

---

# MATRIX-FREE APPROACHES FOR GPU ACCELERATION OF A HIGH-ORDER FINITE ELEMENT HYDRODYNAMICS APPLICATION USING MFEM, UMPIRE, AND RAJA

---

Arturo Vargas<sup>†</sup>, Thomas M. Stitt<sup>†</sup>, Kenneth Weiss<sup>†</sup>, Vladimir Z. Tomov<sup>\*</sup>,  
Jean-Sylvain Camier<sup>\*</sup>, Tzanio Kolev<sup>\*</sup>, Robert N. Rieben<sup>†</sup>,

Weapons and Complex Integration, Lawrence Livermore National Laboratory, Livermore, CA, USA<sup>†</sup>,  
Center for Applied Scientific Computing, Lawrence Livermore National Laboratory, Livermore, CA, USA<sup>\*</sup>

## ABSTRACT

With the introduction of advanced heterogeneous computing architectures based on GPU accelerators, large-scale production codes have had to rethink their numerical algorithms and incorporate new programming models and memory management strategies in order to run efficiently on the latest supercomputers. In this work we discuss our co-design strategy to address these challenges and achieve performance and portability with MARBL, a next-generation multi-physics code in development at Lawrence Livermore National Laboratory. We present a two-fold approach, wherein new hardware is used to motivate both new algorithms and new abstraction layers, resulting in a single source application code suitable for a variety of platforms. Focusing on MARBL's ALE hydrodynamics package, we demonstrate scalability on different platforms and highlight that many of our innovations have been contributed back to open-source software libraries, such as MFEM (finite element algorithms) and RAJA (kernel abstractions).

**Keywords** High-order, Finite Elements, GPUs, Performance Portability

## 1 Introduction

The computational landscape in high-performance computing (HPC) has undergone a dramatic change in recent years. General-purpose CPUs with large caches and sophisticated pipeline executions have been largely replaced by low-power data-parallel GPU accelerators in the vast majority of leadership class US supercomputers. While this hardware shift has paved the way for large-scale simulation codes towards exascale computing [1], it has come at a significant development cost, since HPC applications have had to rethink their numerical algorithms and incorporate new programming models and memory management strategies in order to run efficiently on these advanced architectures.

In this paper we discuss our experience with the GPU port of one such HPC application, the MARBL code, and review our co-design strategy to address the challenges of performance and portability across different computing architectures. We present a two-fold approach, wherein new hardware is used to motivate both new algorithms and abstraction layers, resulting in a single source application code suitable for a variety of platforms. Our work includes the development of GPU-oriented matrix-free algorithms, porting and optimization via mini-apps, and the utilization of open-source software libraries, such as MFEM (finite element algorithms) [2], RAJA (kernel abstraction) [3] and Umpire (memory management) [4]. We demonstrate scalability on different platforms and highlight that many of our innovations have been contributed back to the MFEM and RAJA libraries.

Our focus in this work is on MARBL, a next-generation multi-physics simulation code developed at Lawrence Livermore National Laboratory for simulating high energy density physics (HEDP) and focused experiments driven by high-explosive, magnetic or laser based energy sources for applications including pulsed power and inertial confinement fusion (ICF) [5]. MARBL is built on modular physics and computer science components and makes extensive use of high-order finite element numerical methods. Compared to standard low-order finite volume schemes, high-order

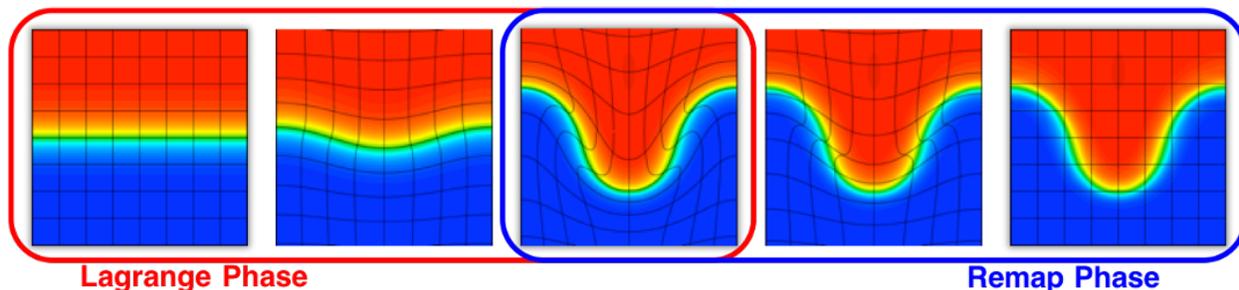


Figure 1: MARBL follows a three-phase high-order ALE algorithm: Lagrange, Mesh Optimization, Remap.

numerical methods have more resolution/accuracy per unknown and have higher FLOP/byte ratios meaning that more floating-point operations are performed for each piece of data retrieved from memory. This leads to improved strong parallel scalability, better throughput on GPU platforms and increased computational efficiency.

