CASC Newsletter

Volume 16 | December 2025

IN THIS ISSUE:

From the Director

Contact: Jeff Hittinger

"There is nothing so stable as change." – ***Bob Dylan***

Change is a constant in our lives. Research is an engine of change, and CASC researchers are drivers of change. In this edition of the CASC newsletter, we highlight four efforts led by CASC researchers that are making positive changes in areas of importance to the Lab. From new algorithms and solvers in contact mechanics and quantum control, to new approaches to performance optimization, to new data-driven approaches to physical modeling in carbon capture, CASC researchers continue to reshape the possibilities of high performance computing (HPC), as they have done for nearly 30 years.

Some change is bittersweet, however. As I have taken on the role of the Computing Principal Directorate's Deputy Associate Director for Science & Technology, this CASC newsletter will be my final opportunity to directly share with you the great achievements of this world-class organization. It has been an honor to serve as the CASC Director for nearly seven years, and it saddens me to be leaving this amazing division. Periodic changes in leadership, however, can re-energize an organization, and I am confident that CASC's new

Director, Kathryn Mohror, will be an agent of change that leads CASC to new heights of success. Excelsior!

Collaborations | Scalable Methods and Solvers for Contact Mechanics

Contact: Cosmin Petra

Contact mechanics research is important because it lies at the heart of how real-world surfaces touch, transmit forces, deform, wear, and fail. State-of-the-art computational techniques for engineering contact mechanics problems have serious limitations, such as lack of scalability, lack of robustness, and low accuracy, when applied to large-scale systems with complex contact nonlinearities. Contact analyses appear in a large number of engineering applications: complex container assemblies, pressure vessels, weapon assemblies, additively manufactured padding materials, and others. The underlying mathematical problems consist of nonsmooth, highly nonlinear, and poorly conditioned systems of equations.

As current state-of-the-art, for example, LLNL's Diablo library has developed several strategies for simulating contact. However, to ensure robustness and to compute accurate Lagrange multipliers, Diablo relies on parallel direct solvers, which have poor scalability and, hence, limit the resolution and extent of contact simulations.

The development of computationally scalable methodologies for contact problems is currently the focus of a multi-directorate team involving CASC researchers Cosmin Petra, Tzanio Kolev, Socratis Petrides, Jingyi Wang, and Tucker Hartland alongside Engineering colleagues Mike Puso, Jerome Solberg, Eric Chin, and Michael Tupek, under an LDRD-ER project. The team has developed a novel homotopy Newton continuation algorithm [1] for contact mechanics that employs a combination of trust-region, filters, and wide neighborhoods of the homotopy path to efficiently and robustly drive the outer nonlinear loop.

Newton continuation linearization systems, however, are extremely challenging, mainly due to ill-conditioning specific to homotopy methods. Efficient algebraic multigrid (AMG) techniques for such systems are also part of the focus of this LDRD project. For interior-point (IP) homotopy [2], the team has successfully deployed a specially designed AMG technique developed by Petrides, Kolev, and Hartland. Based on a two-level approach, where AMG is employed as a smoother and a direct solver is used on the subspace defined by the degrees of freedom involved in the contact surface, the new IP-AMG preconditioner

effectively remedies the ill-conditioning and achieves scalable performance for a wide range of real-world problems, as shown in Figure 1.

