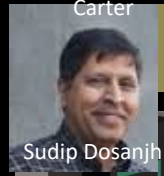
Overview of R&D Activities in Scalable Solvers Group

SIMBER Kick-off --- December 2, 2021

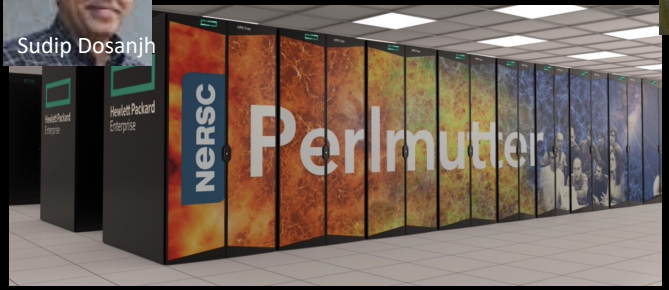
Sherry Li, Group Leader
Scalable Solvers Group
Applied Mathematics and Computational Research Division
Lawrence Berkeley National Laboratory

Computing Sciences at Berkeley Lab


Jonathan Carter

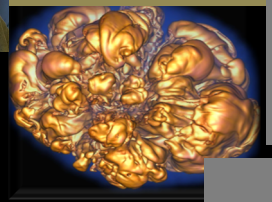

Sudip Dosanjh

NERSC



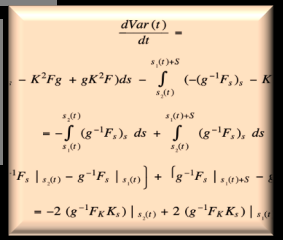

David Brown

Applied Math and Computational Research



Computational Science

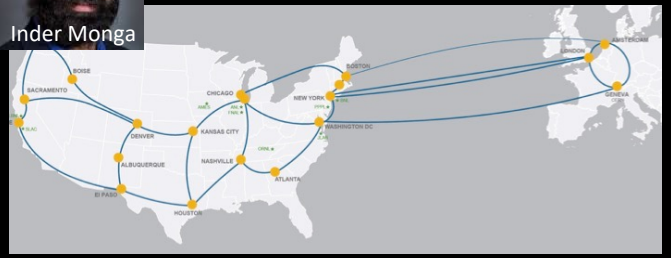
Applied Mathematics

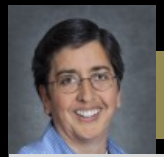


Computer Science


Inder Monga

ESnet

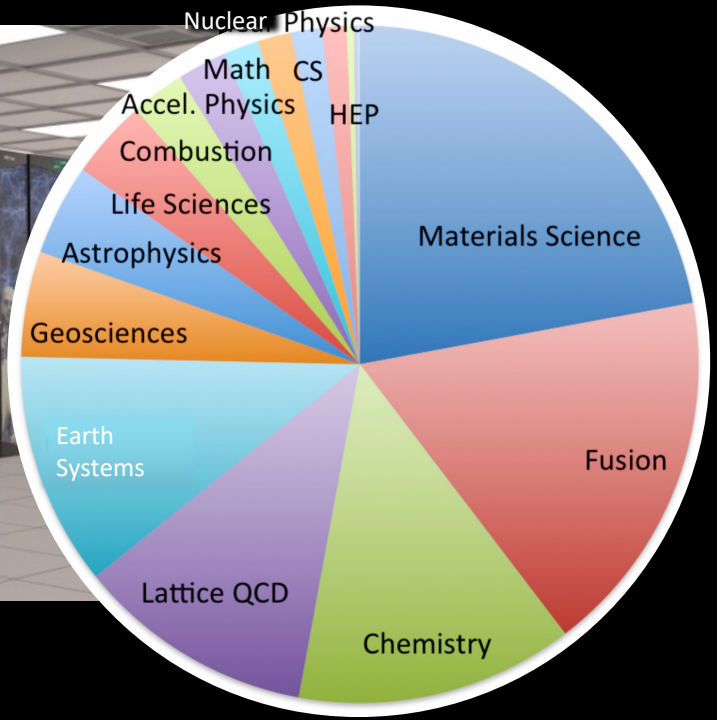



Deb Agarwal

Scientific Data

Data Science & Technology

NERSC: The most widely-used facility in DOE



NERSC is dedicated to the DOE Office of Science

- Thousands of users, 2000 publications/year, 700 codes
- Long history of data-intensive as well as compute-intensive science

