Performance-Portable Solid Mechanics via Matrix-Free $p$-Multigrid

Jed Brown
University of Colorado, Boulder
ejd@jedbrown.org

Valeria Barra
California Institute of Technology
valeria@caltech.edu

Natalie Beams
University of Tennessee
nbeams@icl.utk.edu

Leila Ghaffari
University of Colorado, Boulder
leila.ghaffari@colorado.edu

Matthew Knepley
University of Buffalo
knepley@gmail.com

William Moses
Massachusetts Institute of Technology
wmoses@mit.edu

Rezgar Shakeri
University of Colorado, Boulder
rezgar.shakeri@colorado.edu

Karen Stengel
University of Colorado, Boulder
karen.stengel@colorado.edu

Jeremy L. Thompson
University of Colorado, Boulder
jeremy.thompson@colorado.edu

Junchao Zhang
Argonne National Laboratory
jczh@anl.gov

Abstract—Finite element analysis of solid mechanics is a foundational tool of modern engineering, with low-order finite element methods and assembled sparse matrices representing the industry standard for implicit analysis. We use performance models and numerical experiments to demonstrate that high-order methods greatly reduce the costs to reach engineering tolerances while enabling effective use of GPUs; these data structures also offer up to 2x benefit for linear elements. We demonstrate the reliability, efficiency, and scalability of matrix-free $p$-multigrid methods with algebraic multigrid coarse solvers through large deformation hyperelastic simulations of multiscale structures. We investigate accuracy, cost, and execution time on multi-node CPU and GPU systems for moderate to large models (millions to billions of degrees of freedom) using AMD MI250X (OLCF Crusher), NVIDIA A100 (NERSC Perlmutter), and V100 (LLNL Lassen and OLCF Summit), resulting in order of magnitude efficiency improvements over a broad range of model properties and scales. We discuss efficient matrix-free representation of Jacobians and demonstrate how automatic differentiation enables rapid development of nonlinear material models without impacting debuggability and workflows targeting GPUs. The methods are broadly applicable and amenable to common workflows, presented here via open source libraries that encapsulate all GPU-specific aspects and are accessible to both new and legacy code, allowing application code to be GPU-oblivious without compromising end-to-end performance on GPUs.

Index Terms—portable, scalable, implicit solvers, matrix-free, solid mechanics, HPC

I. INTRODUCTION

Solid mechanics simulations provide vital information for many engineering applications, using a large amount of computational resources from workstation to supercomputing scales. The industry standard for implicit analysis uses assembled sparse matrices with low-order elements, typically $Q_1$ hexahedral and $P_2$ tetrahedral elements [1], [2], with the linear systems solved using sparse direct solvers, algebraic multigrid, or multilevel domain decomposition. This approach has two fundamental inefficiencies: poor approximation accuracy per Degree of Freedom (DoF) and high computational and memory cost per DoF due to choice of data structures and algorithms. High-order finite elements implemented in a matrix-free fashion with appropriate preconditioning strategies can overcome these inefficiencies.

Solid mechanics models invariably have many stress singularities due to boundary conditions and possible reentrant corners; therefore, $h$-refinement with any finite element basis order $p$ will converge at the same low-order of accuracy. Typically, $hp$-adaptive methods [3] are used to resolve these singularities and enable geometric convergence. Such methods are available in niche commercial products such as StressCheck [4] as well as open source finite element libraries [5], [6], but are rarely used in production computational engineering. This is attributed to accuracy requirements and constant factors: a low-order discretization can usually reach engineering tolerances with a coarse enough mesh that the modeling and implementation complexity of $hp$-adaptive methods are not justifiable, despite their clear asymptotic benefit.

Non-adaptive high order finite elements reduce complexity by being drop-in substitutes for low order elements if one can mesh the geometry more coarsely. Quadratic [7] and higher order [8] elements are often shown to be more accurate per DoF for large deformation analysis despite the presence of singularities preventing any asymptotic benefit. However, such methods are rarely used due to high computation and memory costs for assembly and solution of the linear systems. Sparse matrices for high-order elements have more nonzero entries per DoF: a $Q_2$ hexahedral element contributes 27 nodes per row while $Q_2$ elements have an average of 64 nodes per row, so every sparse matrix-vector product is more than twice as expensive per DoF. Note that vertex separator size stays constant in $h$ versus $p$ refinement to the same number of DoFs, so sparse direct solvers have the same size supernodes and thus asymptotic complexity [9], although the memory use and leaf cost increases. Meanwhile, algebraic multigrid (AMG) setup
costs increase due to sparse matrix-matrix products, and the resulting solvers are observed to converge more slowly for high-order discretizations [10]–[12], even when using specialized methods [13]. A practical alternative to algebraic multigrid is p-multigrid [14], which is observed to be robust for finite elements discretizations on unstructured meshes [11], [12] and pairs naturally with efficient matrix-free data structures [15].

Krylov solvers and preconditioners rely on matrix-vector operations that perform two floating point operations (FLOPs) per stored nonzero. For sparse matrix representations, each nonzero requires 12 bytes (or 16 if using 8-byte integers) to store double precision real values and their indices, yielding an arithmetic intensity [16] of 1 FLOPs. In this new performance model, equipped with matrix-free methods [18]–[20] enable greatly reduced memory bandwidth requirements, often in exchange for modestly more FLOPs. In particular, matrix-free p-multigrid methods and AMG coarse solvers, high-order methods become cheaper per DoF than low-order methods (assembled or not), enabling significantly faster and cheaper simulations at engineering tolerances.

In this paper, we demonstrate that high order methods improve accuracy per DoF for hyperelastic simulations of multiscale structures, even at coarse tolerances in the presence of singularities. We also demonstrate that such models can be solved robustly on a range of modern architectures at a fraction of the cost of per DoF of linear elements, using abstractions amenable to encapsulation and productive development. These benefits are multiplicative, reducing the cost of implicit finite element analysis by up to an order of magnitude in terms applicable to existing analysis pipelines. This paper is organized as follows: section II introduces the hyperelastic formulation and finite element representation, section III describes the solver design and implementation, section IV investigates accuracy in terms of mesh resolution and number of DoFs, section V investigates efficiency per DoF and solver robustness, and section VI discusses implications and opportunities for further work.

