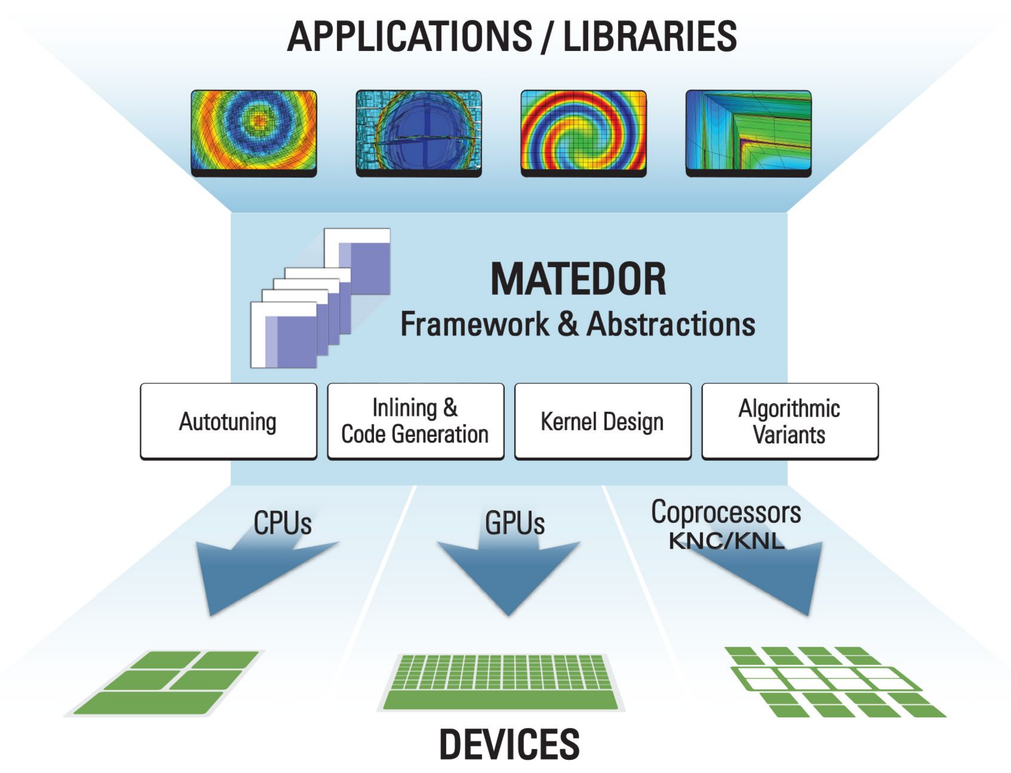
The MAtrix, TEnsor, and Deep-learning Optimized Routines (MATEDOR) project provides software technologies and standard APIs, along with a sustainable and portable library for large-scale computations whose individual components rely on very small matrix or tensor computations. The main target is the acceleration of applications from important fields that fit this profile, including deep learning, data mining, astrophysics, image and signal processing, hydrodynamics, and more.

Standard Interface for Batched Routines

Working closely with interested application developers, we defined modular, language agnostic interfaces that can be implemented so as to work seamlessly with the compiler and be optimizable using techniques such as code replacement and inlining. This provides the application developers, compilers, and runtime systems with the option of launching batched workloads using a single call according to the standard interface. This would allow the entire linear algebra (LA) community to collectively address a wide range of small matrix or tensor problems. Success in such an effort was possible through innovations in the interface design, computational and numerical optimizations, as well as packaging and deployment at the user end to trigger final stages of tuning at the moment of execution.

Sustainable and Performance-Portable Software Library

We demonstrated the power of the MATEDOR interface by delivering a high-performance numerical library for batched LA subroutines autotuned for the modern processor architecture and system designs. The MATEDOR library includes LAPACK routine equivalents for many small dense problems, tensor, and application-specific operations (e.g, for deep-learning). These routines are constructed as much as possible out of calls to batched BLAS routines and their look-alikes required in sparse computation context.



Enabling Technologies

MATEDOR develops enabling technologies for very small matrix and tensor computations, including (1) **autotuning**, (2) **inligning**, (3) **code generation**, and (4) **algorithmic variants**. We define the success of the research conducted and the software developed under the MATEDOR project as being able to automate these four aspects to allow for both flexibility and close-to-optimal performance of the final code that gets used by the domain scientist.

Standard APIs (for Batched BLAS and LAPACK):

Proposed API is very similar to the standard BLAS/LAPACK API

```
void
blas_dgemm_batched(
            blas_trans_t transA, blas_trans_t transB,
            blas_int_t m, blas_int_t n, blas_int_t k,
            double alpha,
            double const * const * dA_array, blas_int_t ldda,
            double const * const * dB_array, blas_int_t lddb,
            double beta,
            double **dC_array, blas_int_t lddc,
            blas_int_t batchCount, blas_int_t *info );
```

Community Effort and Activities Towards Standardization

Batched BLAS BoF 2017 @ SC17
https://sc17.supercomputing.org/presentation/?id=bof147&sess=sess370

Batched BLAS Workshop 2017 @ Georgia Tech
http://bit.ly/Batch-BLAS-2017

Batched BLAS Symposium @ SIAM PP 2018
https://www.siam.org/meetings/pp18/

SIAM Conference on Parallel Processing for Scientific Computing
March 7-10, 2018
Waseda University
Tokyo, Japan

Batched BLAS Symposium @ SIAM CSE 2019
https://www.siam.org/conferences/cm/conference/cse19

SIAM Conference on Computational Science and Engineering
February 25 – March 1, 2019
Spokane Convention Center,
Spokane WA, USA

Broader Impact

MATEDOR is application-motivated and designed to impact application areas from deep-learning, to data mining, metabolic networks, CFD, solvers, image and signal processing, and others that need small matrix/tensor computations.

