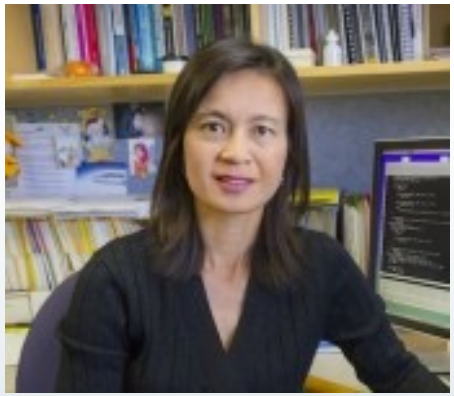


Sherry Li, Lead, Scalable Solvers Group

Areas: Parallel computing, high performance algebraic solvers, sparse matrix computations, combinatorial scientific computing, mathematical software

Products:

- (Co)authored over 130 papers.
- Math software: SuperLU, XBLAS, ARPREC/QD, STRUMPACK, etc.

Experience

- Ph.D. Computer Science, UC Berkeley (1996)
- Computer System Engineer (1996-2001)
- Staff Scientist (2001-2014)
- Senior Scientist (2014-)

My summer internships are immensely valuable

- 2nd summer in PhD program
  - Floating point Group, Sun Microsystems (Oracle)
  - Vectorize transcendental functions (sin, exp, etc.)
- 3rd summer – led to PhD thesis
  - Xerox Palo Alto Research Center (PARC)
    - Mentor: John Gilbert
    - Other leading researchers: Rob Schreiber, Ed Rothberg
  - Develop the first supernodal sparse LU code
    - Continued to parallelize it after returned to school

Scalable Solvers Group

https://crd.lbl.gov/divisions/amcr/applied-mathematics-dept/scalable-solvers/

The group develops fast, parallel algorithms and software for solving the linear and eigenvalue algebraic systems, and deliver the solvers tools to the broad community through libraries and collaboration with domain scientists.

Xiaoye Sherry (Sherry) Li
Senior Scientist & Group Lead
+1 510 486 6684 | XSLi@lbl.gov

Mark F Adams
Research Scientist
MFAdams@lbl.gov

Pieter Ghysels
Research Scientist
+1 510-486-5594 | PGhysels@lbl.gov

Osni Marques
Staff Scientist
+1 510 486 5290 | OAMarques@lbl.gov

Michael L. Minion
Staff Scientist
MLMinion@lbl.gov

Yang Liu
Research Scientist
510-486-5283 | liuyangzhuan@lbl.gov

Yu-Hang (Maxin) Tang
Research Scientist, Career-Track
tang@lbl.gov

Roel Van Beeumen
Research Scientist
+1 (510) 495-2189 | RVanBeeumen@lbl.gov

Chao Yang
Senior Scientist
+1 510 486 6424 | CYang@lbl.gov

Postdoctoral Researchers

Wajih Boukaram
Postdoctoral Fellow
+1 (510) 486-6684 | wajih.boukaram@lbl.gov

Daan Camps
Postdoctoral Fellow
dcamps@lbl.gov

Lisa Claus
Postdoctoral Scholar
LClaus@lbl.gov

Alice Gatti
Postdoctoral Scholar
agatti@lbl.gov

Hengrui Luo
Postdoctoral Scholar
hrluo@lbl.gov

Jordi Wolfson-Pou
jwolfsonp@lbl.gov

Jia Yin
Postdoctoral Scholar
jiayin@lbl.gov

Faculty Scientsts

Zhaojun Bai
Faculty Scientist, UC Davis
+1 510 495 2851 | zbai@ucdavis.edu

James Demmel
Faculty Scientist, UC Berkeley
+1 510 495 2851 | demmel@berkeley.edu

John Gilbert
Faculty Scientist, UC Santa Barbara
+1 510 495 2851 | gilbert@cs.ucsb.edu

Algebraic solvers are fundamental tools

Black-box solvers

Purely algebraic, matrix input

$Ax = b,\ \ Ax = \lambda x$

Application-specific linear algebra tools

Specialized to accelerator, chemistry, fusion, materials, ML, nuclear physics, quantum comput., transportation, . . .