II. MATHEMATICAL MODEL

A. Variational Form for Hyperelasticity

In hyperelasticity, one seeks the displacement field \( \mathbf{u}(\mathbf{X}) \) expressing the current (deformed) configuration \( \mathbf{x} = \mathbf{X} + \mathbf{u} \) in terms of the initial configuration \( \mathbf{X} \). An isotropic Neo-Hookean material is defined by its strain energy density

\[
\psi(\mathbf{e}) = \frac{\lambda}{2} (\log J)^2 - \mu \log J + \frac{\mu}{2} (\operatorname{tr} \mathbf{b} - 3) = \frac{\lambda}{2} (\log J)^2 - \mu \log J + \mu \operatorname{tr} \mathbf{e},
\]

where \( \mathbf{b} = (\nabla_X \mathbf{x})(\nabla_X \mathbf{x})^T \) is the left Cauchy-Green tensor, \( J = \det (\nabla_X \mathbf{x}) \), \( \mu \) and \( \lambda \) are the Lamé parameters, and \( \mathbf{e} \) is the Green-Euler strain tensor.

\[
e \equiv \frac{1}{2} (\mathbf{b} - \mathbf{I}) = \frac{1}{2} \left( \nabla_X \mathbf{u} + (\nabla_X \mathbf{u})^T + (\nabla_X \mathbf{u})(\nabla_X \mathbf{u})^T \right).
\]

For a domain \( \Omega_0 \subset \mathbb{R}^3 \) with boundary \( \partial \Omega_0 \) and the finite element space \( \mathcal{V} \subset H^1 (\Omega_0) \), the variational problem finds a solution \( \mathbf{u} \in \mathcal{V} \) such that [10], [21]

\[
\int_{\Omega_0} \nabla_x \mathbf{v} : \mathbf{\tau} \ dV = \int_{\Omega_0} \mathbf{v} : \rho_0 \mathbf{g} \ dV + \int_{\partial \Omega_0} \mathbf{v} \cdot \mathbf{t} \ dS \quad \forall \mathbf{v} \in \mathcal{V},
\]

(2)

where \( \rho_0 \) is the initial mass density, \( \mathbf{g} \) is the body force per unit mass, \( \mathbf{t} \) is the prescribed traction, \( \nabla_x \) denotes spatial derivative with respect to the current configuration, and \( \mathbf{\tau} \) is the Kirchhoff stress given by [22]

\[
\mathbf{\tau} = \frac{\partial \psi}{\partial \mathbf{e}} = \mu (\mathbf{b} - \mathbf{I}) + \lambda \log J \mathbf{I} = 2\mu \mathbf{e} + \lambda \log J \mathbf{I}.
\]

(3)

In order to solve (2) using a Newton iteration algorithm, we need the Jacobian form of \( a(\mathbf{u}, v) \) as [21]

\[
da(\mathbf{v}, d\mathbf{u}; \mathbf{u}) = \int_{\Omega_0} \nabla_x \mathbf{v} : \left( \mathbf{d}\mathbf{\tau} - \mathbf{\tau} \ (\nabla_x d\mathbf{u})^T \right) dV,
\]

(4)

where

\[
d\mathbf{\tau} - \mathbf{\tau} \ (\nabla_x d\mathbf{u})^T = \nabla_x d\mathbf{u} \mathbf{\tau} + \lambda \operatorname{tr} d\mathbf{e} \mathbf{I}
\]

\[
+ 2 (\mu - \lambda \log J) d\mathbf{e},
\]

(5)

with

\[
d\mathbf{e} = \frac{1}{2} \left( \nabla_x d\mathbf{u} + (\nabla_x d\mathbf{u})^T \right).
\]

(6)

B. Matrix-free Finite Element Formulation

The residual (2) and Jacobian (4) forms require derivatives with respect to (solution dependent) current configuration \( \mathbf{x} \). For efficient matrix-free discretization [23], [24], we pull these forms back to reference coordinates \( \xi \) by way of the chain rule \( \nabla_x(\cdot) = \frac{\partial(\cdot)}{\partial \xi} \left( \frac{\partial \xi}{\partial \xi_X} \left( \frac{\partial \xi_X}{\partial \mathbf{x}_X} \right) \right) \), where the part in brackets will move into the variational forms evaluated at quadrature points and \( \nabla_\xi \frac{\partial}{\partial \xi} \) is applied by batched element algebra. We explain this approach for a general Dirichlet problem: find \( \mathbf{u} \in \mathcal{V}_0 \) such that

\[
\langle \mathbf{v}, f(\mathbf{u}) \rangle = \int_{\Omega_0} [\mathbf{v} : f_0(\mathbf{u}, \nabla_x \mathbf{u}) + \nabla_x \mathbf{v} : f_1(\mathbf{u}, \nabla_x \mathbf{u})] dV = 0, \quad \forall \mathbf{v} \in \mathcal{V}_0
\]

(7)

(cf. (2) without traction, where \( f_0 = -\rho_0 \mathbf{g} \) and \( f_1 = \mathbf{\tau} \)). The discrete form of (7) is given by

\[
\mathbf{F}(\mathbf{u}) = \sum_e (\mathbf{X}^e)^T \left[ (Bf_0^e)^T W^e \Lambda (f_0^e(u^e, \nabla_x u^e)) + \sum_{i=1}^{\dim} (Bf_i^e(x^e, u^e)) (f_1^e(u^e, \nabla_x u^e)) \right],
\]

(8)
where \( \mathcal{E}_e \) is the element \( e \) restriction operator that separates DoFs on the elements they belong to, and \( \Lambda \) represents pointwise function evaluation. The diagonal weighting \( W^e = \text{det}(\nabla_x(X)) \Lambda (\hat{W} \otimes \hat{W}) \) are quadrature weights mapped to the physical element. Both \( f_0 \) and \( f_1 \) come from the constitutive law and its tangent where \( u^e = B^T_1 \mathcal{E}_e \mathbf{u} \) and

\[
\nabla_x u^e = \left[ B^e_{x,i} (\mathcal{E}_e u) \right]_{i=1}^{\text{dim}} = \sum_{j=1}^{\text{dim}} \left[ B_{\xi,j} (\mathcal{E}_e u) \right] \frac{\partial \xi_j}{\partial x},
\]

where \( \mathbf{u} \) is the assembled solution vector, \( \text{dim} \) the dimensionality of the problem (for our use cases \( \text{dim} = 3 \)), and

\[
B_1 = B_1 \otimes B_1 \otimes B_1, \quad B_{\xi,1} = B_\xi \otimes B_1 \otimes B_1, \\
B_{\xi,2} = B_1 \otimes B_\xi \otimes B_1, \quad B_{\xi,3} = B_1 \otimes B_1 \otimes B_\xi,
\]

are evaluations on reference elements written in terms of their one dimensional tabulations \( B_1 \) and \( B_\xi \) of shape functions and their derivatives at quadrature points. The representation (9) implies nine tensor contractions to compute a gradient, but this can be reduced to six by first applying \( B_1 \) (3 contractions) and then applying \( B_{\xi,1} \left( B_\xi B_1 \right) \otimes I \otimes I \), and similarly for the other directional derivatives, where \( B_1 \) is the pseudo-inverse satisfying \( B_1^T B_1 = I \) because \( B_1 \) has full column rank. Asymptotically fast structure can also be exploited for simplicial elements [25], [26], but the constants are large enough that direct assembly of the reference element gradient \( B_\xi \) is preferred for the modest basis order considered here.