Prior to the work described in this paper, MARBL’s Arbitrary Lagrangian-Eulerian (ALE) multi-material compressible hydrodynamics capability was an MPI-only massively parallel code based on full matrix assembly. In the remaining sections, we describe the process of restructuring MARBL’s ALE component to an MPI + X code, where X is a shared memory threading model. In addition, we discuss the algorithmic refactoring which was done to convert the full matrix assembly approach into a more efficient and high-performance version based on “matrix-free” partial assembly techniques. A key detail in our success was the co-design effort between the application and tool libraries (MFEM, RAJA and Umpire) to steer future directions in library developments. One of the main lessons learned in this effort was that customization of established portability tools were crucial to developer productivity and application performance.

The rest of this article is organized as follows: We begin with an overview of the high-order finite element ALE hydrodynamics approach in MARBL, and describe key concepts of the finite element formulation and algorithmic ideas in Section 2. We then discuss the approach for achieving performance and portability in Sections 3 and 4. Our two-pronged approach includes algorithmic improvements, in particular matrix-free partial assembly, as well as performance improvements achieved by exposing key programming features through our portable kernel layer abstraction and memory-based optimizations through the Umpire library. We demonstrate the overall performance and scalability of MARBL across both GPU and CPU platforms in Section 5, review our mini-app driven development in Section 6, and showcase these developments in the context of high-order 3D multi-material simulations in Section 7. Finally, we provide some concluding remarks in Section 8.

## 2 Overview of ALE

MARBL is a multi-physics code including three temperature radiation-magneto-hydrodynamics with thermonuclear burn for fusion plasmas. Central to it is the multi-material Arbitrary Lagrangian Eulerian (ALE) compressible hydrodynamics capability (based on the BLAST code, see for example [6]) which we will focus on in this paper. The high-order finite element ALE approach of MARBL splits the computation into three phases as illustrated in Figure 1. The first phase (Lagrange phase) evolves the simulation within a Lagrangian frame of reference until the mesh reaches a level of distortion, the mesh is then relaxed (mesh optimization phase), and the solution is then remapped onto the optimized mesh (remap phase). We provide a high-level overview of each of the phases in this section and review key algorithmic implementation details in Section 3.

### 2.1 High-Order Lagrange Phase

The Lagrange phase solves the time dependent Euler equations of compressible hydrodynamics within a Lagrangian frame of reference [7]. In the case of multi-materials,  $k = 1, \dots, N_{\text{Materials}}$ , we solve the following conservation laws

from hydrodynamics:

$$\rho \frac{dv}{dt} = \nabla \cdot \sum_k \eta_k \sigma_k \quad (1)$$

$$\frac{d\eta_k}{dt} = \alpha_k \quad (2)$$

$$\frac{d(\eta_k \rho_k)}{dt} = -\eta_k \rho_k \nabla \cdot v \quad (3)$$

$$\eta_k \rho_k \frac{de_k}{dt} = \eta_k \sigma_k : \nabla v - \bar{p} \alpha_k \quad (4)$$

$$\frac{dx}{dt} = v \quad (5)$$

Here  $\rho_k$  corresponds to material density,  $v$  is the common material velocity,  $\eta_k$  serves as an indicator function which keeps track of material volume fraction per zone,  $\alpha_k$  is the rate of change in time of  $\eta_k$ ,  $e_k$  corresponds to specific internal material energy,  $\bar{p}$  is a common thermodynamic pressure, see [8], and  $x$  is the mesh node location. The formulation makes use of the material derivative

$$\frac{d\alpha}{dt} = \frac{\partial \alpha}{\partial t} + v \nabla \cdot \alpha,$$

as well as a total stress tensor for each material,  $\sigma_k$ , which is the sum of the physical stresses and stress due to artificial viscosity.

