STANDARD INTERFACE FOR BATCHED ROUTINES

Working closely with affected application communities, we will define modular, language-agnostic interfaces that can be implemented to work seamlessly with the compiler and be optimizable using techniques such as code replacement and inlining. This will provide the developers of applications, compilers, and runtime systems with the option of expressing as a single call to a routine from the new batch operation standard, and would allow the entire linear algebra (LA) community to collectively attack a wide range of small matrix or tensor problems. Success in such an effort will require innovations in interface design, computational and numerical optimization, as well as packaging and deployment at the user site to trigger final stages of tuning at the moment of execution.

SUSTAINABLE AND PERFORMANCE-PORTABLE SOFTWARE LIBRARY

We will demonstrate 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 will include LAPACK routine equivalents for many small dense problems, tensor, and application-specific operations, e.g., for deep learning; these routines will be constructed as much as possible out of calls to batched BLAS routines and their look-alikes required in sparse computation context.

APPLICATIONS / LIBRARIES

MATEDOR Framework & Abstractions


BREADTH OF MATEDOR’S IMPACT ON APPLICATION DOMAINS

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++
Batched DGEMM acceleration on V100 GPU in DNN computational backends

DEEP NEURAL NETWORKS AND DATA ANALYTICS

This work is partially supported by NSF Grant No. OAC 1740250 and CSR 1514286, NVIDIA, and the Department of Energy under the Exascale Computing Project (17-SC-20-SC and LLNL subcontract under DOE contract DE-AC52-07NA27344).