Pulling (8) back to reference coordinates, we have

\[
F(u) = \sum_e (\mathcal{E}_e)^T \left[ B_1 \begin{bmatrix} f_0 & f_1 \end{bmatrix} \right]^T W^e \Lambda \left[ \begin{bmatrix} f_0 & f_1 \end{bmatrix} \right] \mathcal{E}_e \mathbf{u},
\]

(10)

where \( f_0 \left( u^e, \nabla_\xi u^e \right) = f_0 \left( u^e, \nabla_\xi \mathcal{E}_e u \right) \) and

\[
f_1 \left( u^e, \nabla_\xi u^e \right) = (\nabla_\xi \mathbf{u})^T f_1 \left( u^e, \nabla_\xi \mathcal{E}_e u \right).
\]

While this interface brings the isoparametric mapping into quadrature functions, the work outside these quadrature routines shares the same data and can be batched over elements \( \nabla_\xi u^e = B_\xi \mathcal{E}_e \mathbf{u} \), leading to improved vectorization and data reuse. Moreover, this abstraction provides ready access to element length measures (useful in stabilized methods for transport-dominated processes [27]) and allows optimized data representations, such as bypassing initial configuration because (5) can be evaluated strictly in current configuration, a technique equivalent to that of [10].

The Jacobian action can be computed [23] similar to the residual (10),

\[
J \, d\mathbf{u} = \sum_e (\mathcal{E}_e)^T \left[ B_1 \begin{bmatrix} f_{0,0} & f_{0,1} \\
B_{\xi,1} \end{bmatrix} \right]^T W^e \Lambda \left[ \begin{bmatrix} f_{1,0} & f_{1,1} \end{bmatrix} \right] B_{\xi} \mathcal{E}_e \, d\mathbf{u}
\]

(11)
using Newton-CG, each Newton step uses Krylov Subspace (KSP) and Preconditioning (PC) tools. When method to find \( \Psi \) in optimization [31], but without explicit evaluation of the function, \( F \), that the residual is the functional gradient of a latent objective operator as the fine to coarse grid restriction operator to pre-smoothen the solution. We use the transpose of the prolongation operator, and \( \hat{\mathbf{u}} \) grid prolongation operator, and \( \mathbf{M} \) is the interpolation kernel from the lower order to the higher order finite element, defined by (9), and \( \mathbf{M}_f \) is a pointwise scaling factor for the multiplicity of nodes shared between elements on the fine grid.

The Jacobian on each level is represented by the 17 scalar values per quadrature point, \( \nabla_x \xi, \tau, \log J \), and the quadratic weight, as used in (5) and (11); cf. Table IV. Coarse level discretizations are defined using the same quadrature points and Jacobian representation with coarser basis functions, which is an exact Galerkin method. An alternative needing somewhat more memory, but less memory bandwidth to apply coarse operators, would be to rediscretize by re-evaluating the nonlinearity on a smaller set of quadrature points (sufficient for the lower order polynomials of the coarse space). Note that coarsening from \( Q_2 \) to \( Q_1 \) elements reduces the number of DoFs by a factor of 8 and reduces the number of nonzeros per row (asymptotically on 3D models) by a factor of 64/27. When using direct solvers in 3D, this reduces the vertex and edge elements sharing between elements on the fine grid.

PDE-based models contain symmetry/conservation structure, which is subject to change by a limited subset of stakeholders, and material models (extending (3)) requiring frequent extensions by scientists and engineers who are not sophisticated numerical analysts or software developers. It is thus important that materials developers have a simple, debuggable environment for development and testing. libCEED [15] provides fast algebra for element-based computation on CPUs and GPUs, meant for easy embedding in existing applications and enabling high performance on multiple architectures (Figure 3) from a single source code, with run-time selection of the backend.

The CPU backends call conventionally compiled functions to apply residuals (10) and Jacobians (11) at quadrature points, thereby enabling a rich debugging experience. CPU backends implement the element action \( B \) using tensor contractions with architecture-specific vectorization (e.g., AVX intrinsics, LIBXSMM [33]). The GPU backends used in these experiments create a fused kernel containing the entire \( \mathbf{E}^T \mathbf{B}^T \mathbf{D} \mathbf{B} \mathbf{E} \mathbf{u} \) of (12). The source code for the application of the weak form at the quadrature points, \( \mathbf{D} \), is transformed

\[
P^p_f = \sum_e (\mathbf{E}_f^e)^T \Lambda \left( m_f^{-1} \right) \mathbf{B}_{ctof}^e \mathbf{E}_f^e,
\]
into an appropriate CUDA or HIP device function to be called as part of the fused operator. The resulting kernel is compiled at runtime via NVRT/hipRTC, inlining the above call and making loop bounds and memory access offsets compilation constants, thereby improving register allocation and performance. In this implementation, there is no difference in user code to run on the different architectures. Indeed, the target CPU or (AMD or NVIDIA) GPU may be selected at run-time and need not be uniform across an application. All architecture-specific code is contained within libCEED backends and within PETSc and hypre [34] numerical kernels that are entirely independent from the application.

D. Parallelism and GPUs

While libCEED provides fast algebra on individual CPUs and GPUs, it is important that all problem-sized data stay resident on GPUs throughout the parallel solve. PETSc provides matrix and vector operations on the GPU, including the Galerkin product $P^T \text{ctof} A J \text{ctof} P$ in algebraic multigrid setup, with ability to use external libraries like Kokkos [35] and hypre [34] as well as CUDA and ROCm vendor libraries. Communication and computation are overlapped where possible, with message packing taking place on the GPU along with persistent nonblocking sends and receives using GPU-aware MPI, all via the “star forest” [36] abstraction.

Compressed sparse row (CSR) type matrices are desirable for algebraic multigrid setup and solves, and have historically been created by preallocating and adding values to logically dense blocks using PETSc’s MatSetValues, typically one block per element in a finite element computation. This interface keeps memory utilization low, but is too fine-grained for efficient computing on GPUs and requires a binary search to find the insertion location in the CSR matrix.

We have developed a new interface in PETSc, based on a split-phase COO specification that enables efficient GPU assembly with strong encapsulation. Previous literature [37–40] on GPU-based sparse matrix assembly, including those using COO format [40], used coloring, atomics, or avoid assembling the global matrix to get around data races related to multiple finite elements summing into the same nonzero entries. PETSc’s new COO-based assembly avoids data races completely with new algorithms, and handles MPI parallelism. The classic COO format consists of three arrays, $\text{row}[\cdot], \text{col}[\cdot], \text{val}[\cdot]$, of equal length, in which the assembled matrix $A$ is defined as the sum of each contribution $\text{val}[k]$ to entry $a_{\text{row}[k],\text{col}[k]}$. It is common in nonlinear and transient PDE solves that one needs to assemble a matrix with the same nonzero pattern but different numeric values. PETSc’s interfaces splits COO assembly into a symbolic MatSetPreallocationCOO in which the $\text{row}[\cdot], \text{col}[\cdot]$ parts are provided, followed by one or more calls to MatSetValuesCOO in which the numeric array $\text{val}[\cdot]$ is provided on-device.