Improve algorithmic efficiency, parallelism, and solution quality

- Multilevel, multigrid, hierarchical algorithms
- Reduce communication / synchronization
- Increase concurrency
- Improve convergence
- HPC-aware: GPUs, …

Expertise, capabilities

(Most software packages are open source, BSD License)

- **Dense linear algebra** (LAPACK/ScaLAPACK, ButterflyPACK)
- **Sparse linear solvers**
  - Direct solvers (STRUMPACK, SuperLU, symPACK)
  - Multigrid (GAMG in PETSc)
  - Algebraic preconditioner (STRUMPACK)
  - Communication-reducing Krylov solvers
- **Eigenvalue calculations**
  - Lanczos / Arnoldi iterative eigensolver (BLZPACK, PARPACK)
  - Non-Hermitian eigensolver for interior eigenvalues (software: GPLHR)
  - Application-specific structured eigensolvers
    - Electronic structure, quantum chemistry, nuclear physics (PEXSI, BSEPAC, SpectrumSlicing)
    - Linear, nonlinear, parameterized eigenvalue problems
- **Multi-linear algebra (tensor) (**FunFact**)**
- **High-precision floating-point arithmetic** (QD, ARPREC, XBLAS)
- **High-order PDE solvers, parallel-in-time PDE solvers** (PFASST)
- **Machine learning for sciences** (GAP, GPTune)
- **Quantum computing algorithms** (QFT, QPIXL)

R&D in fast solvers

Linear solvers, eigensolvers, preconditioners, …

"Fast" == asymptotically lower arithmetic and/or communication

Areas:

- **Hiding/avoiding communication/synchronization**
  - Latency / bandwidth / flops / memory
- **Randomization: sampling, projection**
- **Low-rank approximations**
  - Exploit localization
- **Hierarchical & multilevel methods**
  - Multigrid, $\mathcal{H}^2$/HSS matrices, FMM, Butterfly, FFT

Linear Solvers

Batched all GPU controlled solvers for many small systems in PETSc

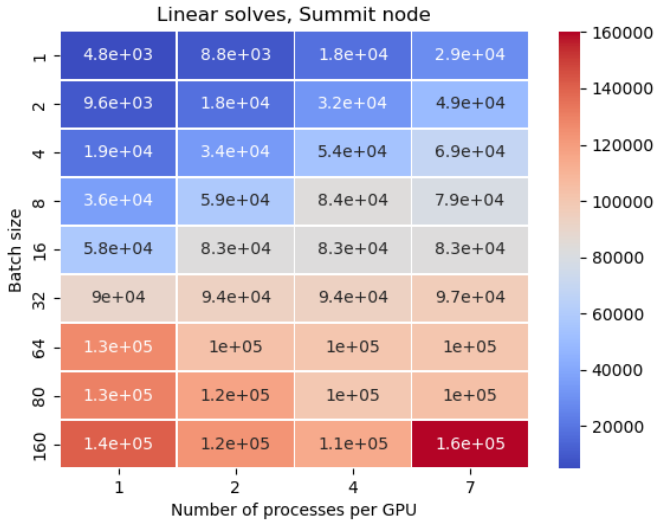
Mark F. Adams

Objectives

- **Provide support for many small system solves on GPUs**
- **Port subset of PETSc solvers to new all GPU solver infrastructure in PETSc**
- **Batch systems to amortize kernel launch cost for linear solver**
  - And for vector operations in nonlinear solvers and time integrators
- **Future: extend all-GPU solvers up solver stack**
  - Nonlinear solvers

Impact

- **Chemistry applications, like combustion have a solve at each vertex**
- **Plasma collision operators have many small solves per vertex [1,2,3]**
- **Continue to provide performant solvers to PETSc users**
- **Support single level domain decompositions solvers (smoothers)**
  - PCPatch in PETSc



Accomplishments

- **Developed early implementations of all-GPU direct and iterative solvers for small systems [3]**

[1] E. Hirvijoki, M.F. Adams, Physics of Plasmas, 24, 3, 2017
[2] M.F. Adams, et. al., SIAM J. Sci. Comp. 39 (6), 2017
[3] M.F. Adams, et. al., Submitted IPDPS 2022

Throughput (solves / second) on one Summit node of linear solver with hybrid asynchronous and batched dispatch

Mixed Precision Accuracy and Speed in SuperLU

Sherry Li

Scientific Achievement

- First sparse direct solver for multi-GPUs that can use single precision LU factorization (for speed), followed by double precision iterative refinement (IR) to recover accuracy
- 30-40% faster than double precision code on 60 GPUs

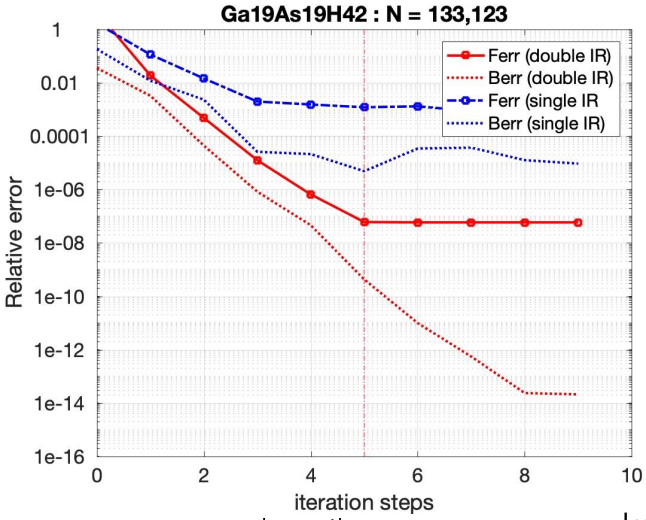
Significance and Impact

In past few years, hardware vendors have started providing faster units for low-precision arithmetic in response to the ML community's demand. Developing new numerical algorithms to harness this power is beneficial to the scientific codes. In addition, using a lower-precision format leads to smaller memory footprint and lower communication volume.