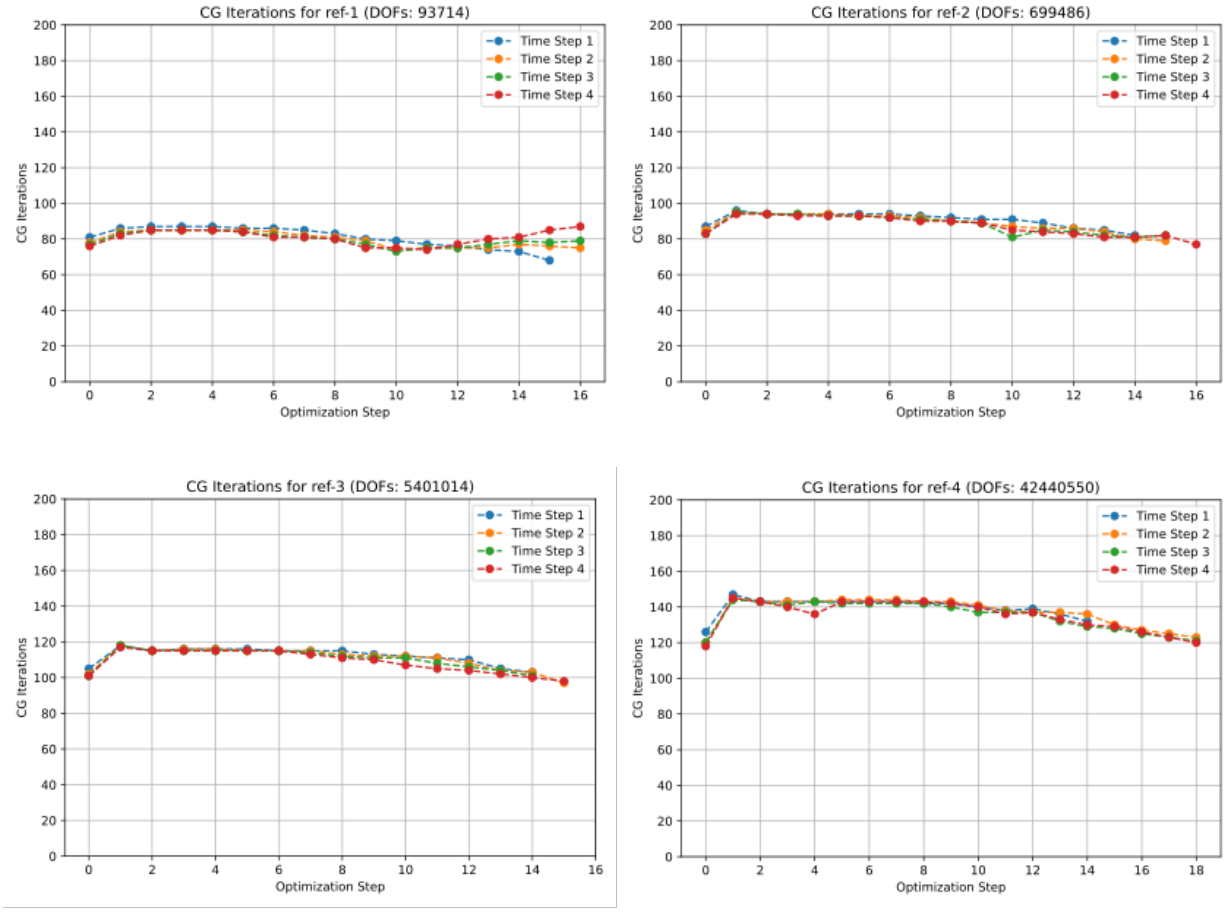

Figure 1: The above plots illustrate the number of conjugate gradient (CG) iterations for each time step across various levels of mesh refinements for a contact problem simulation. The problem involves two bodies, where a smaller block is pushed against a larger block.

The team assembled a scalable contact solver stack that uses MFEM for discretization, *hypre* for AMG operations, Tribol for mortar contact mechanics computations, and a customized IP solver. A second software instantiation of the MFEM-based IP solver is available wherein Diablo is used for both finite element discretization and contact mechanics computations. Two example problems are shown in Figure 2. In the contact simulations, the number of IP outer loop iterations and AMG preconditioned conjugate gradient inner linear solver iterations are small and remain small upon mesh refinement.

Thus far, the simulations have been run with tens of millions of degrees of freedom and thousands of MPI processes.

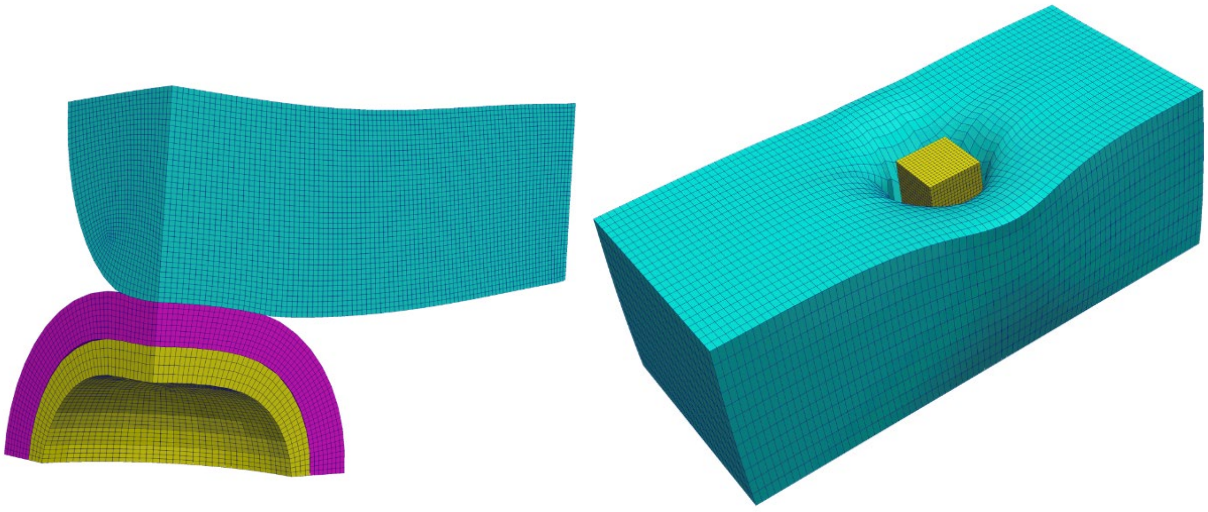

Figure 2: Examples of two contact analysis simulations performed with the MFEM-Tribol stack. The left picture simulates a smaller block being pushed against and then slid along the larger block, which is also softer. The picture on the right shows the simulation of an elongated beam pressing on two concentric spherical shells (three body contact).

Design optimization of structures driven by contact mechanics simulations is another research thrust of the LDRD-ER. Examples of such optimization problems are the design of high current joint shown in Figure 3 and of a Marman clamp problem shown in Figure 4. A significant collaborative effort involving CASC and Engineering researchers Wang, Solberg, Puso, and Petra developed accurate mathematical models for these two problems and investigated efficient numerical solution techniques. The salient challenge in tackling optimization problems driven by contact analyses arises from the lack of smoothness of design objective(s) caused by changing contact regions throughout optimization iterations.

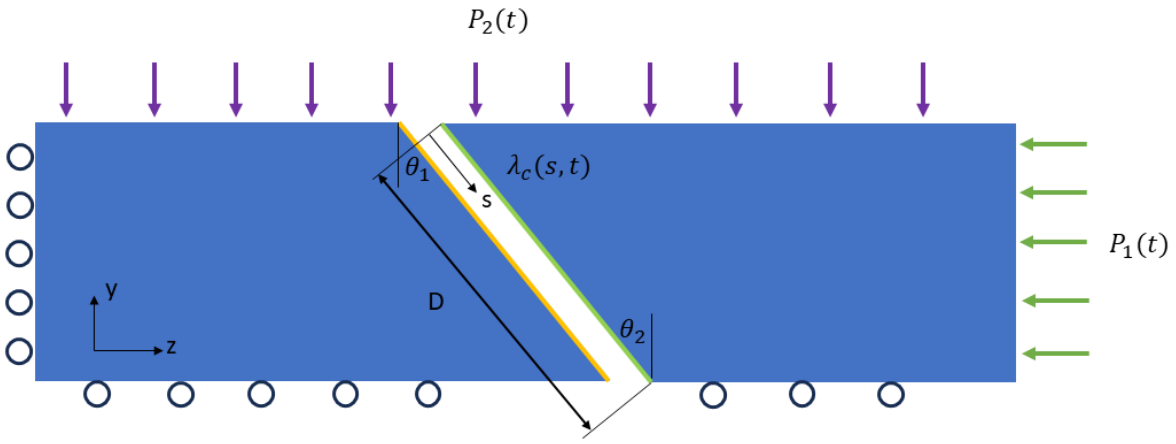

Figure 3: 2D illustration of a high current joint contact design problem. Two time-dependent loads are considered. The design variables are the two angles of the wedges and the maximum preload value P_1. The boundary conditions are displayed in the figure. The problem minimizes the maximum preload while maintaining sufficient stress towards the top of the contact region, via constraints.