Building from the MFEM finite element library, MARBL uses arbitrary high-order finite elements to solve Equations (1)–(5). A continuous Galerkin discretization is used for the kinematic variables (mesh velocity and position), while a discontinuous Galerkin discretization is used for the thermodynamic variables (indicator functions, density, and energy). The approach presented here can be viewed as a generalization of the traditional staggered-grid discretization for hydrodynamics (SGH) [9].

The finite element algorithm to solve Equations (1)–(5) requires solving a global linear system for the momentum conservation equation. The key operators in this algorithm are the “mass” matrix and a “force” matrix connecting the kinematic and thermodynamic variables. Computation of a hyper-viscosity term for shock capturing requires an additional “stiffness matrix” operator as described in [10]. The traditional approach of fully assembling corresponding matrices for these operators has been found to have poor scaling characteristics motivating the development of a matrix-free approach.

## 2.2 Mesh Optimization

The mesh optimization algorithm in MARBL is based on the Target-Matrix Optimization Paradigm (TMOP) [11, 12] for high-order meshes. The method is based on node movement and does not alter the mesh topology. Optimal mesh node positions are determined by minimizing a nonlinear variational functional. One of the main advantages of this approach is that all computations can be expressed through finite element operations, making the algorithm well-suited for matrix-free partial assembly implementation as discussed in Section 3.4. The TMOP objective function has the form:

$$F(x) = \int_{\Omega} \mu(T(x)) + \gamma \int_{\Omega} \left( \frac{x - x_0}{d(x_0)} \right)^2, \quad (6)$$

where  $x$  is the vector of mesh positions,  $T(x)$  is a Jacobian matrix that represents the transformation between user-specified target elements and physical elements, and  $\mu$  is a nonlinear function that measures mesh quality. The second term in (6) is used to limit the node displacements based on the maximum allowed displacements  $d(x_0)$  that are defined with respect to the initial mesh position  $x_0$ . The normalization constant  $\gamma$  balances the magnitudes of the two terms; its value is computed once on the initial mesh.

The objective  $F(x)$  is minimized by solving the nonlinear problem  $\partial F(x)/\partial x = 0$  with the Newton’s method. This requires the ability to compute the gradient  $\partial F(x)/\partial x$  and to apply/invert the Hessian  $\partial^2 F(x)/\partial x^2$  of the objective functional. In addition, TMOP has the ability to perform simulation-based mesh adaptivity by querying the values of discrete function  $\xi$  provided by the simulation [13]. For example,  $\xi$  can prescribe local mesh size based on the current state of the physical system. Since  $\xi$  is only defined with respect to the initial mesh, the adaptivity process includes a procedure to solve a pseudo-time advection equation, namely,

$$\frac{d\xi}{d\tau} = v \cdot \nabla \xi, \quad (7)$$

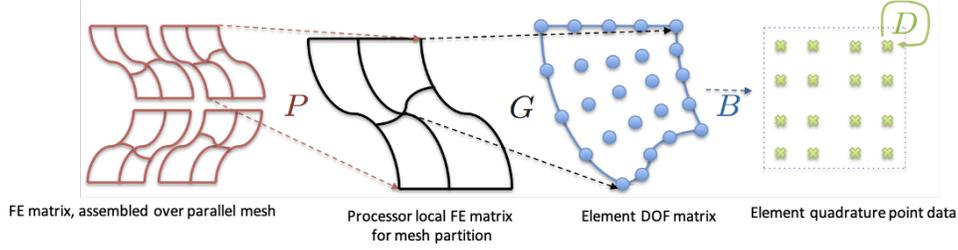


Figure 2: Finite element operator decomposition from global operator to element quadrature data.

which is a method to advect the function  $\xi$  between different meshes, as done in the remap phase of the code, see Section 2.3. All methods described in this section are based on standard finite element operations that can be performed in matrix-free manner.

### 2.3 Remap Phase

Following the mesh optimization, the ALE remap phase is responsible for transferring field quantities from the current Lagrangian mesh to the optimized mesh. The transfer process is posed as the advection of field values between the two meshes over a fictitious time interval,  $\Delta\tau$ . For each material  $k$ , we form and solve the following set of advection equations:

$$\frac{d(\rho v)}{d\tau} = u \cdot \nabla (\rho v) \quad (8)$$

$$\frac{d\eta_k}{d\tau} = u \cdot \nabla \eta_k \quad (9)$$

$$\frac{d\langle \eta \rho \rangle_k}{d\tau} = u \cdot \nabla \langle \eta \rho \rangle_k \quad (10)$$

$$\frac{d\langle \eta \rho e \rangle_k}{d\tau} = u \cdot \nabla \langle \eta \rho e \rangle_k \quad (11)$$

$$\frac{dx}{d\tau} = u. \quad (12)$$

Here  $u$  corresponds to the mesh displacement, and the field variables are the same as those in the Lagrange phase Equation(1)–(5).