Research Details

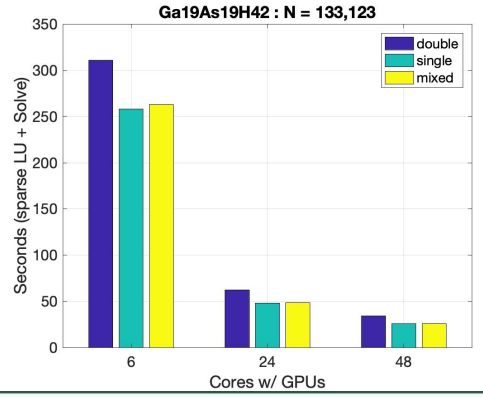
- In Ax=b, user inputs {A,b} and solution x are in single precision
- Perform sparse LU factorization ($O(n^2)$ operation) in single precision, then perform triangular solve and a few steps of iterative refinement ($O(n^{4/3})$ operation) in double precision
- Use equilibration and componentwise scaling, the IR can deliver reliable error bounds, both normwise and componentwise. Solution accuracy guaranteed ~10^{-6}, independent of condition number.

Quantum chemistry: LU size 1.6 billion
Convergence of iterative refinement



$$\text{Berr} := max_i \frac{|b-Ax||_i}{(|A||x||+|b|)_i} \qquad \text{Ferr} := max_i \frac{|x_i - \hat{x}_i|}{|x|_i}$$

Parallel runtime on 8 nodes Summit



X.S. Li, "Accuracy of a sparse direct solver with lower precision factorization on GPUs", SIAM LA 2021

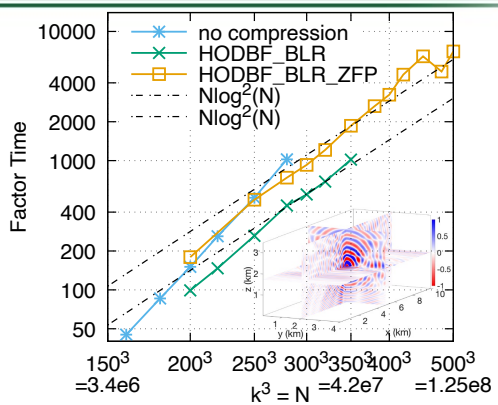
U.S. DEPARTMENT OF ENERGY | Office of Science
ECP EXASCALE COMPUTING PROJECT
FASTMATH
BERKELEY LAB

Butterfly-based Direct Solvers and Preconditioners

Yang Liu, Pieter Ghysels, Lisa Claus, Tianhuan Luo, Sherry Li

Scientific Achievement

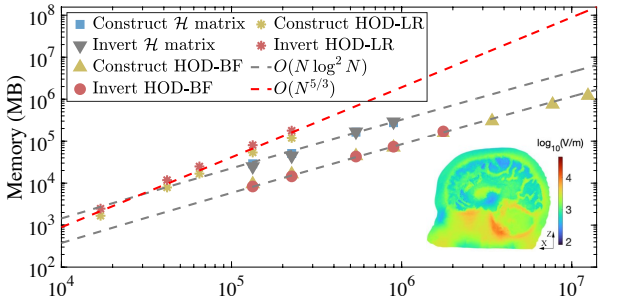
- We developed hybrid algorithms to push the capability of STRUMPACK for high-frequency wave equations from 300^3 to 500^3.
- We developed butterfly-enhanced integral equation (IE) alternatives to STRUMPACK requiring 5x lower mesh density.
- We added support for high-order basis functions for accelerator cavity modeling



HODBF Multifrontal for 3D Helmholtz

Significance and Impact

Direct and preconditioned iterative solutions of high-frequency wave equations are critical components for many ECP and SciDAC applications, including MFEM at LLNL, accelerator modeling at SLAC, EM simulation codes at Sandia. Their fast solutions require leveraging the lately developed numerical linear algebra tool, e.g., butterfly, to significantly reduce the solution time.