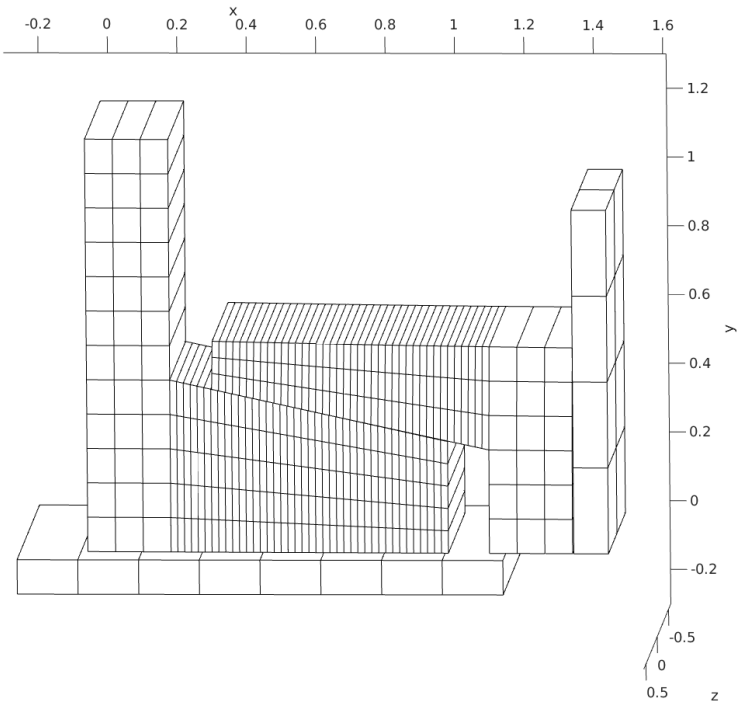

Figure 4: An example of the deformed clamp problem consisting of a flange (left), a retainer (middle), a band (on the right). The bottom plate is used to constrain the bottom surface of the flange via contact, due to the symmetry of the clamp system with respect to the y-plane. An initial displacement of the band is used to induce contact between the retainer and the flange. We aim to minimize the compliance while requiring enough pressure on the bottom-left of the flange. The shape of the contact surfaces between the flange and the retainer is optimized, and the mesh around potential contact region is refined.

The team has shown that gradient-based sensitivities can be effectively used upon introduction of additional stabilization in the problem formulation. For example, traditional derivative-based optimizers can effectively find accurate optimal contact design parameters for both high current joint and clamp problems, as shown in Figures 5 and 6.

The team is currently investigating and developing derivative-free numerical optimization methods as a more general and stable approach for nonsmooth problems. A preliminary constrained Bayesian optimization algorithm that uses Gaussian process surrogate models and special purpose merit functions for the constraints showed great potential when applied to the high current joint and clamp problems [3], with performance similar to gradient-based approaches. Ongoing efforts aim to further refine numerical methods and modelling for optimization with contact analyses [4], with the goal of identifying general and robust numerical approaches to be incorporated in the Lab's efforts in design optimization, such as the Livermore Design Optimization (LiDO) framework.

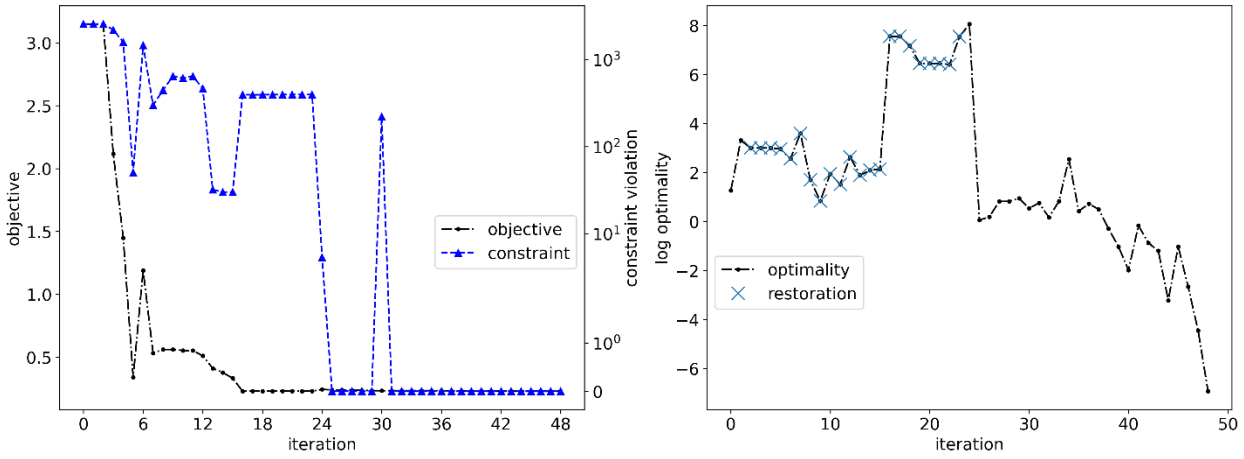

Figure 5: Sensitivity-based optimization history results for the high current joint contact design problem. The objective value (left y-axis) and constraint violation (right y-axis) are shown on the left. Ipopt optimality history is shown on the right.