For the transfer of the momentum field (8) we follow a standard continuous Galerkin discretization. Since only the action of the operators are needed, we bypass the assembly of the global matrices. For the transfer of the remaining fields, it is imperative that the transfer procedure is conservative, accurate, and is bounds preserving, i.e. the solution remains in a set of admissible bounds. In our target application of remapping multi-material fields, preservation of solution bounds plays a critical role at material interfaces in order to achieve high-fidelity simulations. For ALE remap we use the approach described in [14]. The underlying strategy is to begin with a low-order solution which is guaranteed to be within bounds, and use a separate high-order solution to make corrections to the low order solution. The correction strategy draws on principles from the flux corrected transport (FCT) work [15, 14]. These nonlinear corrections require algebraic modifications to the fully assembled advection matrix in the DG-FCT scheme, so the current implementation cannot be done matrix-free. However, we are actively researching and developing a new DG remap method which retains all of the numerical benefits but can be done in a high-performance matrix-free manner.

## 3 Matrix-Free Partial Assembly

For many applications, the assembly of large global sparse matrices can be a performance bottleneck. Computation with the common CSR matrix format also leads to indirection access, which further limits performance. In most cases, the assembly of the finite element operators in the ALE phases can be avoided as it is only the action of the operator that is needed. We leverage this by using matrix-free algorithms based on the partial-assembly approach described in [2] and available in the MFEM library. Under partial assembly, FEM operators are decomposed into a sequence of operators with cascading scope only requiring quadrature-specific data. For completeness we provide a brief overview of the approach; we refer the reader to [2] for additional details and to the MFEM library [16] for open-source implementation.

### 3.1 Operator Decomposition

Following the nomenclature introduced in [2], for a given global finite element operator ( $A$ ) distributed over a number of processors, we recognize that its assembly or evaluation can be decomposed into a product of operators with cascading scope corresponding to parallel ( $P$ ), mesh ( $G$ ), basis ( $B$ ), and quadrature-specific parts ( $D$ ).

$$A = P^T G^T B^T D B G P$$

Starting with a distributed mesh (as depicted in Figure 2), the role of the  $P/P^T$  operators is to perform the transfer from a global to a processor local domain. The operators  $G/G^T$  perform an additional restriction/prolongation at the element level, while  $B/B^T$  hold the values of the basis functions sampled at quadrature points. For simplicity we assume that test and basis functions are the same, but note that this is not required. Lastly,  $D$  corresponds to element-specific quadrature data. Under the partial-assembly methodology each of these operators are precomputed and stored.

### 3.2 Accelerated Matrix-Vector Evaluation Using Sum-Factorization

Building on the concepts of operator decomposition, we restrict our discussion to quad and hexahedral elements thereby enabling us to exploit the tensor product structure of the basis functions. Continuing with the nomenclature introduced in Section 3.1, we focus on the element local matrix-vector product,

$$y = B^T D B x. \tag{13}$$

Using the element level mass matrix,  $M_E$ , to lead our discussion we express the construction as

$$(M_E)_{i,j} = \sum_{k=1}^{nq} \alpha_k \phi_j(q_k) \phi_i(q_k) |J_E(q_k)| \quad i, j = 1, \dots, nq.$$

Here  $\alpha_k$  corresponds to quadrature weights,  $\phi_i$  are the basis functions sampled at quadrature points,  $q_k$ , and  $|J_E(q_k)|$  is the geometric factor between reference to physical element. When using quad and hex elements, the multi-dimensional basis functions are simply tensor products of their one-dimensional counterparts. In 2D the basis functions have the form

$$\phi_i(q_k) = \phi_{i_1}^{1d}(q_{k_1}) \times \phi_{i_2}^{1d}(q_{k_2}).$$

Recognizing this decomposition enables us to decompose the operator  $B$  in Equation 13 into

$$B = B_{k_1, k_2, i_1, i_2}^{2d} = B_{k_1, i_1}^{1d} B_{k_2, i_2}^{1d} = \phi_{i_1}^{1d}(q_{k_1}) \times \phi_{i_2}^{1d}(q_{k_2}),$$