Scalable Solvers Group

https://crd.lbl.gov/departments/applied-mathematics/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: GPU, KNL, …

Expertise, capabilities

(Most software packages are open source, use 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
- **High-precision floating-point arithmetic** (QD, ARPREC, XBLAS)
- **High-order PDE solvers, parallel-in-time PDE solvers** (PFASST)
- **Machine learning for sciences** (GPTune, DRL-Graph-Partition)
- **Quantum computing algorithms** (QFT)

Linear Solvers

Batched all GPU controlled solvers for many small systems in PETSc

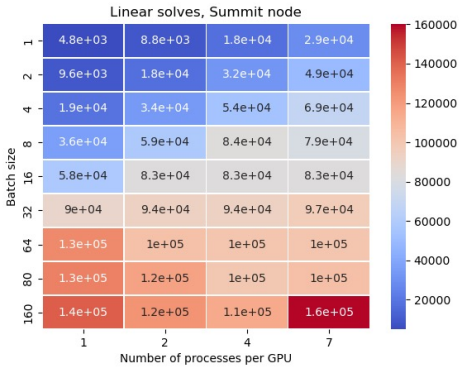
Mark F. Adams

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

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



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

SuperLU: Sparse Triangular Solve on multi-GPUs

Nan Ding, Yang Liu, Sam Williams, Sherry Li

Scientific Achievement

- First multi-node, multi-GPU implementation of a supernodal sparse triangular solve; good performance on up to 12 GPUs.
- Up to 6.1x faster than NVIDIA cuSPARSE SpTRSV

Significance and Impact

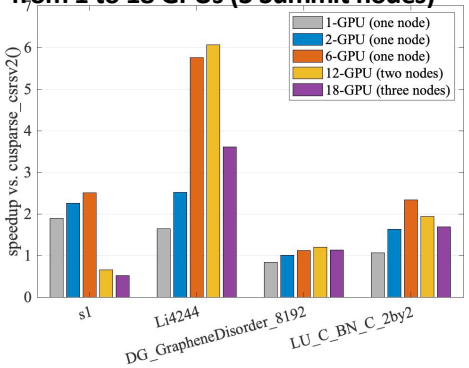
SuperLU preconditioners are essential for several ECP and SciDAC application codes. The preconditioned iterative solver's performance is dominated by triangular solve. Our multi-GPU implementation of SpTRSV enables the application codes to harness the GPU power.

Research Details

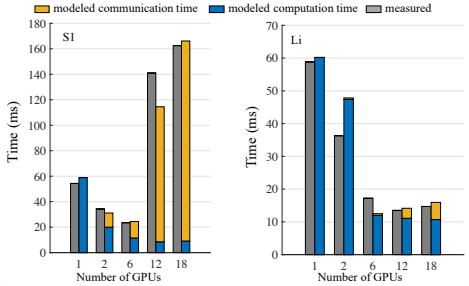
- Leverage the advantage of GPU-initiated data transfers using NVSHMEM (stems from OpenSHMEM standard)
- Use producer–consumer paradigm to manage the computation and synchronization using two CUDA streams
- Developed an accurate performance model based on the critical path in the computational DAG, taking into account flops, memory bandwidth and inter-process communication

N, Ding, Y. Liu, S. Williams, X.S. Li, "A Message-Driven, Multi-GPU Parallel Sparse Triangular Solver", SIAM ACDA 2021 Proceedings.

Multi-GPU lower triangular solve speedup over *cusparse_csrsv2()*. Up to 6.1x speedup from 1 to 18 GPUs (3 Summit nodes)



Performance model accurately captures the scaling trend of multi-GPU SpTRSV.