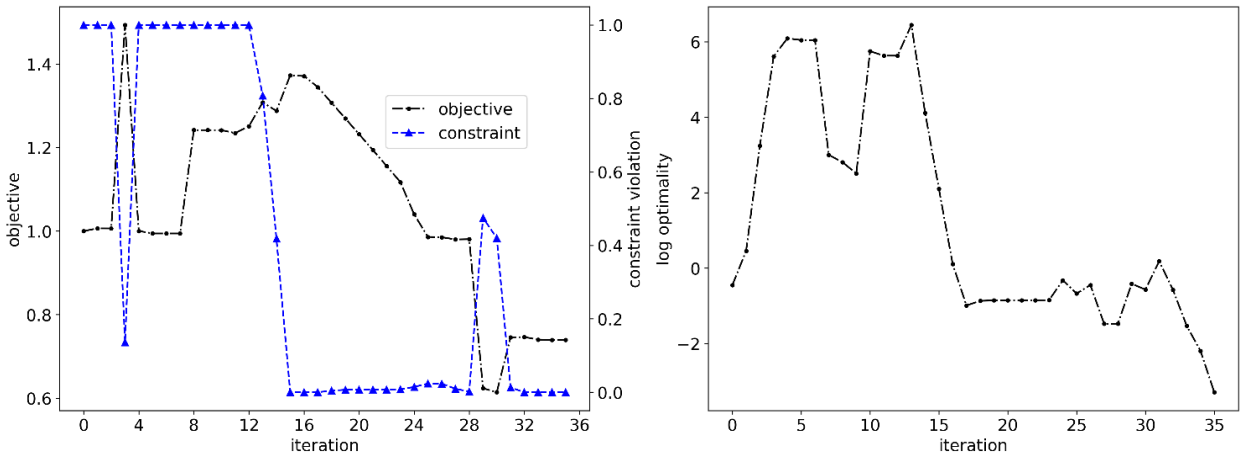

Figure 6: Same as Figure 5 but for the Marman clamp problem.

[1] C. Petra, N. Chiang, J. Wang, T. Hartland, E. Chin, and M. Puso. "A filter trust-region Newton continuation method for nonlinear complementarity problems." Submitted to *Optimization Methods and Software*, 2024.

[2] T. Hartland, C. Petra, N. Petra, and J. Wang. "A scalable interior-point Gauss-Newton method for PDE-constrained optimization with bound constraints." arXiv, 2024.

[3] J. Wang, J. Solberg, M. Puso, E. Chin, and C. Petra. "Design optimization in unilateral contact using pressure constraints and Bayesian optimization." arXiv preprint: 2405.03081, 2024.

[4] J. Wang, N. Chiang, C. Petra, and J. Peterson. "Constrained Bayesian optimization with merit functions." arXiv preprint: 2403.13140, 2024.

Lab Impact | Extending the Boundaries of Performance Optimization for HPC GPU Codes

Contact: Konstantinos Parasyris and Giorgis Georgakoudis

Optimizing scientific software is a substantial challenge. Computational scientists must leverage domain expertise to write their software while navigating the multifaceted complexities of making that software run fast. In today's heterogeneous HPC ecosystem—composed of diverse architectures and sophisticated compiler, runtime, and programming language environments—applications are prone to inflated execution times due to hard-to-find "performance defects."

Deriving actionable insight to improve performance still requires substantial human expertise. Existing performance analysis tools report hotspots, where software spends most time. The process of analyzing profiling data is time-consuming and does not scale; doing performance optimization and re-evaluating changes through those existing methods are non-scalable too. Domain experts must execute their large applications multiple times to evaluate the effectiveness of their manual optimizations and to assess if it helps performance, making the process labor-intensive and inefficient.

For example, transcribing a non-portable CUDA implementation of the HPGMG mini-application to the portable OpenMP offloading model showed that the performance of the initial OpenMP implementation was 12x slower on Cori-GPU (NERSC) and 5.7x slower on Summit (ORNL) compared to the original CUDA version. Another example—manifesting on the Seismic Waves, fourth order algorithm (SW4CK)—relates to GPU occupancy and register allocation schemes. The compiler tries to maximize register usage in GPU-kernels, but in some cases that hurts GPU occupancy resulting in poor execution time. In both cases, developers waste significant time reverse engineering the software stack to detect/fix defects. Once the root cause is discovered, simple code modifications can significantly improve performance.

CASC researchers Konstantinos Parasyris and Giorgis Georgakoudis provide software developers with new state-of-the-art performance analysis and optimization tools that cut through the complexities of heterogeneous hardware and complex software stacks. Those tools must be fast and easy-to-use and integrate with HPC applications.

Such a tool is Mneme [1], which leverages a "Record-Replay" capability to dissect application execution into smaller, analyzable pieces—called code snippets as presented in Figure 7. Those snippets optimized independently without the resource and time demands of running the entire application. Mneme composes with AI-driven search algorithms, such as Bayesian optimization, to explore variations of those code snippets to minimize their execution time. Once the optimization process completes, Mneme suggests code variations to the developer for improving overall application performance. For instance, Mneme determined launch parameters for GPU kernels, achieving an end-to-end speedup of up to 1.53x for an application, while performing this analysis orders of magnitude faster than traditional approaches.