In MatSetPreallocationCOO, which is done on-host, we analyze the coordinates, exchange information about remote entries, finalize the sparsity pattern of diagonal and off-diagonal blocks, and preallocate memory for them on-device. This phase prepares to ignore negative indices (convenient for boundary conditions) and sum duplicate entries, as well as planning how to send remote entries to their destination, including which entries in $\text{val}[\cdot]$ should be packed into send buffers. In both PETSc’s native and GPU formats as well as hypre’s ParCSR, matrices are distributed row-wise across processes with diagonal (intra-process coupling) and off-diagonal (inter-process coupling) blocks stored separately in CSR format. The arrays $\text{row}[\cdot], \text{col}[\cdot]$ can be freed after the planning stage.

The $\text{val}[\cdot]$ array is populated on-device using a libCEED kernel that performs the $B^T DB$ portion of the operator (12) (cf. [41]) as a triple matrix product, formulated such that each thread accumulates contributions for a particular element-based non-zero without forming an intermediate matrix. When the number of basis nodes per element is low (up to and including $Q_2$ hexahedra), a two-dimensional thread block processes the row and column combinations in an element’s output matrix; when this design would exceed the allowed number of threads per block, the assembly switches to a one-dimensional thread block with an additional loop in the kernel. Runtime compilation through NVRT/hipRTC ensures that all loop bounds are compile-time constants. Each accumulated value is then assigned to the $\text{val}[\cdot]$ array at a specified index determined by element and component ordering, and the final array is provided to MatSetValuesCOO. Each entry (with nonnegative indices) is destined for the owned diagonal, owned off-diagonal block, or send buffer. The implementation first calls a kernel to fill the send buffer and initiate the MPI communication, two asynchronous kernels for nonzeros in the diagonal and off-diagonal blocks, in which each thread accumulates into a single nonzero, and after completing communication, two similar kernels unpack entries from the receive buffer.
IV. ACCURACY

A. Pareto optimality at engineering tolerances

Real-world structural mechanics problems have numerous reentrant corners and Dirichlet (fixed/clamped) to Neumann (free or applied traction) boundary condition transitions, each of which result in stress singularities. Geometric convergence can be attained for such problems using $hp$-adaptive finite element methods [3], but such methods are rare in industrial practice because adequate tolerances can be achieved on coarser meshes. This can be because the quantity of interest is not so sensitive or because unresolved features (beveling or bolts/washers) or physical yielding will alleviate the singularity in quantities of interest such as the von Mises stress. Using high order finite elements on coarse meshes exposes some nuance, which we explore by way of a representative example. Consider a unit cube with radius 0.3 cylindrical hole, fixed to a rigid boundary on one end and with applied tangential traction on the other. Figure 4 shows the deformed state and strain energy function for a Neo-Hookean material with Young’s modulus 2.4 and Poisson ratio 0.4, and applied traction of 0.2.

We perform a convergence study using linear and high-order geometry meshes produced by Gmsh [42], which can generate arbitrary order curved meshes. Figure 5 shows the relative error in predicted total strain energy $\Psi$ (reference value computed on a highly-resolved mesh) versus DoFs for $h$ and $p$ refinement of the 36-element (3 layers deep) mesh evident in Figure 4 (mesh A) as well as a more resolved mesh B. We observe that $p$ refinement of very coarse meshes is the most efficient path to accuracy. Our experience is consistent with prior empirical studies [43] that a quadratic solution space can be paired with linear geometry, but we also find that further $p$ refinement is actively harmful as non-physical singularities are resolved.

When high order elements are used on the coarsest possible meshes (one element thick), the number of DoFs is often an order of magnitude less than would be required of linear elements to achieve the same accuracy. Therefore, high order methods have better accuracy constants (thus favoring $p$ refinement), and yet they are rarely used in practice, mainly because the assembly and linear algebra are so much more expensive (no improvement in asymptotics). Specifically, the FLOPs per DoF of naive matrix assembly for order $p$ polynomial basis functions in $d$ dimensions scales as $(p+1)^{2d}$ (this can be reduced by specialized methods [44]) and the number of nonzeros per DoF in the assembled matrix scales with $p^d$. The former was a historical bottleneck while the latter is fundamental given the high relative expense of data motion on modern hardware. In contrast, when applying operators matrix-free with quadrature-point data, the storage per DoF declines (and is asymptotically constant) with increasing order $p$. In the following sections, we show that solve costs decrease with increasing $p$ using matrix-free $p$-multigrid and thus Figure 5 is in fact generous to the low-order methods.

B. Schwarz Primitive extrusions under load

Volumetric extrusions of triply periodic minimal surfaces have garnered interest during the additive manufacturing revolution for a range of applications from tissue membranes [45] to metallurgy [46]. We consider the Schwarz Primitive surface, which exhibits interesting geometric and material nonlinearities. Prior finite element analysis of such models [47] using voxelized meshes [48] found that about 30k low order (Abaqus C3D8R) elements were needed to achieve an engineering tolerance of 1%. We consider conformal meshes...
that attain comparable accuracy with fewer DoF and much fewer elements. To generate such meshes, start with a 24-element 2D manifold mesh of a single unit cell embedded in 3D, replicated to the prescribed extent in each embedding dimension. This mesh is partitioned and distributed using ParMETIS, then refined with new nodes projected to the closest point on the implicit surface

\[ \cos 2\pi x + \cos 2\pi y + \cos 2\pi z = 0. \]

The resulting manifold mesh is extruded normal to this surface to the prescribed thickness and number of layers. Figure 6 shows such a model loaded to about 12\% strain on an extent (8, 8, 8) model with about 11.8 million DoF (MDoF). Larger and smaller models are created by changing the extent, keeping the applied surface traction constant so the deformation is similar. These models, which are available in PETSc-3.17, provide excellent tests for solvers since they exhibit all compressive and bending modes, nonlinearities are activated at local and global scale, coarsening is inherently unstructured, and scaling is done by making the domain larger while achieving the same accuracy tolerances, in contrast to the common practice of refining a simpler domain to achieve unrealistically tight accuracy tolerances.