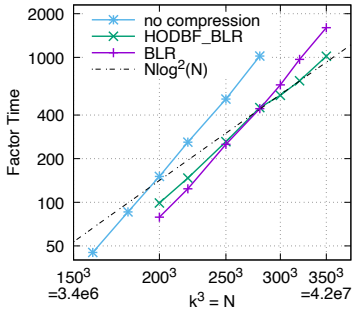
U.S. DEPARTMENT OF ENERGY | Office of Science
ECP EXASCALE COMPUTING PROJECT
FASTMATH
BERKELEY LAB

STRUMPACK: Next-generation fast sparse solver and preconditioner

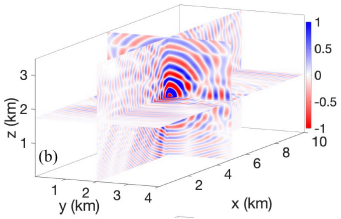
Pieter Ghysels, et al.

Scientific Achievement

STRUMPACK is a modern factorization based sparse solver that overcomes the bottlenecks in traditional sparse direct solvers by relying on rank-structured matrix approximations. Based on sparse LU factorization, the preconditioners in STRUMPACK are robust for a wide range of numerical problems, including highly ill-conditioned and indefinite linear systems.



Factor time of our HOD-BF + BLR multi-frontal solver on 64 Cori Haswell nodes.



Pressure field in the Marmousi2 velocity model on a 190x216x516 grid computed by our solver

Significance and Impact

Modeling of complex multiscale and multiphysics problems often leads to large and numerically challenging linear systems. The STRUMPACK solver can be run as a black-box exact solver, or as a robust, efficient and scalable preconditioner, with a minimum number of tuning parameters.

Research Details

- STRUMPACK's direct solver supports both Nvidia and AMD GPUs through CUDA and HIP. Intel GPU support is experimental, using SYCL/DPC++.
- The preconditioners are based on approximate multifrontal LU factorization using several rank-structured matrix compression formats: Block Low Rank, Hierarchically Off-Diagonal Low Rank, Hierarchically Semi-Separable, Butterfly, …

Ghysels, P. and R. Synk (2020). "High performance sparse multifrontal solvers on modern GPUs". Submitted.
Liu, Y., P. Ghysels, L. Claus, and X.S. Li (2020). "Sparse approximate multifrontal factorization with butterfly compression for high frequency wave equations". arxiv.org/abs/2007.00202

Eigen Solvers

Eigenvalue Calculation via Spectrum Slicing

Chao Yang, David Williams-Young

Scientific Achievement

Developed parallel implementation of shift-invert and polynomial filtering-based spectrum slicing (SS) algorithm for computing eigenvalues of symmetric matrices

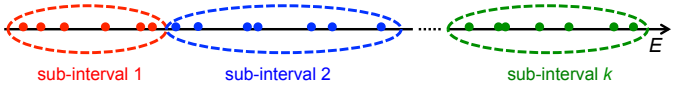
Significance and Impact

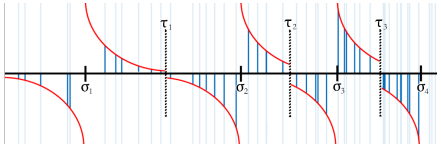
Enable large-scale Kohn-Sham density functional theory based electronic structure calculation for materials design

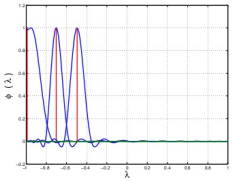
Research Details

- Develop effective spectrum partition schemes based on density of state estimation and K-means clustering
- Develop validation schemes that require minimal communication
- Develop effective load balancing scheme
- Integration with PARSEC and DGDFT

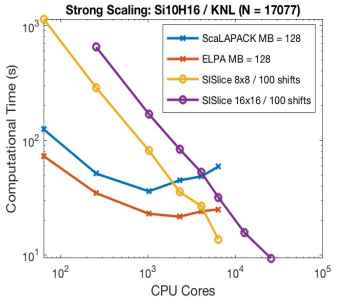
- D. Williams-Young, P. Beckman and C. Yang, "A Shift Selection Strategy for Parallel Shift-Invert Spectrum Slicing in Symmetric Self-Consistent Eigenvalue Computation", submitted 2019.
- K. Liou, C. Yang and J. Cheliokowsky, "Scalable Implementation of Polynomial Filering for Density Functional Theory Calculation in PARSEC", submitted, 2020



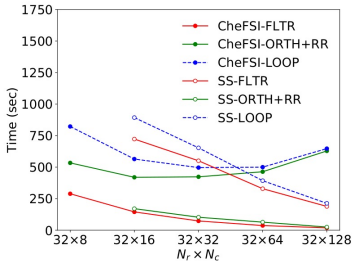




Strong scaling



Shift-invert SS



Polynomial filtering SS

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.