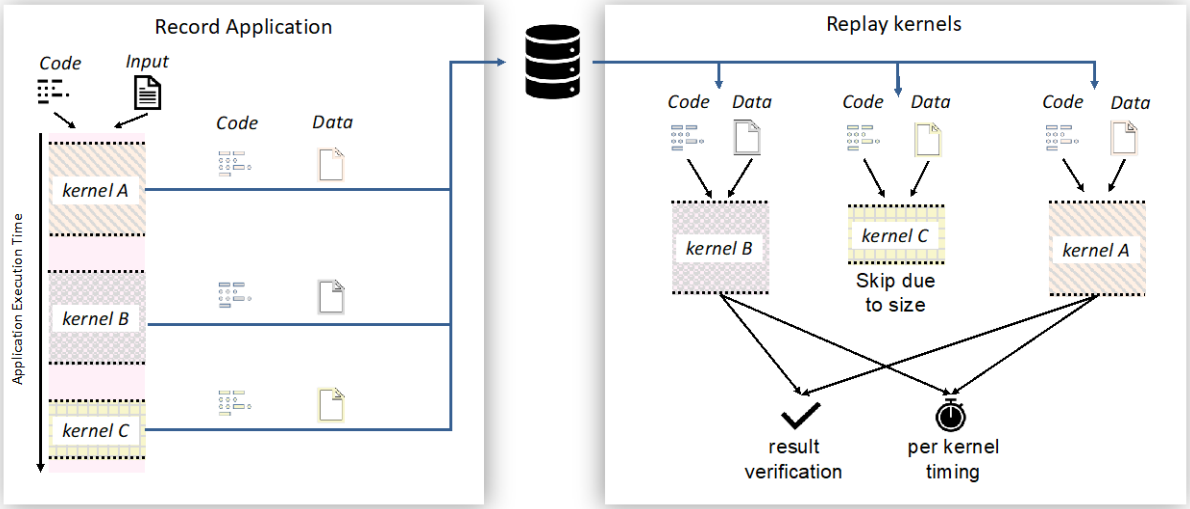

Figure 7: Left: Mneme records invocations and the state of the GPU hardware. Right: Mneme can execute (replay) the individual kernels as independent binaries and optimize them.

The current paradigm in HPC software development is to write applications in a high performance programming language, like C/C++ and Fortran, and then hand its source code over to an optimizing compiler to translate it to machine instructions. This process is ahead-of-time (AOT) compilation, meaning the code is compiled prior to its execution. Optimizing compilers automatically analyze code to apply optimizing transformations to it, but their capability is limited by the code information available during AOT. For example, actual values of program variables, such as tensor sizes or number of particles in a simulation, are unknown until executing the application.

If runtime information were available during compilation, an optimizing compiler would generate faster code. Based on this hypothesis, they developed Proteus [2], a tool that compiles just-in-time (JIT)—meaning, at the point of execution—and includes runtime information to generate faster code than typical AOT compilation. Proteus offers a developer-friendly interface in C/C++ to empower application developers to annotate functions for JIT compilation and specify which function parameters to optimize for. Leveraging those straightforward user annotations, Proteus automatically extracts code and runtime information to JIT optimize code, powered by LLVM, while reducing JIT compilation overheads through sophisticated caching techniques.

Figure 8 shows an evaluation of Proteus on the Tioga cluster with AMD MI250X GPUs, a precursor to El Capitan, showing the end-to-end speedup of several HPC kernels over

typical AOT compilation. Proteus demonstrates significant speedup, up to 2.8x, thanks to reducing computation instructions and improving register utilization.

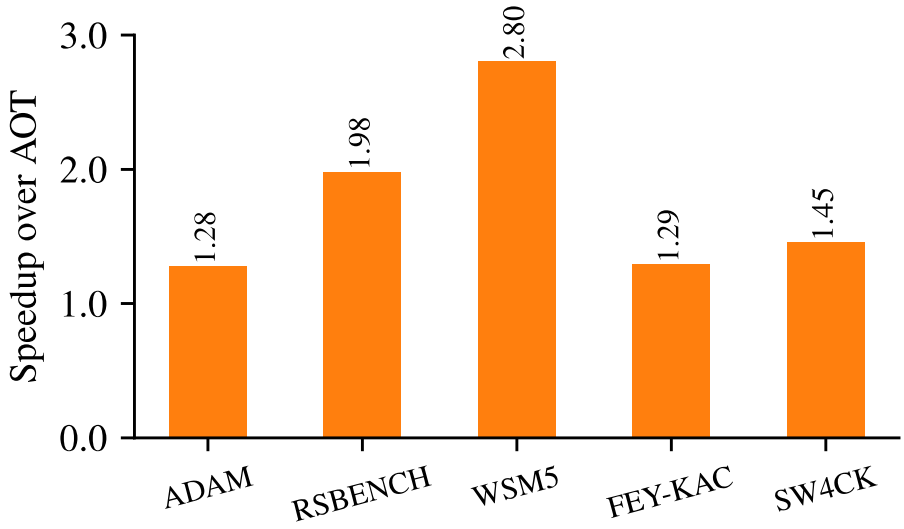

Figure 8: Performance improvements of Proteus versus AOT compilation on Tioga.

Mneme and Proteus offer novel performance optimization tools designed for ease-of-use, scalability, and efficiency. By enabling targeted optimizations through Record-Replay capabilities, Mneme reduces the resource intensity of performance tuning and accelerates the path to optimized code. Proteus, with the use of runtime information, supports compiler flexibility, allowing for code specialization that maximizes performance across varying runtime conditions. Both tools leverage LLVM, a modern compiler used by most applications inside LLNL and by vendors. Together, these tools empower developers to overcome the limitations of traditional manual optimization methods, facilitating solutions for more efficient and scalable scientific software solutions in HPC.

[1] K. Parasyris, G. Georgakoudis, E. Rangel, I. Laguna, and J. Doerfert. “Scalable tuning of (OpenMP) GPU applications via kernel record and replay.” *International Conference for High Performance Computing, Networking, Storage and Analysis*, 2023.

[2] G. Georgakoudis, K. Parasyris, and D. Beckingsale. “Proteus: portable runtime optimization of GPU kernel execution with just-in-time compilation.” To appear at the *International Symposium on Code Generation and Optimization*, 2025.