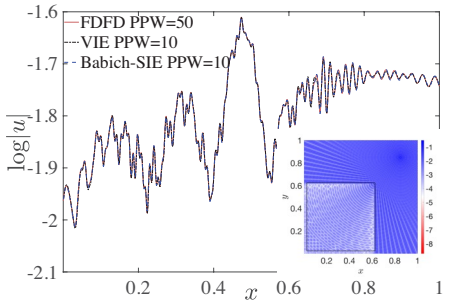


Memory scaling of VIE for biomedical applications

Research Details

- We leveraged hybrid HOD-BF, BLR and ZFP algorithms for large-scale FDFD-discretized wave equations.
- We developed HOD-BF enhanced VIE and Babich ansatz-based SIE formulations free of numerical dispersion.
- We added high-order mesh and basis functions for modeling accelerator cavities.

Y. Liu, P. Ghysels, L. Claus, X.S Li, SIAM. J. Sci. Comput., 2021
S. Sayed, Y. Liu, L. Gomez, A. C. Yucel, IEEE TAP, 2021
Y. Liu, J. Song, R. Burridge, J. Qian, SIAM. J. MMS, submitted



FDFD, VIE and SIE require different points-per-wavelength (PPW)

U.S. DEPARTMENT OF ENERGY | Office of Science
LDRD LABORATORY DIRECTED RESEARCH & DEVELOPMENT
FASTMATH
ECP EXASCALE COMPUTING PROJECT
BERKELEY LAB

Eigen Solvers

GPU Implementation of Eigensolver in MFDn

Chao Yang

Scientific Achievement

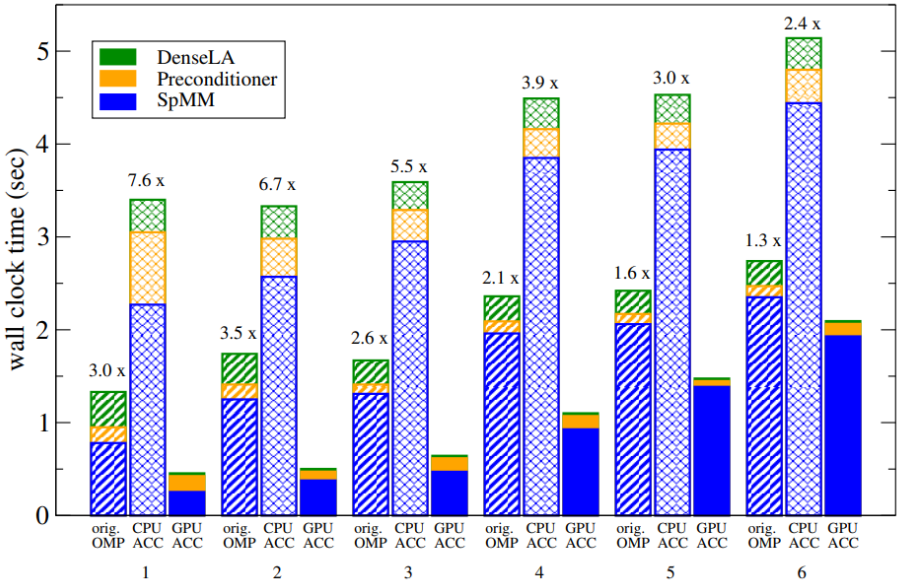
We developed a GPU implementations of eigensolvers for large sparse eigenvalue problems arising from nuclear structure calculations in the MFDn code using OpenACC.

Significance and Impact

Such a solver will enable researchers to study properties of light nuclei isotopes on DOE leadership machines such as the Perlmutter machine at NERSC.

Research Details

- Use OpenACC directives, CuBLAS, CuSolver and CUDA-Aware MPI to port the CPU version of the Lanczos and LOBPCG solver to GPUs
- Use loop fusion, vectorization and atomic updates to improve the performance
- Demonstrate performance improvement on several benchmark problems.



Performance improvement of the GPU implementation of the LOBPCG eigensolver in MFDn over CPU implementation on 6 benchmark problems of dimension between 3M to 122 M.

- B. Cook, P.J. Fasano, P. Maris, C. Yang, D. Oryspayev, *Accelerating quantum many-body configuration interaction with directives,* WACCPD 2021, in press, Lecture Notes in Computer Science.
- P. Maris, C. Yang, D. Oryspayev, B. Cook, *Accelerating an Iterative Eigensolver for Nuclear Structure Configuration Interaction Calculations on GPUs using OpenACC,* Journal of Computational Science, 59, 101554. 2022. https://doi.org/10.1016/j.jocs.2021.101554

Fast Iterative Eigenvalue Solver for Many-body Localization

Chao Yang

Scientific Achievement

Optimized the performance an iterative eigensolver capable of computing interior eigenvalues of Heisenberg spin ½ models with more than 28 spins. (The previous record was 26 spins)

Significance and Impact

Such a solver will enable EFRC researchers to study localization and thermalization properties of quantum materials that depend on the interplay between many-body interaction and disorder.