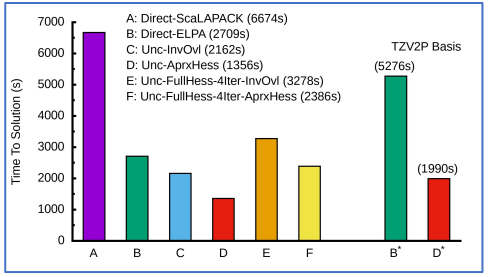
Significance and Impact

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

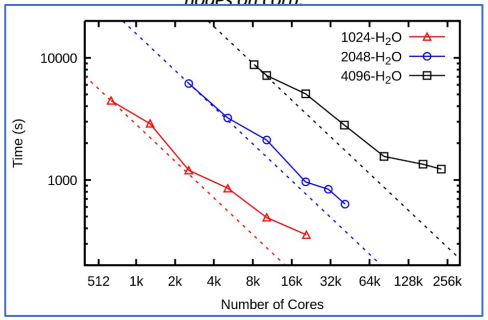
Research Details

- 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)
  - $\min_{\mathrm{X}} \mathrm{Tr}\,[S^{-1}\mathrm{X}^T H \mathrm{X}],\ S = \mathrm{X}^T\mathrm{X},\ \Psi = \mathrm{X}S^{-\frac{1}{2}}$
  - $S^{-1} \approx (2I - S)$ (1^{st} 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.

U.S. DEPARTMENT OF ENERGY | Office of Science

FASTMATH

Quantum Computing Algorithms

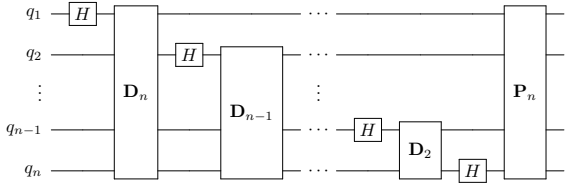
Quantum Fourier Transform Revisited

Scientific Achievement

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

Significance and Impact

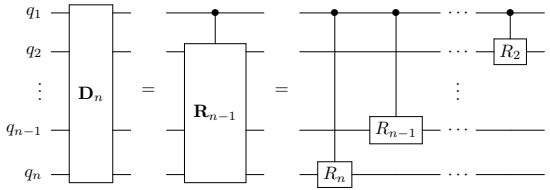
Proves the linear algebra relation between FFT and QFT with little knowledge of quantum computing and by only using elementary properties of Kronecker products of matrices.



Quantum circuit of the DFT matrix decomposition

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 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.

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

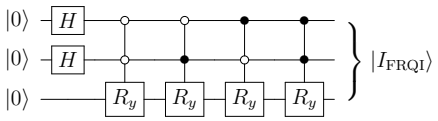
The QPIXL framework provides optimal circuit implementations that significantly reduce the gate complexity and are practical in the NISQ era.

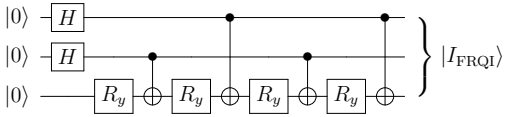
Research Details

- 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

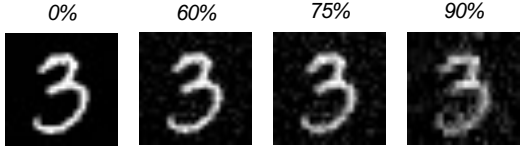
M.G. Amankwah, D. Camps, E.W. Bethel, R. Van Beeumen, and T. Perciano
Quantum pixel representations and compression for N-dimensional images
arXiv:2110.04405, *2021.*