where  $B^{1d}$  are the one-dimensional basis functions. Finally, by arranging  $D$  as a two dimensional array, computing the action of  $M_E$  may be expressed as a sequence of tensor contractions

$$\begin{aligned} y &= (M_E)_{i_1, i_2, k_1, k_2} x_{i_1, i_2} \\ &= B_{k_1, i_1}^{1d} B_{k_2, i_2}^{1d} (D_E)_{k_1, k_2} B_{k_1, j_1}^{1d} B_{k_2, j_2}^{1d} x_{j_1, j_2}. \end{aligned}$$

These tensor contractions can be implemented as a sequence of dense matrix-matrix multiplications suitable for single instruction multi-data (SIMD) parallelism.

To illustrate the advantages of the partial assembly approach, Table 1 compares storage and algorithmic complexity between full assembly and partial assembly of matrix-vector evaluations per element. As an example, when  $d = 3$  and  $p = 3$ , the storage for 1000 elements would require around 729K floating point values, while the partial assembly approach would only require storing 27K floating point values; similar trends follow when considering floating-point operations for assembly and applying the action of the operator. In summary, partial assembly provides clear advantages in terms of both memory requirements and floating-point operations.

Table 1: Comparison of storage and assembly/evaluation FLOPs required for full and partial assembly algorithms on tensor-product element meshes (quadrilaterals and hexahedra). Here,  $p$  represents the polynomial order of the basis functions and  $d$  represents the number of spatial dimensions. The number of DOFs on each element is  $\mathcal{O}(p^d)$  so the ‘‘sum factorization full assembly’’ and ‘‘partial assembly’’ algorithms are nearly optimal.

| Method                                | Storage               | Assembly                | Evaluation             |
|---------------------------------------|-----------------------|-------------------------|------------------------|
| Traditional full assembly + matvec    | $\mathcal{O}(p^{2d})$ | $\mathcal{O}(p^{3d})$   | $\mathcal{O}(p^{2d})$  |
| Sum factorized full assembly + matvec | $\mathcal{O}(p^{2d})$ | $\mathcal{O}(p^{2d+1})$ | $\mathcal{O}(p^{2d})$  |
| Partial assembly + matrix-free action | $\mathcal{O}(p^d)$    | $\mathcal{O}(p^d)$      | $\mathcal{O}(p^{d+1})$ |

### 3.3 Partial Assembly in Lagrange Phase

Applying the finite element discretization to Equations (1)–(5) leads to the so-called ‘‘Semi-discrete’’ set of equations, which under the appropriate initial conditions and boundary conditions can be solved to a final time with numerical integration techniques.

$$M_v \frac{dv}{dt} = - \sum_k F_k 1 \quad (14)$$

$$\frac{d\eta_k}{dt} = \alpha_k \quad (15)$$

$$\eta_k \rho_k |J| = \eta_{k_0} \rho_{k_0} |J_0| \quad (16)$$

$$M_{\mathcal{E},k} \frac{de_k}{dt} = F_k^T v - c_k \quad (17)$$

$$\frac{dx}{dt} = v \quad (18)$$

The key operators in equations (14)–(18) are the Mass matrix,  $M_v$ , arising from the conservation of momentum in (14), and the force matrix,  $F_k$ , connecting the kinematic and thermodynamic spaces. There is an additional operator required for computing the hyper-viscosity operator as discussed in [10]. Both  $M_v$  and  $F_v$  are global operators and applying their action requires global communication in a distributed memory setting, for example using domain decomposition with MPI. Based on the memory and algorithmic complexity requirements as presented in Table 1, our algorithms bypass the full matrix-assembly, and rely on the partial-assembly approach when applying the action of these operators.