Research Details

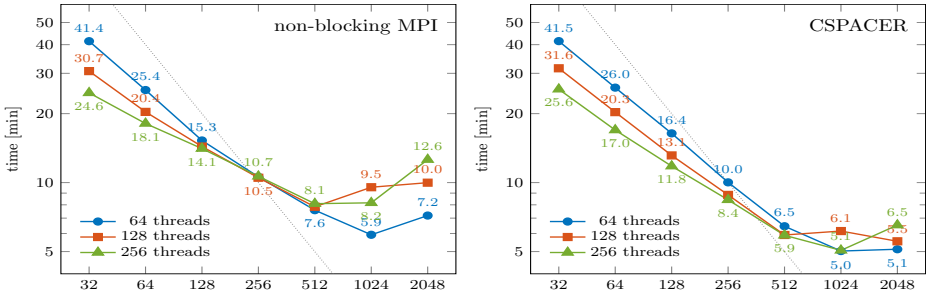
- Use graph partitioning algorithms to reorder the matrix to optimize imbalance
- Use mixed precision to reduce communication volume
- Runtime optimization via Consistent SPACE Runtime
- Scalable up to 1024 KNL Cori nodes: 1 MPI rank and 256 OpenMP threads per node

Spin Model: $H\psi = E\psi$

$$H = \sum_{i=1}^{L-1} \vec{S}_i \cdot \vec{S}_{i+1} - h_i S_i^z$$

$$H = \left[\sum_{i=1}^{L-1} I \otimes \cdots \otimes A_i \otimes I \cdots\right] + D$$

dim: $\binom{L}{L/2}$, e.g. L=32 yields 2.3×10^9



Strong scaling of 1 LOBPCG iteration (26 spins) with 20,000 PCG iterations in single precision.

R. Van Beeumen, K.Z. Ibrahim, G.D. Kahanamoku-Meyer, N.Y. Yao, and C. Yang, "***Enhancing Scalability of a Matrix-Free Eigensolver for Studying Many-Body Localization",*** *International Journal of High Performance Computing Applications, 2022,* https://doi.org/10.1177/10943420211060365

R. Van Beeumen, G. D. Kahanamoku-Meyer, N. Y. Yao and C. Yang, "**A scalable matrix-free iterative eigensolver for studying many-body localization**", HPCAsia2020: Proceedings of the International Conference on High Performance Computing in Asia-Pacific Region, January 2020 Pages 179–187. https://doi.org/10.1145/3368474.3368497

U.S. DEPARTMENT OF ENERGY | Office of Science

FASTMATH

BERKELEY LAB
Lawrence Berkeley National Laboratory

Massively Parallel CG Eigensolvers based on Unconstrained Energy Functionals Methods

Andrew Canning, Mauro Del Ben, Osni Marques

Scientific Achievement

Development of iterative eigensolvers that do not require reorthogonalization of the iterates and lead to better parallel scalability.

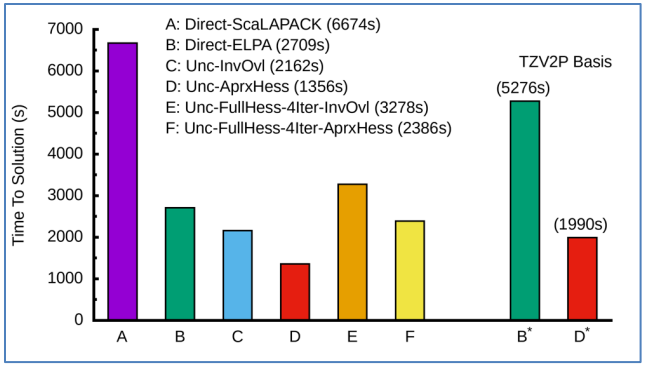
Significance and Impact

This work seeks to improve the performance of electronic structures codes that typically take up to 25% of the workload of NERSC computers.

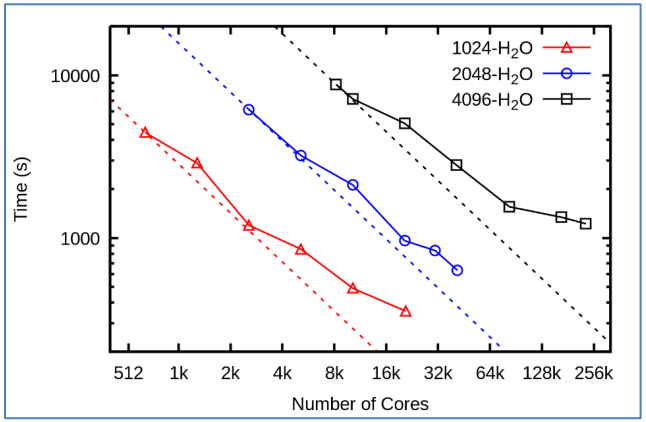
Research Details

- **Constrained (standard) CG:**
  - $\min_{\Psi} \mathrm{Tr}\,[\Psi^T H \Psi]$, $\Psi = [\psi_1, \psi_2, \ldots \psi_N]$, $\Psi^T \Psi = I$
  - Operations on small subspace scale poorly
- **Unconstrained CG method (simplest form)**
  - $\min_{X} \mathrm{Tr}\,[S^{-1} X^T H X]$, $S = X^T X$, $\Psi = X S^{-\frac{1}{2}}$
  - $S^{-1} \approx (2I - S)$ (1st order expansion)
  - No operations on subspace matrix (scales to large core counts)
  - Required the development of novel preconditioners