Table I quantifies the effect of mesh refinement and number of layers on the accuracy of a solve with linear, \( Q_2 \), and \( Q_3 \) elements, all with linear geometry. We see that the \( Q_2 \) and \( Q_3 \) solutions are more accurate per DoF than those with linear elements. Increasing the number of layers in the mesh helps decrease the relative error of the simulation; however, 1 or 2 layers is sufficient for both \( Q_2 \) and \( Q_3 \) elements to give solutions within typical engineering accuracy tolerances.

| Order | Refinement | Layers | Disp. X | Disp. Y | Strain | MDoF |
|-------|------------|--------|---------|---------|--------|------|
| 3     | 3          | 1      | 0.42    | 0.47    | 0.90   | 6.0  |
| 3     | 2          | 1      | 1.60    | 1.09    | 3.96   | 1.5  |
| 3     | 1          | 1      | 1.31    | 2.63    | 10.90  | 0.38 |
| 2     | 3          | 5      | 0.42    | 0.33    | 0.92   | 7.3  |
| 2     | 3          | 2      | 0.60    | 0.77    | 1.07   | 3.3  |
| 2     | 2          | 2      | 2.54    | 2.64    | 4.82   | 0.84 |
| 2     | 2          | 1      | 3.85    | 4.34    | 6.15   | 0.50 |
| 1     | 4          | 5      | 2.15    | 2.94    | 2.23   | 4.0  |
| 1     | 3          | 5      | 6.67    | 8.66    | 7.36   | 1.0  |
| 1     | 3          | 2      | 8.88    | 11.52   | 9.70   | 0.50 |
| 1     | 2          | 2      | 22.62   | 27.80   | 25.22  | 0.12 |

V. Performance

A. Compute environments

We present GPU-based results on LLNL’s Lassen, OLCF’s Summit and Crusher, and NERSC’s Perlmutter. Lassen and Summit are both IBM POWER9 machines with 4 and 6 NVIDIA V100-SXM2 16 GiB GPUs per node, respectively. Crusher is an early-access machine with the same node architecture as the upcoming Frontier. Each node has one 64-core AMD EPYC 7A53 CPU and connected via Infinity fabric to 4 AMD MI250X GPUs, each of which consists of two GCDs that appear as logically separate GPUs with 64 GiB each. The GCDs and GPUs are connected via high-bandwidth Infinity fabric, and 4 Cray network interfaces are connected directly to the 4 (dual GCD) GPUs. Perlmutter, which is presently in early access, consists of nodes with one AMD EPYC 7763 CPU connected via PCIe-4.0 to 4 NVIDIA A100 40 GiB GPUs. The GPUs are connected to each other with NVLink-3 and each node has 2 Cray network interfaces connected to the CPU. To compare performance on these machines, we present achieved throughput (DoF/second) normalized by logical GPUs (each of which has similar power requirements). GPU-aware MPI was used on Lassen (Spectrum MPI) and Crusher (Cray MPI), but was disabled on Summit (Spectrum MPI; because results were slower) and Perlmutter (Cray MPI; because of bugs). Table II describes the environment used on each machine.

This study used the open source packages PETSc-3.17 [28], hypre-2.24 [34], [49], Kokkos-3.6 [35], ParMETIS 4.0.3 [50], libCEED-0.10.1 [15], and Ratel-0.1 [51]. The numerical experiments “preload” by doing a crude tolerance solve that is discarded before starting timers in order to provide consistent timing representative of longer-running simulations. For more accurate profiling of individual events, the profiled...
runs include some unnecessary synchronization with the GPU, introducing a slight latency penalty to the smallest model sizes.

### B. Operator application efficiency

In addition to using less memory, the matrix-free representation is much more efficient per DoF to apply. Figure 7 presents performance by varying the domain size (thus total number of DoFs) on a test that runs 3 Newton steps with 500 iterations of CG per step (preconditioned by Jacobi). The figure reports timing for the matrix multiplication operation only. The assembled matrix for this model averages about 63 nonzeros per row (i.e., per DoF) and the empirical STREAM bandwidth on Lassen is 820 GB/s so we expect the matrix multiply to plateau at $820 \times (63 \cdot 12) \approx 1 \text{ GDoF/s}$ if it was only streaming float64 matrix entries and int32 column indices, without any cost to communicate or access vectors. We see that it achieves nearly that and similar math shows that the high order discretizations (with more nonzeros per DoF) also nearly saturate the STREAM bandwidth. The matrix-free discretizations achieve much higher throughput because they store less data per DoF. This model at order 2 has about 30 DoF per element and each element has 27 quadrature points that must store 17 float64 values each, resulting in about $140 \text{ B/DoF}$ (including the vectors) for the matrix-free operator, and a predicted streaming peak of a bit under $6 \text{ GDoF/s}$. About half of that is achieved, with the discrepancy attributable to atomic writes, vector zeroing and copies/packing related to communication and boundary conditions, and the computation to compute gradients and apply the quadrature points operations.

Note that latency is an ever-present specter, with efficiency still rising at the point when GPU memory capacity is reached. Moreover, many applications have strict time-to-solution requirements imposed by business, policy, or human timelines, and thus it is informative to report the time at which, say 80% of peak efficiency is achieved. In the subsequent section, we will place time on the $x$ axis, allowing us to compare efficiency of different machines and different parallel scale.

### C. Nonlinear solves

To test the efficiency of end-to-end nonlinear solves, we choose the model from Figure 6 with (nondimensionalized) parameters thickness 0.2, Young’s modulus 1, and Poisson ratio 0.3, fixed to the left wall with a compressive traction of 0.02 applied from the right. This produces approximately 12% strain at every resolution, which is just shy of where plastic yielding occurs for photopolymer additively manufactured products of these models [47]. This model requires 5 to 7 Newton iterations across the range of resolutions, with each linear solve needing 9 to 25 preconditioned CG iterations to converge to a relative tolerance of $10^{-3}$ in the natural norm, with CG condition number estimates from 9.5 to 61 (mostly less than 15 iterations and condition numbers less than 20; depending on the Newton step).

We sweep through a range of Primitive model extents up to $20^3$ per node of Crusher (184 MDoF), solve each model, and plot efficiency versus time per Newton iteration for $Q_2$ and $Q_3$ elements in Figure 8 and Figure 9. The $Q_2$ model is as depicted in Figure 6 with 2 refinements and 2 extruded layers, while the $Q_3$ model uses only one extruded layer to achieve somewhat better accuracy; see Table I. In such plots, perfect weak scaling would have the 1-node and 8-node curves on top of each other, with strong scaling limits visible in the minimum time at which acceptable efficiency can be achieved. This human-centric figure is meant to assist the analyst with cloud or HPC access in choosing an efficiency-versus-time tradeoff. For example, one may look at Figure 9 and decide that under 2 s per Newton iteration (about 10 s for the total nonlinear solve) delivers an acceptable efficiency-time tradeoff. Examining the Perlmutter curve with about 3 MDoF/s/GPU at 2 s, the target problem would be scaled to about 6 MDoF/GPU. The 1-node and 8-node Perlmutter curves lie on top of each other here, indicating that one can solve a 24 MDoF problem on one node (4 GPUs) with the same efficiency as a 192 MDoF problem on 8 nodes. Note that AMG requires a deeper V-cycle for the larger problem size, but this latency impact is hidden at the 2 s solve time with $Q_3$ elements. Compare with Figure 8, in which there is a slight efficiency penalty to the weak scaling since a greater fraction of the solve time is spent in AMG when using $Q_2$ elements. The solve can be made somewhat faster by using more GPUs (with some drop in efficiency) or more...
Fig. 8. Efficiency per Newton iteration versus time for $Q_2$ finite elements using matrix-free Newton-Krylov with $p$-MG preconditioning and Boomer-AMG coarse solve. Problem sizes (in MDoF) are annotated for the minimum and maximum sizes for each host and number of nodes combination. The impact of latency is ever-present, with memory capacity limiting the right end of each curve.