### 3.4 Matrix-Free Mesh Optimization Phase

The computations in this phase utilize a fully matrix-free approach, i.e., all data that is needed at quadrature points is computed on-the-fly and there is no matrix  $D$ . This decision is motivated by the complex structure of the Hessian of  $F(x)$  and the fact that stored quadrature data cannot be reused in consecutive calculations. The major computational kernels in this phase are related to the first and second derivatives of the  $\mu$ -integral in (6). These are given here to provide the relevant context for the discussion:

$$\frac{\partial F_\mu}{\partial x_{a,i}} = \sum_q^{N_q} w_q \det(W) \sum_{m,n=1}^d \frac{\partial \mu}{\partial T_{mn}} \sum_{l=1}^d \delta_{m,a} \frac{\partial w_i}{\partial \bar{x}_l} (W^{-1})_{ln}, \quad (19)$$

$$\frac{\partial^2 F_\mu}{\partial x_{b,j} \partial x_{a,i}} = \sum_q^{N_q} w_q \det(W) \sum_{m,n=1}^d \sum_{o,p=1}^d \left[ \frac{\partial^2 \mu}{\partial T_{op} \partial T_{mn}} \left( \sum_{l=1}^d \delta_{o,b} \frac{\partial w_j}{\partial \bar{x}_l} (W^{-1})_{ln} \right) \left( \sum_{l=1}^d \delta_{m,a} \frac{\partial w_i}{\partial \bar{x}_l} (W^{-1})_{ln} \right) \right]. \quad (20)$$

The above derivatives are taken with respect to the mesh position vector  $x = (x_1 \dots x_d)^T$  where  $d$  is the dimension and each component's expansion is  $x_a = \sum_i x_{a,i} w_i(\bar{x})$ ,  $a = 1 \dots d$ . These integral computations use a standard Gauss-Legendre quadrature rule with  $N_q$  points,  $w_q$  are the corresponding quadrature weights,  $W$  is a  $d \times d$  matrix, and  $\delta$  is the Kronecker delta.

In addition to the standard coefficients and geometric data that is required to compute the integrals, the quadrature-based GPU calculations include the following: the target matrices  $T(x)$  and the mesh quality metric  $\mu(T)$ ; the first derivatives  $\partial\mu/\partial T$ , which are  $d \times d$  matrices; the second derivatives  $\partial^2\mu/\partial T^2$ , which are 4-tensors of dimension  $d \times d \times d \times d$ .

Matrix-free GPU kernels were developed for the computation of the objective (6), the assembly of the nonlinear residual vector (19), and the action of the Hessian (20). Furthermore, diagonal Jacobi preconditioning of the Hessian operator  $\partial^2 F/\partial x^2$  is performed by a matrix-free computation of the diagonal of (20). All of the above kernels utilize the tensor structure of the basis functions  $w_i$  to perform efficient tensor contractions and achieve the optimal operation complexity, see Section 3.2.

GPU support is also provided for TMOP's mesh adaptivity capabilities. In this case, the construction of the matrix  $T(x)$  depends on a discrete function  $\xi$  that must be advected between different meshes through (7). As (7) is a standard advection equation, we make direct calls to MFEM's existing internal partial assembly kernels for the action of the mass and convection operators.

### 3.5 Partial Assembly in Remap Phase

Lastly, for the Remap phase the set of semi-discrete equations take the form of

$$M_v \frac{dv}{d\tau} = K_v v \tag{21}$$

$$M^* \frac{d\eta_k}{d\tau} = K^* \eta_k \tag{22}$$

$$M^* \frac{d\langle \eta \rho \rangle_k}{d\tau} = K^* \langle \eta \rho \rangle_k \tag{23}$$

$$M^* \frac{d\langle \eta \rho e \rangle_k}{d\tau} = K^* \langle \eta \rho e \rangle_k \tag{24}$$

$$\frac{dx}{d\tau} = x_{\tau=1} - x_{\tau=0} \tag{25}$$

where  $M_v$  and  $K_v$  and the “mass” and “convection” matrices corresponding to the continuous Galerkin discretization. As only the action is required, we follow the operator decomposition in Section 3.1 to bypass matrix assembly. The remaining operators,  $M^*$  and  $K^*$ , are algebraically modified “mass” and “convection” matrices from the Algebraic DG-FCT discretization discussed in Section 2.3. Due to algebraic modifications, the operators currently require full matrix assembly. An active area of research is the development of a matrix-free remap framework for MARBL.

## 4 Platform Portability Through MFEM, RAJA, and Umpire

The effort to refactor MARBL's ALE algorithms and data structures was a multi-year process that required a number of cross-project collaboration; it was an iterative process that provided many opportunities for innovation and application-driven development. In this section, we describe the integration of technology and restructuring of MARBL to improve productivity and code performance.

Since MFEM serves as a toolkit for developing algorithms within MARBL, introducing GPU capabilities into MFEM was a natural starting point for MARBL's platform portability. With the release of MFEM 4.0, there were a number of features that provided the foundation for applications to begin leveraging GPUs, including fundamental finite element routines such as computing geometric factors, quadrature interpolators, as well as matrix-free action for a subset of its finite element operators.