- *Del Ben, Marques, and Canning, Improved Unconstrained Energy Functional Method for Eigensolvers in Electronic Structure Calculations, ICPP2019, 48th International Conference on Parallel Processing, Kyoto, Japan. Best paper in the Applications Track (100+ submissions to the track, ~25% acceptance rate).*
- *Marques, Del Ben and Canning, Massively Parallel Eigensolvers based on Unconstrained Energy Functionals Methods, SC19 poster. Best research poster finalist (200+ posters submitted, 105 accepted, 5 finalists).*



Time to solution for full SCF convergence compared to direct solvers (ScaLAPACK and ELPA) for various preconditioners developed in this study. B and D* are times obtained with a larger basis (about 1.7 times larger than in B and D, with 160 KNL nodes on cori).*



Strong scaling study: time to solution for bulk liquid water with 1024, 2048 and 4096 molecules.

Quantum Computing Algorithms

Quantum Fourier Transform Revisited

Scientific Achievement

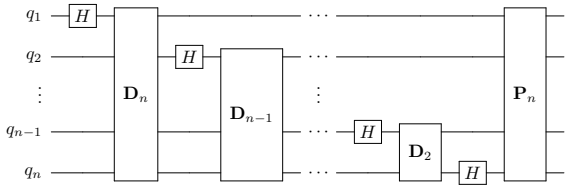
Deriving the quantum Fourier transform (QFT) from the fast Fourier transform (FFT)

Significance and Impact

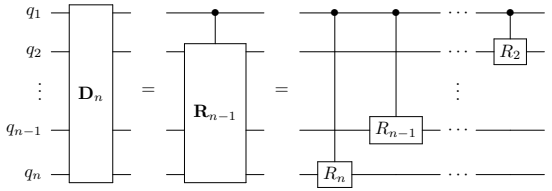
Proves the linear algebra relation between FFT and QFT with little knowledge of quantum computing and by only using elementary properties of Kronecker products of matrices.

Research Details

- FFT algorithm can be derived as a particular matrix decomposition of the discrete Fourier transformation (DFT) matrix
- QFT algorithm can be derived by further decomposing the diagonal factors in the FFT decomposition into products of matrices with Kronecker product structure
- QFT decomposition of the DFT matrix and the corresponding quantum circuit is not unique
- Extended the radix-2 QFT decomposition to a radix-*d* QFT decomposition



Quantum circuit of the DFT matrix decomposition



Quantum circuit representation of the decomposition of the diagonal matrices $\mathbf{D}_n$

D. Camps, R. Van Beeumen, and C. Yang
Quantum Fourier Transform Revisited
https://arxiv.org/abs/2003.03011, 2020.

LDRD

PI: Roel Van Beeumen (LBNL)

QPIXL: Quantum Pixel Representations for Images

M.G. Amankwah, D. Camps, E.W. Bethel, R. Van Beeumen, T. Perciano

Scientific Achievement

Introducing a novel and uniform framework for quantum pixel representations that overarches many of the popular image representations proposed in the recent literature.

Significance and Impact

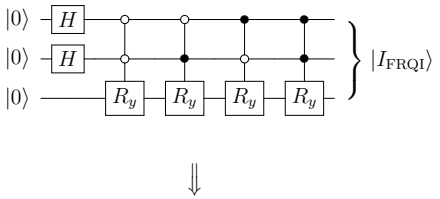
The QPIXL framework provides optimal circuit implementations that significantly reduce the gate complexity and are practical in the NISQ era.

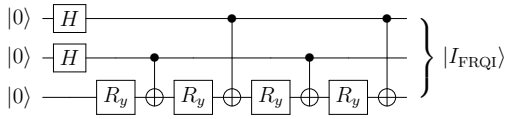
Research Details

- Linear number of gates in terms of the number of pixels
  - Only R_y gates and CNOT gates
  - No need for extra ancilla qubits or multi-controlled gates
- Comprises many of the most popular representations: (I)FRQI, (I)NEQR, MCRQI, (I)NCQI, …
- Efficient circuit and image compression algorithm
- QPIXL++: Quantum Image Pixel Library

M.G. Amankwah, D. Camps, E.W. Bethel, R. Van Beeumen, and T. Perciano
Quantum pixel representations and compression for N-dimensional images
arXiv:2110.04405, 2021.