Original FRQI circuit





QPIXL FRQI circuit



Various compression levels

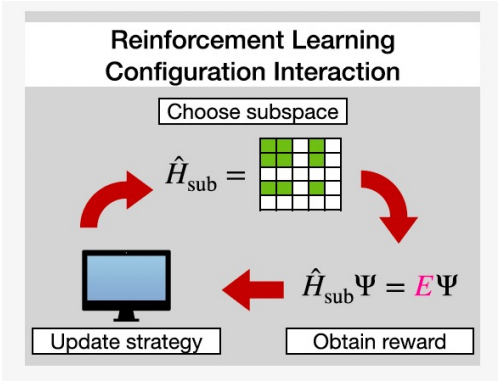
AI/ML Methods

Solving k-sparse eigenvalue problem using reinforcement learning

Scientific Achievement

Developed a reinforcement learning based algorithm to solve a k-sparse eigenvalue problem

$$\min_{\|x\|_0 \le k} \frac{x^T A x}{x^T x}$$



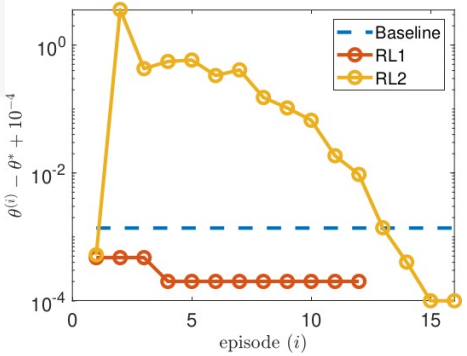
Significance and Impact

The use of RL allows us to improve existing selected configuration interaction (sCI) method for solving many-body eigenvalue problems in which the eigenvector of interest has localization properties.

Research Details

- Use linear feature based Q-learning approach to ultimately rank selected rows/columns of A
- Use an active space based search policy to choose actions (removing and adding row/column)

- L. Zhou, L. Yan, M. A. Caprio and C. Yang, "Solving k-sparse eigenvalue problems with reinforcement learning", to appear CSIAM-AM, 2021
- J. Goings, H. Hu, C. Yang and X. Li, "Reinforcement Learning Configuration Interaction", submitted, 2021



Eigenvalue convergence with respect to training episode numbers

GPTune autotuner: Bayesian optimization with Gaussian Process surrogate modeling

Yang Liu, Sherry Li

Scientific Achievement

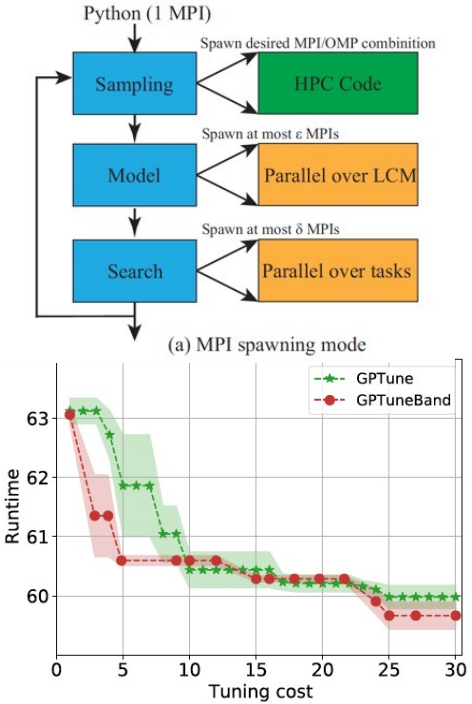
- **Parameter optimization :** $\min_x y(t,x)$, x : parameter configuration
- **Applied to applications: NIMROD (10% speedup), M3D-C1 (15% speedup), MFEM (1.7x speedup) and accelerator cavity modeling.**

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
- Integrated GP surrogate to replace expensive simulation for RF cavity resonance detection



Tuning history of single-fidelity (GPTune) and multi-fidelity (GPTuneBand) tuning of NIMROD runtime with 4 tuning parameters

Y. Liu, W. M. Sid-Lakhdar, O. Marques, X. Zhu, C. Meng, J. W. Demmel, and X. S. Li. GPTune: multitask learning for autotuning exascale applications, PPoPP21, 2021

History of collaboration with Simula

Simula: Johannes Langguth

LBNL: Ariful Azad, Aydin Buluc, Pieter Ghysels, Sherry Li

"A distributed-memory algorithm for computing a heavy-weight perfect matching on bipartite graphs" (HWPM)

Significance:

- A highly scalable parallel algorithm for numerical pivoting in sparse LU factorization
- Available through CombBLAS (https://github.com/PASSIONLab/CombBLAS)
- Used in STRUMPACK and SuperLU