Advancing the Discipline | Leveraging Classical Computing for Quantum Control: The Quandary Software

Contact: Stefanie Guenther and Anders Petersson

At the intersection of computational mathematics and quantum computing, researchers are addressing one of the most critical challenges of the quantum age: designing precise microwave pulses that execute fundamental quantum operations in superconducting quantum computers. These pulses, which control the delicate quantum states of qubits, must be engineered with exceptional precision to achieve high-fidelity operations, while minimizing time and energy costs.

To meet these challenges, CASC researchers Stefanie Guenther and Anders Petersson collaborate with physicists at LLNL's Center for Quantum Science (LCQS), where the Quantum Design and Integration Testbed (QuDIT) is situated. Ongoing research at CASC is centered around leveraging classical HPC to develop scalable tools to solve these demanding optimization problems efficiently.

One of the key contributions to this effort is the development of the Quandary software, a C++ software package tailored to simulate quantum dynamics and optimize pulse shapes to realize single- and multi-qubit operations [1]. Quandary employs gradient-based optimization techniques using the adjoint method, allowing for efficient computation of gradients for designing pulses that facilitate core quantum operations, such as quantum logical gates, state preparation, or rapid qubit state resets, a prerequisite for initializing quantum algorithms [2] (see Figure 9).

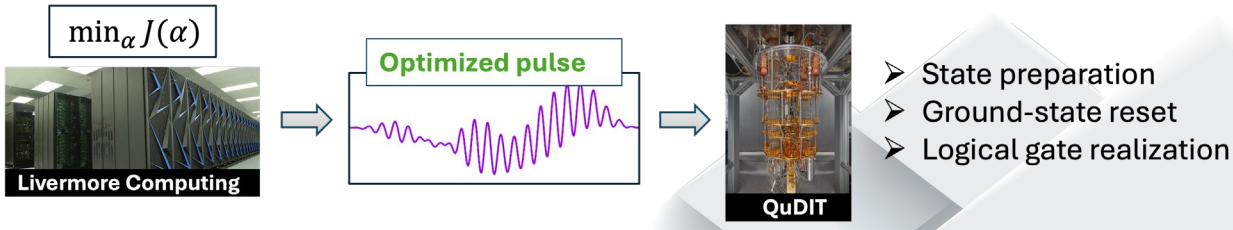

Figure 9: Based on a dynamical model of the quantum system, an optimal control problem is solved on a HPC cluster (left), resulting in an optimized control pulse (center) that then can be applied to the quantum hardware to perform fundamental quantum operations (right).

When employed to multi-qubit transformations and blocks of quantum algorithms, optimal quantum control holds the potential to drastically reduce circuit depth, enabling compression of algorithms into a shorter pulse sequence that provides higher fidelity. Multi-qubit operations necessitate that control pulses adhere to a narrow frequency band, with precisely chosen frequencies that trigger system resonances. Quandary addresses this through a specialized control parameterization based on B-spline basis functions, incorporating carrier waves whose frequencies correspond to the eigenvalue differences of the system's Hamiltonian [3]. This approach further ensures that the generated pulses are effective and finely tuned to the specific dynamics of the quantum system.

Multi-qubit optimization problems, however, are computationally intensive, requiring fine temporal resolution and the management of exponentially scaling state spaces as the number of qubits increases. Quandary leverages the Lab's HPC infrastructure to push the boundaries of quantum control optimization by enabling multiple levels of parallel computing on distributed memory platforms. Recently, they introduced a multiple-shooting algorithm that unlocks parallelism in time, in addition to parallelism in the spatial domain and over individual basis states. This method divides the time-domain of the optimization problem into smaller segments, which can be treated concurrently, thereby drastically reducing computation time for gradient evaluations [4] (see Figure 10).

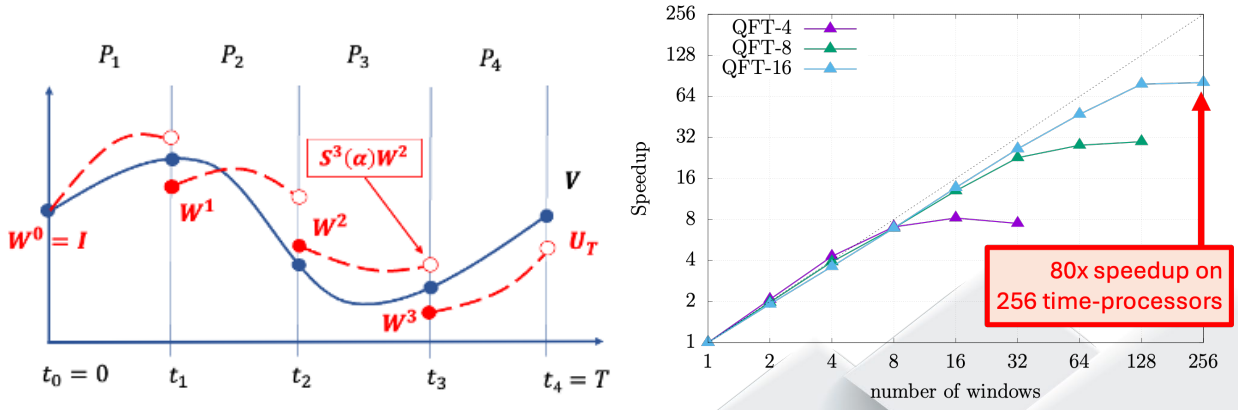

Figure 10: Left: The multiple-shooting approach decomposes the time domain into windows that can be processed concurrently. Intermediate initial states become additional optimization parameters to enforce state continuity. Right: Parallel scaling of gradient computations on LLNL's Dane system, yielding up to 80x speedup from time parallelization on a 4-qubit target gate.