Fig. 9. Efficiency per Newton iteration versus time for $Q_3$ finite elements using matrix-free Newton-Krylov with $p$-$\text{MG}$ preconditioning and Boomer-AMG coarse solve. Problem sizes (in MDoF) are annotated for the minimum and maximum sizes for each host and number of nodes combination. Ideal weak scaling is evident on Perlmutter for Newton step time above 1.8 s where the 1-node and 8-node curves coincide, while communication latency leads to degradation at the smallest problem sizes (2 MDoF/GPU with time around 1 s).

Fig. 10. Flame graph for a typical setup and solve with $Q_2$ elements. The two dominant costs are preconditioner application (left half; part of the linear solve) and setup (center-right third). The coarse solve (Level 0 and above) takes about half the time of the fine (Level 1) and coarse setup (AMG) takes more time than fine $p$-MG setup, despite the fine having 8x more DoF.

conditioner setup ($\text{PCSetUp}$) dominating. The linear solve (Figure 11 and 12) is communication-intensive since each preconditioner application goes through a V-cycle (with an increasing number of levels as the model gets larger).

Figure 13 considers preconditioner setup, which consists of algebraic multigrid analysis and Galerkin products as well as a few Krylov iterations to calibrate the smoothers. Assembly of the coarse Jacobian ($\text{SNESJacobianEval}$; far-right yellow in Figure 10) exhibits nearly perfect problem-size independence at about 20–25 MDoF/s/GPU (with $Q_2$ elements) on Crusher and Perlmutter, and is thus always less than 20% overhead. The relative cost of Jacobian assembly and preconditioner setup both decrease when going to $Q_3$ elements because the coarse problem is a smaller fraction of the fine problem size. When solving transient problems or nonlinear problems by quasi-Newton methods, the preconditioner setup can often be reused across many solves and thus this phase becomes insignificant in comparison to the increasingly dominant linear solve. Although these figures show Newton-Krylov methods due to their familiar properties and diagnostics, we observe...
up to 2x performance improvement when using L-BFGS as described in subsection III-A, and recommend testing it on representative problems.

In general, we observe greater volatility in the “strong scaling” regime at the left edge of Figures 8 to 13. Most configurations reach high efficiency weak scaling (solid and dotted lines very close) as the problem size per GPU increases, leading to Newton solve times increasing to around 2s and higher. This efficient weak scaling is usually realized at smaller (faster) solves than where performance plateaus, indicating that 1-node architectural latencies are a more insidious performance obstacle than multi-node communication. Although Crusher exhibits a regime of efficient scaling, the efficiency degrades at the largest problem sizes. This effect is not present on other machines and our profiling points to network degradation not identifiable in microbenchmarks (or smaller problem sizes) that we hope will be resolved in the MPI implementation or tuning. Specifically, the majority of the degradation is attributable to point-to-point messaging and Jacobian assembly communication that performance models indicate should be cheap relative to volume work because these huge subdomains have low surface area to volume ratio.

**D. Robustness**

We now explore under what circumstances the p-MG limits convergence versus when it is limited by the AMG coarse solve. In order to make direct solvers affordable, we consider the Primitive model with extent (8,2,2) under 0.001 tension and report the iteration count and condition number from the first solve using Hypre, GAMG, and Cholesky. Table III investigates convergence for the mildly stretched elements in the original thickness 0.2 and more stretched in thickness 0.05, both with two layers. We have fixed solver parameters to be representative of problems with both well-shaped and stretched elements; better convergence on stretched models can be obtained by tuning threshold and smoothing parameters for the worst quality elements, at the expense of degraded convergence for better-shaped elements (left column).

**E. Usability via Automatic Differentiation**

Efficient use of matrix-free methods requires quadrature-point based linearization (“partial assembly”) of forward (and possibly adjoint) operators. While many problems have structure [10], [23] that can reduce the memory footprint and operation count, it can be tedious to find these formulations and it is onerous to have to develop the nonlinear residual and Jacobian action synchronously. Automatic differentiation (AD) tools simplify this process, automating the Jacobian action so that only the nonlinear forward model needs to be written by a human. Enzyme [52] is a new LLVM plugin with GPU support that provides split forward and reverse mode AD on LLVM intermediate representation (IR).

We investigate applicability and performance computing the Jacobian action using Enzyme’s new (version 0.0.29) split forward-mode capability to provide derivatives of Neo-Hookean models. Split mode populates a “tape”, which contains opaque intermediate values at quadrature points, and is stored in ordinary libCEED arrays (output from residual computation and input to Jacobian evaluation). The material model expressed in current configuration (3) is too simple for this test so we include tests of the same model expressed in initial configuration: the second Piola-Kirchhoff stress as a function of the Green-Lagrange strain, $S(E)$. This model contains matrix inverses and thus its analytic derivative uses the identity $dC^{-1} = -C^{-1}dC C^{-1}$, but this is not known to Enzyme. Enzyme identifies a straightforward and relatively low-memory representation (small tape) given the provided structure, and the resulting vectorized code is on par with naive hand-written code that doesn’t exploit symmetries and cancellation. Table IV compares total solve time (over many steps) on a small cube mesh with 3630 DoF (fits in cache, thus stresses flops) on a single process of an AMD EPYC 7452. The initial configuration cases need to store initial configuration geometry $\nabla X \xi$ and quadrature-weighted determinants $W$, while the current configuration maps directly to the solution-dependent current configuration $\nabla X \xi$. Since Enzyme is language-agnostic (by virtue of operating on LLVM IR), this opens the door to constitutive modeling in safer/more convenient languages, such as Rust and Julia, with no impact on execution performance or environment.