MFEM also introduced intermediate layers between memory allocations and kernel execution to support GPU acceleration. The intermediate layer for kernels enables users to choose a backend at run-time; supported backends include the RAJA abstraction layer [3], native OpenMP, CUDA, HIP, and OCCA [17]. Additionally, data structures in MFEM such as vectors, arrays, and matrices are equipped with an intermediate layer, which keeps track of host and device pointers and coordinates transfers between different memory space.

Listing 1: C-style for loop

---

```
double* x; double* y;
double a;
double tsum = 0;
double tmin = MYMAX;

for (int i = 0; i < N; ++i) {
    y[i] += a * x[i];
    tsum += y[i];
    if (y[i] < tmin) { tmin = y[i];}
}
```

---

Listing 2: RAJA-style for loop

---

```
double* x; double* y;
double a;
RAJA::ReduceSum<red_pol, double> tsum(0);
RAJA::ReduceMin<red_pol, double> tmin(MYMAX);

RAJA::forall<exec_policy>(
    RAJA::RangeSegment(0, N),
    [=] (int i) {
        y[i] += a*x[i];
        tsum += y[i];
        tmin.min(y[i]);
    });
```

---

#### 4.1 Development with RAJA in MARBL

After updating to MFEM 4.0, MARBL gained the capability to perform basic memory movement between host and device as well as the ability to offload a select number of MFEM routines to a device. To further enable GPU capabilities, the RAJA abstraction layer was chosen to express computational kernels within MARBL.

With the RAJA abstraction layer, loop bodies are captured in C++ lambdas and template specialization via “execution policies” to select an underlying programming model, i.e., Sequential, OpenMP, CUDA, or HIP. To compliment MFEM’s run-time selectivity, development in MARBL is done through a macro layer that forwards the lambda to the RAJA method that matches the MFEM configured backend (i.e. CPU, OpenMP, CUDA, or HIP). Listings 1, 2 and 3 compare a basic loop expressed using standard C, native RAJA, and the MARBL-RAJA abstraction layer, respectively.

For most of the initial refactoring in MARBL, the RAJA “forall” method was the work horse for developing portable kernels. The simplicity of RAJA forall made it easy to integrate, but the simplicity also limited the exposed parallelism, which limited performance and motivated the search for abstractions that could better utilize the hardware. While more detailed threading was available through the RAJA “kernel” interface, we determined that the associated per-kernel policies were too complex to maintain for our ALE hydro application. Through a co-design effort that leveraged

Listing 3: MARBL-style for loop

---

```
double *x; double *y;
double a;
BlastReduceSum tsum(0);
BlastReduceMin tmin(MYMAX);

BLAST_FORALL(i, 0, N, {
    y[i] += a*x[i];
    tsum += y[i];
    tmin.min(y[i]);
});
```

---

Listing 4: CUDA-style hierarchical parallelism

---

```

DenseTensor A(D1D, Q1D, NE);

void kernel(DenseTensor A) {

    const int z = blockIdx.x;
    __shared__ double s_A[max_D1D][max_Q1D];

    // Load into shared memory
    for(int q=threadIdx.y; q<Q1D; q+=blockDim.y){
        for(int d=threadIdx.x; d<D1D; d+=blockDim.x){
            s_A[d][q] = A(d, q, z);
        }
    }

    // Synchronize threads in a thread team
    __syncthreads();
}

```

---

expertise from the RAJA, MFEM, and MARBL teams, we developed the “RAJA Teams” API, which is available in RAJA v0.12+ as an experimental feature.

## 4.2 Enhancing Performance with RAJA Teams

To develop the RAJA Teams abstraction layer, we used the following guiding principles:

**Simplicity:** The API should be simple to understand and use so that developers can focus on algorithms.

**Expose key features:** For GPUs, we found that the use of device shared memory improved performance and an abstraction is needed so algorithms remain correct when executed on the CPU.

**Lightweight:** Abstraction layers should add minimal overhead and allow the compiler to optimize.

We developed the RAJA Teams framework using a bottom-up approach, in which we built an abstraction layer around loop patterns arising from partial-assembly based algorithms for high-order finite elements. Listing 6 illustrates an example of the RAJA Teams API to load quadrature data from a `DenseTensor` into a shared memory array.