Despite these advances, translating optimized pulses from simulation to existing quantum hardware introduces its own challenges. Real quantum devices often exhibit discrepancies between simulated and observed dynamics due to imperfections in the

hardware or inaccuracies and drift in the model of the system's Hamiltonian. To address this, they are exploring machine learning techniques, particularly the universal differential equation framework, to refine the underlying dynamical models [5]. By learning corrections to the Hamiltonian and decoherence operators from experimental data, they can enhance the predictive fidelity of the simulations and improve alignment with experimental outcomes through continuously refined models (see Figure 11). This research direction not only bridges the gap between theoretical predictions and practical implementations, but also advances their understanding of quantum dynamics in real-world quantum devices.

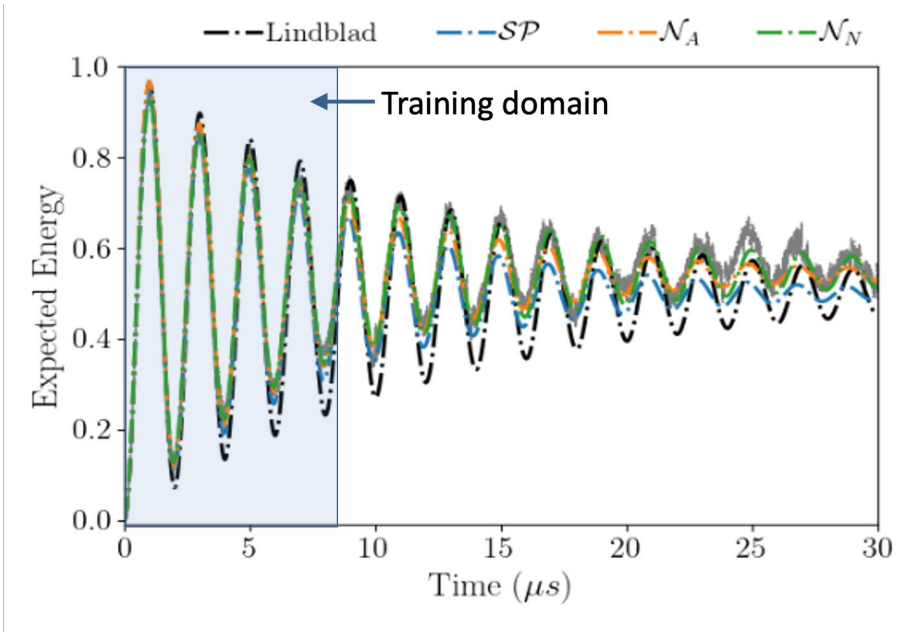

Figure 11: A trained structure-preserving Hamiltonian correction (SP) alongside trained nonlinear neural networks (N_A, N_B) enhance the predictive capabilities of the quantum dynamics described by Lindblad's master equation.

The impact of this research extends beyond pulse design, showcasing the critical role of classical HPC in advancing quantum technologies. As they move forward, the further development of Quandary and the ongoing research to combine advanced optimization methods, HPC, and machine learning will continue to unlock new possibilities for higher fidelity and faster quantum algorithms, bringing them one step closer to realizing the full potential of quantum computing.

[1] S. Günther, N. Petersson, and J. DuBois. "Quandary: an open-source C++ package for high-performance optimal control of open quantum systems." *IEEE/ACM Second International Workshop on Quantum Computing Software (QCS)*, 2021.

[2] S. Günther, N. Petersson, and J. DuBois. "Quantum optimal control for pure-state preparation using one initial state." *AVS Quantum Science*, 3 (4): 043801, 2021.

[3] N. Petersson and F. Garcia. "Optimal control of closed quantum systems via B-splines with carrier waves." *SIAM Journal on Scientific Computing*, 44.6, A3592-A3616, 2022.

[4] N. Petersson, S. Günther, and S. Chung. "A time-parallel multiple-shooting method for large-scale quantum optimal control." *Journal of Computational Physics* (submitted), arXiv preprint: 2407.13950, 2024.

[5] S. Reddy, S. Günther, and Y. Cho. "Data-driven characterization of latent dynamics on quantum testbeds." *AVS Quantum Science*, 6 (3): 033803, 2024.

Machine Learning & Applications | libROM's Breakthroughs in Carbon Capture Simulation

Contact: Youngsoo Choi

Focused on developing efficient data-driven surrogate and reduced-order models (ROM), the libROM team, led by CASC researcher Youngsoo Choi, has been instrumental in advancing computational capabilities for a range of high-impact projects. One such project, the Scaleup LDRD Strategic Initiative, highlights the transformative potential of their innovative approaches.

As the world addresses changes in Earth systems, carbon capture engineering stands at the forefront of technological solutions. High-fidelity multiphysics models play a vital role in designing and optimizing carbon capture systems. However, simulating industry-scale facilities with these models poses significant challenges, even with LLNL's state-of-the-art HPC resources. The memory and computational demands are simply too vast. This bottleneck inspired the libROM team to think outside the box, leading to a novel approach for achieving scalability without compromising accuracy.

The libROM team pioneered the development of component ROMs [1,2]—a framework that can be viewed as a data-driven finite element method. It mirrors the traditional finite element method's process but replaces polynomial basis functions with data-driven basis representations. This innovative technique involves:

1. *Training on small-scale domains:* Generating training data from manageable, small-scale simulations.
2. *Building a solution basis:* Constructing a data-driven solution basis tailored to these small-scale domains.
3. *Assembling for scale:* Combining the small-scale components to simulate large-scale systems efficiently and accurately.

This approach resolves a major challenge in machine learning for physical simulations: the expensive training process for large-scale problems. By training on small-scale domains, the generation of training data finally becomes practical, and the computational cost is drastically reduced. Additionally, the data-driven basis is problem-aware, unlike the polynomial basis in traditional finite element methods, allowing for unprecedented scaling-up effect. Finally, solving the underlying physics at scale through the assembly of small-scale domains ensures that the model retains high fidelity, generalizability, and predictability.