![Fig. 13. Preconditioner setup efficiency spectrum for $Q_2$ finite elements using matrix-free Newton-Krylov with p-MG preconditioning and BoomerAMG coarse solve. The times and efficiencies are per Newton iteration. Problem sizes (in MDoF) are annotated for the minimum and maximum sizes for each host and number of nodes combination.](image-url)
TABLE IV

| Problem          | Storage | Scalars | Time (s) |
|------------------|---------|---------|----------|
| current          | $W; \nabla X \xi$, $\tau$, $A \log J$ | 17 | 7.097    |
| initial native   | $\nabla X \xi$, $W; \nabla X u$ | 19 | 11.556   |
| initial tuned    | $\nabla X \xi$, $W; \nabla X u$, $C^{-1}$, $\lambda \log J$ | 26 | 9.498    |
| initial AD       | $\nabla X \xi$, $W; \nabla X u$, $S$, tape | 31 | 10.661   |

VI. DISCUSSION

High order methods have thus far made little impact on industrial practice of structural engineering primarily due to performance consequences of traditional sparse matrix abstractions, which offer increasingly poor utilization of modern hardware. We have shown that this performance landscape is inverted by changing data structures to matrix-free representations with linearization defined at quadrature points, for which high order methods are significantly cheaper per DoF. The 1-node Crusher examples solve problems of similar size (hundreds of MDoF) and complexity to the implicit structural mechanics problems in the 2002 [53] and 2004 [54] Gordon Bell Prizes, which are still considered large in industrial and research structural mechanics practice. Our methods enable pragmatic use of $Q_2$ and $Q_3$ elements while delivering much faster time to solution. While we have focused our study here on problems of size less than 2 GDoF to maximize interpretability and relevance to practitioners, the algorithms are scalable to much larger problems and node counts.

Similar structure has previously been exploited by [55] to reduce memory requirements by not storing the fine-grid matrix for bone structure analysis, while still constructing prolongation operators for smoothed aggregation AMG preconditioning. This work was restricted to linear elasticity on voxelized meshes and the setup and solve time was somewhat longer than with standard assembled methods. Other recent work [10] (based on the fast matrix-free operators [56] in the deal.II library) used geometric $h$-multigrid for high order finite elements applied to hyperelasticity on CPUs, showing excellent performance for a matrix-free methods relative to assembled methods. In particular, the matrix-free iteration counts were found to be much smaller than AMG applied directly to the assembled high-order discretization, and each iteration was cheaper by virtue of the matrix-free data structures. In that work, the high order discretization was preserved on nested coarse grids, which limits applicability to problems with high geometric complexity. Non-nested geometric multigrid [57]–[59] could be extended to high order elements, but studies of such methods in complex geometry [58] have encountered robustness problems relative to algebraic multigrid.

The matrix-free $p$-multigrid approach presented here offers the robustness of low-order algebraic multigrid with much higher efficiency per DoF and simulation time to reach engineering tolerances. The use of high order elements has the additional benefit that coarser meshes can be used, thus reducing preprocessing time and I/O costs related to element topology, though it requires more attention to element quality at the mesh generation stage. We find that quadratic geometry is often sufficient for large deformation with quadratic and cubic solution spaces, and thus these methods can be used with existing meshing and visualization tools, though tailoring to $p$-version finite element efficiency [60] is beneficial. When cubic and higher order meshes are needed, one can use Gmsh [42] to generate arbitrary order meshes, but many popular mesh formats support at most second order elements and there is a need for improving data representation standards and postprocessing/visualization tools to better support high order geometry and solution fields [61]–[63].

We find that the algorithms here provide substantial benefit already for quadratic elements, and thus is a viable drop-in technique any time quadratic geometry can be used, and sometimes even for linear geometry elements. One can switch from linear to quadratic elements on the same mesh for about double the cost, despite 8 times more DoFs. The method is applicable for almost any problem in which the coarsest geometry-resolving mesh is not accurate enough for simulation with linear elements. Despite equivalent asymptotic convergence in the presence of singularities for high order methods, the combination of constants for approximation and algorithmic implementation efficiency often leads to an order of magnitude reduction in cost to reach engineering tolerances, offering a transformative opportunity to make batch simulations interactive and greatly expand the use and fidelity of solid mechanics simulation in science and industry. While the methods require revisiting the traditional centrality of sparse matrices for implicit finite element analysis, most of the algorithmic structure remains intact (with a new economy that inherits key instances of conventional wisdom to enable further efficiency gains). Note that libCEED was designed for use in legacy software; adoption by conventional CPU-based implicit FEA software is mostly a matter of calling material models in libCEED Q-functions and modest data structure abstraction in the solver via standard interfaces provided by PETSc and similar libraries.

One limitation to the matrix-free $p$-multigrid technique is that Chebyshev/Jacobi smoothing degrades for highly stretched elements, such as appear in volumetric discretizations of shell structures. One needs either semi-coarsening or block/line smoothers to make multigrid convergence uniform on such models, neither of which is especially convenient in the present framework. We note that shell structures usually have small vertex separators and thus direct solvers and parallel adaptive BDDC solvers such as PETSc’s PCBDDC [64] offer sharp convergence guarantees at manageable cost.

ACKNOWLEDGMENT

This research is supported by the Exascale Computing Project (17-SC-20-SC), a collaborative effort of two U.S. Department of Energy organizations (Office of Science and the National Nuclear Security Administration) responsible for...
the planning and preparation of a capable exascale ecosystem, including software, applications, hardware, advanced system engineering and early testbed platforms, in support of the nation’s exascale computing imperative. This research is supported by the U.S. Department of Energy, Office of Science, Office of Advanced Scientific Computing Research under contract DE-AC02-06CH11357 and Award Number DE-SC0016140. The authors acknowledge support by the Department of Energy, National Nuclear Security Administration, Predictive Science Academic Alliance Program (PSAAP) under Award Number DE-NA0003962. This research used resources of the Oak Ridge Leadership Computing Facility at the Oak Ridge National Laboratory, which is supported by the Office of Science of the U.S. Department of Energy under Contract No. DE-AC05-00OR22725. This research used resources of the Livermore Computing Facility at the Lawrence Livermore National Laboratory. This research used resources of the National Energy Research Scientific Computing Center, which is supported by the Office of Science of the U.S. Department of Energy under Contract No. DE-AC02-05CH11231. We thank PETSc, hypre, and libCEED developers, especially Mark Adams, Barry Smith, Stefano Zampini, Victor Paludetto Magri, Veselin Dobrev, Yohann Dudouit, and Tzanio Kolev, for collaboration on software development and algorithmic tuning. This research used Paraview, Seaborn, Matplotlib, and Pandas for data analysis and visualization.