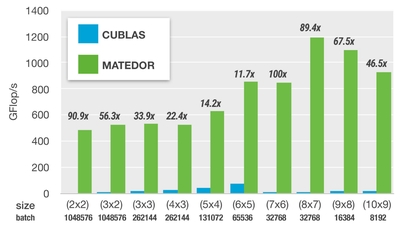
PUBLICATIONS

1. A. Abdelfattah, S. Tomov and J. Dongarra, "**Fast Batched Matrix Multiplication for Small Sizes Using Half-Precision Arithmetic on GPUs**," 2019 IEEE International Parallel and Distributed Processing Symposium (IPDPS 2019)
2. A. Abdelfattah, A. Haidar, S. Tomov, and J. Dongarra, "**Tensor Contractions using Optimized Batch GEMM Routines,**" March 26-29 2018, GPU Technology Conference (GTC), Poster, San Jose, CA. [Online]. Available: http://icl.cs.utk.edu/magma/software/
3. L. Ng, K. Wong, A. Haidar, S. Tomov, and J. Dongarra, "**MagmaDNN High-Performance Data Analytics for Manycore GPUs and CPUs**," December 2017, MagmaDNN, 2017 Summer Research Experiences for Undergraduate (REU), Knoxville, TN. [Online]. Available: http://icl.cs.utk.edu/magma/software/
4. A. Haidar, A. Abdelfattah, M. Zounon, S. Tomov, and J. Dongarra, "**A Guide For Achieving High Performance With Very Small Matrices On GPU: A case Study of Batched LU and Cholesky Factorizations**," IEEE Transactions on Parallel and Distributed Systems, vol. PP, no. 99, pp. 1–1, 2017.
5. A. Abdelfattah, A. Haidar, S. Tomov, and J. Dongarra, "**Batched one-sided factorizations of tiny matrices using GPUs: Challenges and countermeasures,**" Journal of Computational Science, 2018. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S1877750317311456
6. I. Yamazaki, A. Abdelfattah, A. Ida, S. Ohshima. S. Tomov, R. Yokota, J. Dongarra, " **Performance of Hierarchical-matrix BiCGStab Solver on GPU clusters,**" 2018 IEEE International Parallel & Distributed Processing Symposium.

Breadth of MATEDOR's Impact on Application Domains

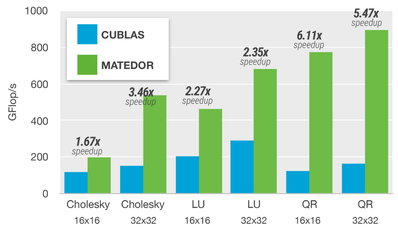
Tensor Contractions High Order FEM & Applications

Tensor Contractions: computing $B^T D (BAB^T) B$, double precision, Tesla V100 GPU



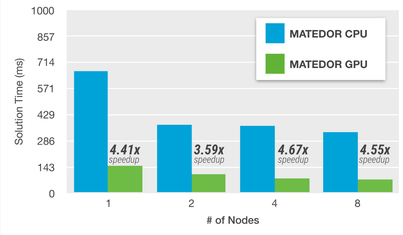
Sparse/Dense Solvers & Preconditioners

Batch Matrix Factorization, 100k matrices, double precision, Tesla V100 GPU



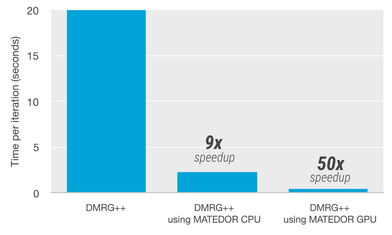
Hierarchical Linear Solvers on GPU clusters

Hierarchical Linear Solver, 2 P100 GPUs per node



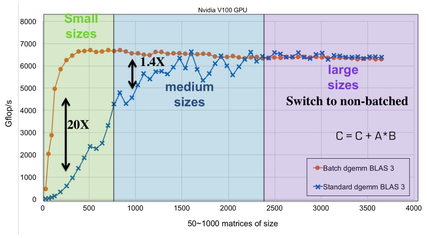
Density Matrix Renormalization Group DMRG++

DRMG++ Acceleration using MATEDOR Batched computations



Deep Neural Network & Data Analytics

Batched DGEMM acceleration on V100 GPU in DNN computational backends



INNOVATIVE COMPUTING LABORATORY

THE UNIVERSITY OF TENNESSEE KNOXVILLE

NSF