The component-wise ROM approach was tested on critical physics problems relevant to carbon capture, including Stokes [3] and Navier-Stokes flows in porous media [4] and time-dependent Burgers flow [5]. The team demonstrated scale-up capabilities by orders of magnitude—achieving 25- to 1,000-fold speedup and scale-up effect while maintaining accuracy (see Figure 12).

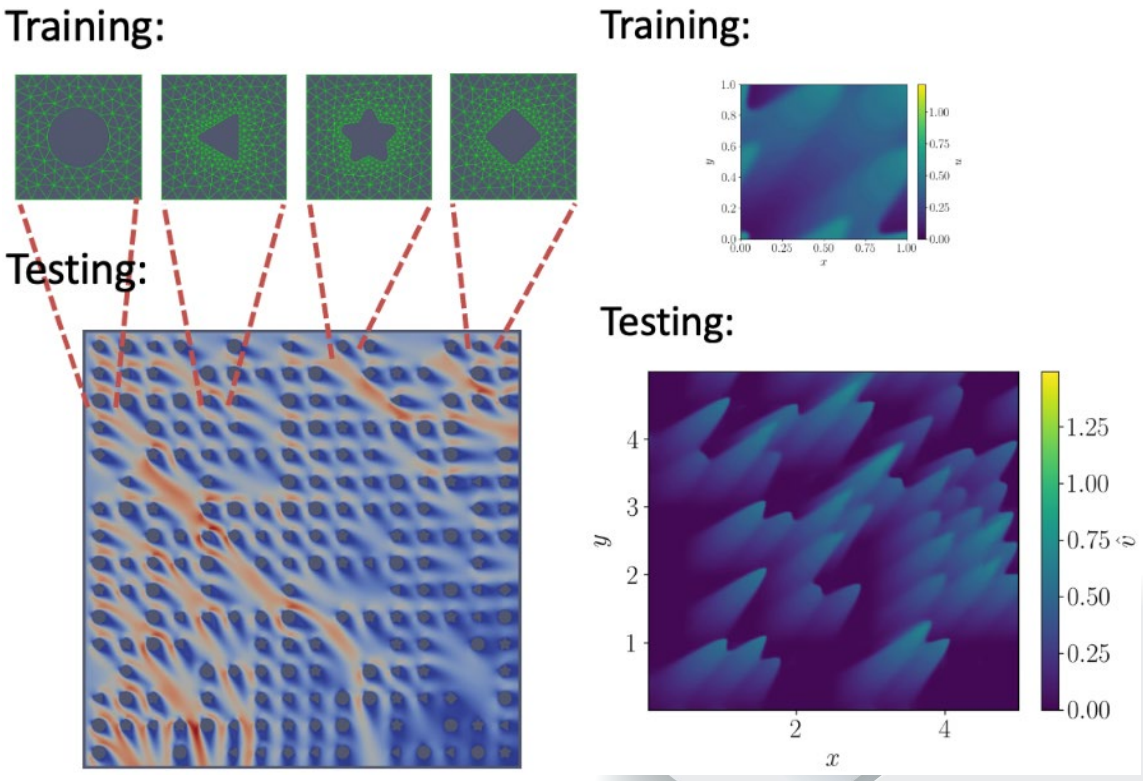

Figure 12: Left: Navier-Stokes porous media flow achieving scaleup of 256x and <2% relative error [2]. Right: Time-dependent 2D Burgers achieving speed-up of 700x and ~1% relative error [3].

While this innovative scaling-up methodology has profound implications for designing more efficient and sustainable carbon capture facilities, fully enabling industry-scale

carbon capture process simulations will require extending the component ROM framework to handle multiscale, multiphysics problems. These include:

- Mass transfer
- Chemical reactions
- Heat transfer
- Multi-phase flow
- Turbulence modeling

The success of this project signals a promising future for reduced-order modeling. The libROM team envisions extending this approach to other domains where computational efficiency is paramount. From fluid dynamics to energy systems, the possibilities are vast. libROM (website) is at the cutting edge of data-driven modeling, pushing the boundaries of what's computationally possible. By combining expertise in machine learning, applied mathematics, and scientific computing, the team continues to deliver scalable solutions to some of the most complex challenges in modern science and engineering.

[1] S. McBane and Y. Choi. "Component-wise reduced order model lattice-type structure design." *Computer Methods in Applied Mechanics and Engineering*, 381: 113813, 2021.

[2] S. McBane, Y. Choi, and K. Willcox. "Stress-constrained topology optimization of lattice-like structures using component-wise reduced order models." *Computer Methods in Applied Mechanics and Engineering*, 400: 115525, 2022.

[3] S. Chung, Y. Choi, P. Roy, T. Moore, T. Roy, T. Lin, D. Nguyen, C. Hahn, E. Duoss, and S. Baker. "Train small, model big: scalable physics simulators via reduced order modeling and domain decomposition." *Computer Methods in Applied Mechanics and Engineering*, 427: 117041, 2024.

[4] S. Chung, Y. Choi, P. Roy, T. Roy, T. Lin, D. Nguyen, C. Hahn, E. Duoss, and S. Baker. "Scaled-up prediction of steady Navier-Stokes equation with component reduced order modeling." arXiv preprint: 2410.21534, 2024.

[5] I. Zanardi, A. Diaz, S. Chung, M. Panesi, and Y. Choi. "Scalable nonlinear manifold reduced order model for dynamical systems." arXiv preprint: 2412.00507, 2024.

CASC Newsletter Sign-up

Was this newsletter link passed along to you? Or did you happen to find it on social media? Sign up to be notified of future newsletters.

This work was performed under the auspices of the U.S. Department of Energy by Lawrence Livermore National Laboratory under Contract DE-AC52-07NA27344. LLNL-MI-2012485, LLNL-MI-2012441. Edited by Ming Jiang.