REFERENCES

[1] Simulia Dassault Systemes. Abaqus Standard. https://www.3ds.com/products-services/simulia/products/abajas/abaqusstandard/, 2021.
[2] Ansys. Ansys Mechanical. https://www.ansys.com/products/structures/ansys-mechanical, 2021.
[3] Ivo Babuška and Manil Suri. The p and h – p versions of the finite element method, basic principles and properties. SIAM Review, 36(4):578–632, 1994.
[4] ESRD. StressCheck Professional. https://www.esrd.com/products/stresscheck-professional/, 2021.
[5] Wolfgang Bangerth and Oliver Kayser-Herold. Data structures and requirements for hp finite element software. ACM Transactions on Mathematical Software (TOMS), 36(1):1–31, 2009.
[6] P. Frauenfelder and C. Lage. Concepts—An Object-Oriented Software Package for Partial Differential Equations. Mathematical Modelling and Numerical Analysis, 36(5):937–951, 2002.
[7] Teseo Schneider, Yixin Hu, Xifeng Gao, Jérémy Dumas, Denis Zorin, and Daniele Panozzo. A large-scale comparison of tetrahedral and hexahedral elements for solving elliptic pdes with the finite element method. ACM Transactions on Graphics, 41(3), mar 2022.
[8] Alexander Düster, Stefan Hartmann, and Ernst Rank. p-fem applied to finite isotropic hyperelastic bodies. Computer Methods in Applied Mechanics and Engineering, 192(47-48):5147–5166, 2003.
[9] A. George, J. Liu, and E. Ng. Computer Solution of Sparse Linear Systems. Oak Ridge National Laboratory, 1994.
[10] Denis Davydov, Jean-Paul Pelteret, Daniel Arndt, Martin Kronbichler, and Paul Steinmann. A matrix-free approach for finite-strain hyperelastic problems using geometric multigrid. International Journal for Numerical Methods in Engineering, 121(13):2874–2895, 2020.
[11] Alexander Weber, Johannes Mueller-Roemer, Christian Altenhofen, André Stork, and Dieter Fehlner. Deformation simulation using cubic finite elements and efficient p-multigrid methods. Computers & Graphics, 53:185–195, 2015.
[12] Hari Sundar, Georg Stadler, and George Biros. Comparison of multigrid algorithms for high-order continuous finite element discretizations. Numerical Linear Algebra with Applications, 22(4):664–680, 2015.
[13] JJ Heys, TA Manteuffel, Steve F McCormick, and LN Olson. Algebraic multigrid for higher-order finite elements. Journal of Computational Mathematics, 20(4):520–532, 2005.
[14] Einar M Roquist and Anthony T Patera. Spectral element multigrid. I. formulation and numerical results. Journal of Scientific Computing, 2(4):389–406, 1987.
[15] Jed Brown, Ahmad Abdelfatah, Valeria Barra, Natalie Beans, Jean-Sylvain Camier, Veselin Dobrev, Yohann Dudouit, Leila Ghaffari, Tzanio Kolev, David Medina, et al. libceed: Fast algebra for high-order element-based discretizations. Journal of Open Source Software, 6(65):2945, 2021.
[16] Samuel Williams, Andrew Waterman, and David Patterson. Roofline: an insightful visual performance model for multicore architectures. Communications of the ACM, 52(4):65–76, 2009.
[17] Karl Rupp. CPU-GPU-MIC comparison charts. https://github.com/karlrupp/cpu-gpu-mic-comparison, 2020.
[18] Tzanio Kolev, Paul Fischer, Misun Min, Jack Dongarra, Jed Brown, Veselin Dobrev, Tim Warburton, Stanimire Tomov, Mark S. Shepard, Ahmad Abdelfatah, Valeria Barra, Natalie Beans, Jean-Sylvain Camier, Noel Chalmers, Yohann Dudouit, Ali Karakus, Stefan Kerkemeier, Tzanio Kolev, Yu-Hsiang Lan, David Medina, Elia Merazzi, Aleksandr Obabko, Will Pazner, Thilina Rathnayake, Cameron W. Smith, Lukas Spies, Kasia Swirydowicz, Jeremy Thompson, Ananias Tomboulides, and Vladimir Tomov. Efficient exascale discretizations: High-order finite element methods. International Journal of High Performance Computing Applications, 2021.
[19] Ahmad Abdelfatah, Valeria Barra, Natalie Beans, Ryan Bleie, Jed Brown, Jean-Sylvain Camier, Robert Carson, Noel Chalmers, Veselin Dobrev, Yohann Dudouit, Paul Fischer, Ali Karakus, Stefan Kerkemeier, Tzanio Kolev, Yu-Hsiang Lan, Elia Merazzi, Misun Min, Malachi Phillips, Thilina Rathnayake, Robert Rieben, Thomas Stitt, Ananias Tomboulides, Stanimire Tomov, Vladimir Tomov, Atturo Vargas, Timothy Warburton, and Kenneth Weiss. GPU algorithms for efficient exascale discretizations. Parallel Computing, 108:102841, 2021.
[20] Dave A. May, Jed Brown, and Laetitia Le Pourhiet. Tat3D: High-performance methods for long-term lithospheric dynamics. In Proceedings of SC14: International Conference for High Performance Computing, Networking, Storage and Analysis. ACM, 2014.
[21] Arash Mehrabian, Jed Brown, Henry Tufo, Jeremy Thompson, Rezgar Shakeri, and Richard Regueiro. Efficient parallel scalable matrix-free 3d high-order finite element simulation of neo-hookean compressible hyperelasticity at finite strain. In ASME International Mechanical Engineering Congress and Exposition, volume 85860, page V012T12A027. American Society of Mechanical Engineers, 2021.
[22] Gerhard A Holzapfel. Nonlinear solid mechanics: a continuum approach for engineering science. Meccanica, 37(4):489–490, 2002.
[23] Jed Brown. Efficient nonlinear solvers for nodal high-order finite elements in 3D. Journal of Scientific Computing, 45(1-3):48–63, 2010.
[24] M. G. Knepley, J. Brown, K. Rupp, and B. F. Smith. Achieving high performance with unified residual evaluation. arXiv:1309.1204, September 2013.
[25] Robert C Kirby. Fast simplicial finite element algorithms using Bernstein polynomials. Numerische Mathematik, 117(4):631–652, 2011.
[26] Jesse Chan and T. Warburton. GPU-accelerated Bernstein–Bézier discontinuous Galerkin methods for wave problems. SIAM Journal on Scientific Computing, 39(2):A626–A654, 2017.
[27] Thomas J R Hughes, Guglielmo Scovazzi, and Tayfun E Tezduyar. Stabilized methods for compressible flows. Journal of Scientific Computing, 43:334–368, 2010.
[28] Satish Balay, Shrirang Abhyankar, Mark F. Adams, Steven Benson, Jed Brown, Peter Brune, Kris Buschelman, Emil Constantinescu, Lisandro Dalcin, Alp Dener, Victor Eijkhout, William D. Gropp, Václav Hapla, Tobin Isaac, Pierre Jolivet, Dmitry Karpeev, Dinesh Kaushik, Mark Adams, Barry F. Smith, Stefano Zampini, Hong Zhang, Hong Zhang, and Richard Tran Mills, Mark S. Shepard, Ahmad Abdelfatah, Valeria Barra, Natalie Beans, Jean-Sylvain Camier, Noel Chalmers, Yohann Dudouit, Ali Karakus, Stefan Kerkemeier, Tzanio Kolev, Yu-Hsiang Lan, David Medina, Elia Merazzi, Misun Min, Malachi Phillips, Thilina Rathnayake, Robert Rieben, Thomas Stitt, Ananias Tomboulides, Stanimire Tomov, Vladimir Tomov, Atturo Vargas, Timothy Warburton, and Kenneth Weiss. PETSc/TAO users manual. Technical Report ANL-1703E, Argonne National Laboratory, 2022.