Original FRQI circuit





QPIXL FRQI circuit

Various compression levels

QPIXL++: https://github.com/QuantumComputingLab/qpixlpp

FunFact: A DSL+Python framework for fast-prototyping tensor decomposition models

Daan Camp, Yu-hang Tang

Scientific Achievement

Developed a Python package that can dramatically simplify the design of custom matrix and tensor factorization models.

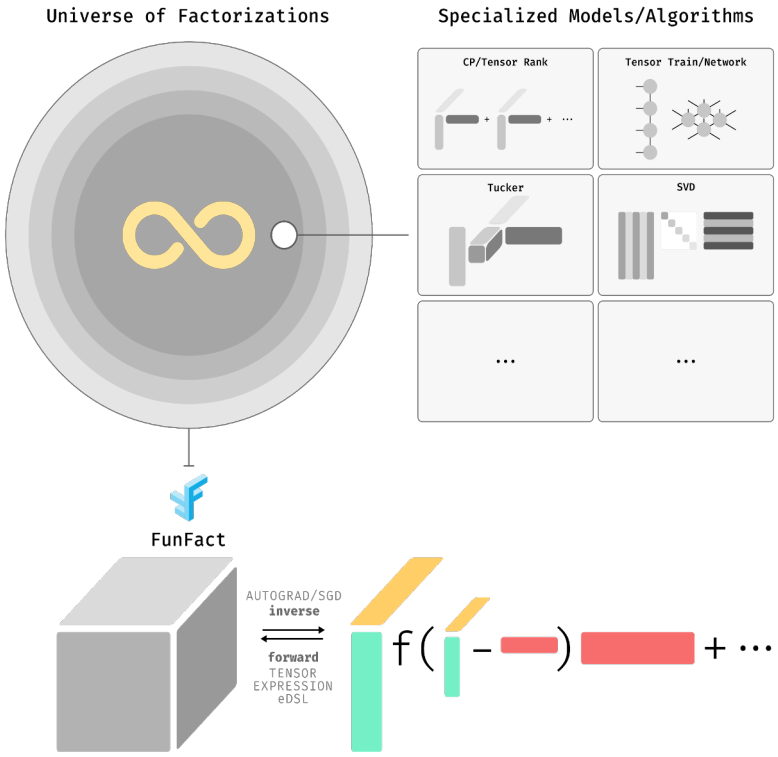
Significance and Impact

Lowered the technically barrier of entrance for the creation of new tensor decomposition models.

Shortened the time-to-algorithm for custom, non-conventional tensor decomposition models from weeks/months to a matter of minutes.

Research Details

- Implemented a powerful programming interface that augments the NumPy API with Einstein notations for writing concise tensor expressions.
- Given an arbitrary forward calculation scheme, automatically solve the corresponding inverse problem using stochastic gradient descent, automatic differentiation, and multi-replica optimization.
- Application areas include tensor decomposition, quantum circuit synthesis, and neural network compression.
- GPU- and parallelization-ready.



FunFact allows users to explore the vast universe of tensor decomposition models that consists of deeply nested structures and generalized contraction operators, etc.

Camps, **Tang***, *Manuscript under review.*

Work was performed at Lawrence Berkeley National Laboratory

AI/ML Methods

Deep Learning and Spectral Embedding for Graph Partitioning

Alice Gatti, Pieter Ghysels

Scientific Achievement

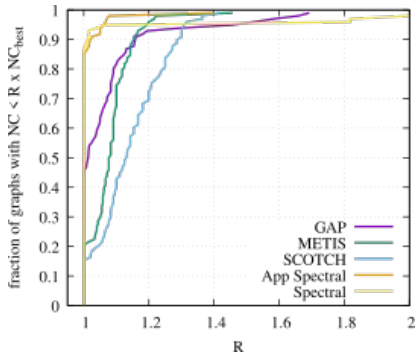
We developed a novel deep learning graph partitioning algorithm. Graph partitioning is a common problem in scientific computing and is NP complete. Heuristic approximations are available but are increasingly becoming bottleneck operations in large scale simulations.

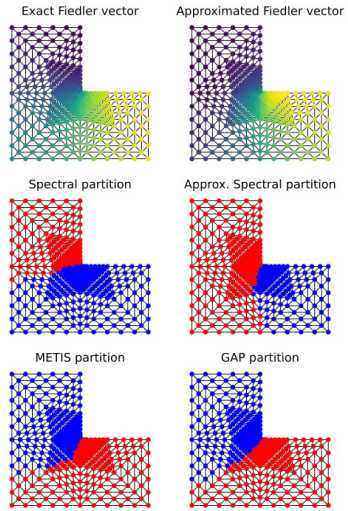
Significance and Impact

We propose a partitioning algorithm based on a multilevel graph convolutional deep neural network that runs entirely on GPU, effectively utilizing modern high performance computing hardware, and resulting in high quality graph partitions.

Research Details

- An embedding module, using a multilevel graph convolutional network inspired by multigrid, produces an approximate spectral embedding. The loss function is based on the eigenvector residual.
- The partitioning module outputs partition probabilities given the approximate spectral embedding. The loss function corresponds to the expected value of the normalized cut.
- We train on small (< 5K nodes) Delaunay graphs, FEM graphs, and a variety of problems in the SuiteSparse Collection. The trained algorithm generalizes extremely well to much larger graphs (up to 10M nodes).





Left: Performance profile for different partitioning methods. Our novel method GAP (Generalizable Approximate Partitioning) performs, in a large fraction of cases, better than either METIS or SCOTCH (state-of-the-art partitioners), while being much faster than spectral partitioning. Right: Comparison of several embeddings (top row) and resulting partitions for a simple planar graph from a finite element discretization: the Fiedler vector is produced by a spectral embedding while the approx. Fiedler vector, the output from our multilevel graph convolution neural network embedding module, is a very good approximation, and is much cheaper to compute.

A. Gatti, Z. Hu, T. Smidt, E.G. Ng, P. Ghysels. Proceedings of the 2022 SIAM Conference on Parallel Processing for Scientific Computing. Pp. 25-36.
DOI: https://doi.org/10.1137/1.9781611977141.3

U.S. DEPARTMENT OF ENERGY | Office of Science
BERKELEY LAB

GPTune autotuner: Bayesian optimization with Gaussian Process surrogate modeling

Younghyun Cho, Jim Demmel, Yang Liu, Henrui Luo, Sherry Li

Scientific Achievement

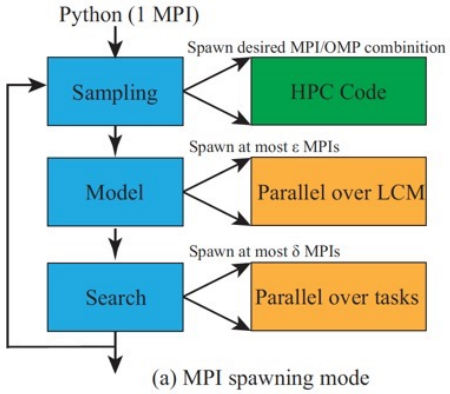
- **Optimization :** $\min_x y(t,x)$, x : parameter configuration
- **Applicable to any black-box software**

Significance and Impact

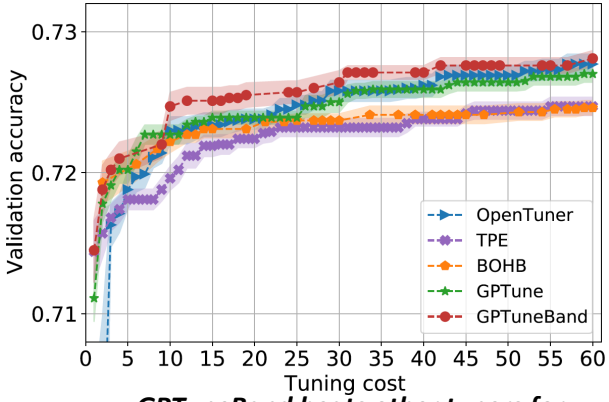
Gaussian process (GP) models can act as surrogates for code performance or first-principle physics for many expensive SciDAC and ECP applications. Our work leverages multi-task and multi-fidelity GP models to allow accurate surrogates.

Research Details

- Features: multi-task, multi-objective, and multi-fidelity
- Added multi-objective tuning features to allow memory/time tradeoff
- Supported multi-task and transfer learning features to leverage correlation between tuning tasks to improve model accuracy
- History database for crowd-tuning
- GPTune has been applied to Hypre, MFEM, STRUMPACK, SuperLU_DIST, PLASMA, SLATE, ScaLAPACK, NIMROD, M3D-C1, IMPACT-Z, CNN, GCN, kernel ridge regression, sketching-based linear square solvers.



(a) MPI spawning mode



GPTuneBand beats other tuners for tuning GCN on the Citeseer dataset

Y. Cho, J. W. Demmel, X. S. Li, Y. Liu, and H. Luo, IEEE MCSoC, 2021
X. Zhu, Y. Liu, P. Ghysels, D. Bindal, and X. S. Li, SIAM PP, 2022
H. Luo, J.W. Demmel, Y. Cho, X. S. Li, and Y. Liu, JMLR, submitted

U.S. DEPARTMENT OF ENERGY | Office of Science | FASTMATH | ECP EXASCALE COMPUTING PROJECT | BERKELEY LAB