The RAJA Teams API is composed of three core concepts. The first is the **Launch Method**, which is responsible for creating a kernel execution space. The launch method uses a run-time value, `ExecPlace`, and a grid configuration of teams and threads to switch between host or device execution. The teams and threads concept is based on CUDA’s and HIP’s programming model in which computation is performed on a predefined grid composed of threads, which are then grouped into blocks (teams in RAJA). Second, the **Launch Context** in the launch lambda is used for control flow within the kernel, e.g. to perform thread synchronizations as would be done with CUDA’s or HIP’s `__syncthreads()` routine. Lastly, within the kernel execution space, algorithms are expressed in terms of a nested loop hierarchy. Hierarchical parallelism is naturally expressed through RAJA **Loop Methods**, where, in CUDA, ‘outer’ loops are mapped to block-based policies and inner loops may be mapped to thread policies.

In addition, because the utilization of GPU shared memory is an important cornerstone for kernel optimizations, the RAJA teams API includes a macro, `RAJA_TEAM_SHARED`, which expands out to device shared memory or host stack memory, depending on the desired execution space.

To simplify development within MARBL, we also developed a macro layer over the RAJA Teams API similar to our RAJA forall macro wrapper. Listings 4, 5 and 6 compare the native, RAJA, and MARBL kernel APIs. As can be seen in Listing 6, the macros allow us to remove much of the boilerplate code that would be common to all Teams-based kernels over mesh elements and quadrature points.

## 4.3 Enhancing Memory Management with Umpire

During MARBL’s initial GPU refactoring effort, we quickly realized that device memory allocations (e.g. `device mallocs` and `device frees`) were drastically limiting run-time improvements; many kernels used local stack-allocated or fixed lifetime temporary (scratch) memory that introduced expensive global synchronizations in every time step. With

Listing 5: RAJA-style hierarchical parallelism

---

```

using namespace RAJA;
using namespace RAJA::expt;
DenseTensor A(D1D, Q1D, NE);

RangeSegment team_range(0, NE);
RangeSegment q_range(0, Q1D);
RangeSegment d_range(0, D1D);

launch<launch_policy>(
    ExecPlace, // runtime selection
    Grid(Teams(NE), Threads(Q1D, Q1D)),
    [=] RAJA_HOST_DEVICE (LaunchContext ctx) {
// Loop over elements
loop<team_x>(ctx, team_range, [&](int z) {
    RAJA_TEAM_SHARED double s_A[max_D1D][max_Q1D];

// Load into shared memory
loop<thread_y>(ctx, q_range, [&](int q) {
    loop<thread_x>(ctx, d_range, [&](int d) {
        s_A[d][q] = A(d, q, z);
    });
});

// Synchronize threads in a team
ctx.teamSync();
});
});

```

---

Listing 6: MARBL-style hierarchical parallelism

---

```

DenseTensor A(D1D, Q1D, NE);

// Loop over elements
BLAST_FORALL_TEAM(ctx, z, NE, D1D, Q1D, {
    BLAST_SHARED double s_A[max_D1D][max_Q1D];

// Load into shared memory using 2D thread loop
TEAM_LOOP_2D(ctx, d, q, D1D, Q1D, {
    s_A[d][q] = A(d, q, z);
});

TEAM_SYNC;
});

```

---

application run time in mind we decided to move these large localized allocations into permanent memory (allocations lasting for the duration of the simulation). Finding performance-impacting allocations was tedious as our approach relied on being able to match up `cudaMalloc` API calls in the profiler timeline with our in-source annotations.

Although this initial approach eliminated the runtime impacts of repeated allocations, it increased our overall device memory usage, which impacted our ability to run larger problems. Since our initial performant implementation resulted in many temporary buffers being moved into permanent memory, we devised a strategy in which we could reuse memory buffers in parts of the code where short term scratch buffers were required. To assist in the effort, we integrated the Umpire memory resource library [4]. With Umpire, developers can create multiple device allocators and configure them to use either the standard device malloc or special purpose allocators like pools. With pools a large chunk of memory is allocated upfront and subsequent allocations are pulled from the pool. The total size of pools in Umpire can change over time, either when more memory is needed than the pool currently has or when a coalesce operation is requested. Pools are specifically designed to enable fast memory allocations and deallocations and are ideal for temporary or scratch memory that may be needed in kernels.

As previously mentioned, MARBL is built on MFEM and nearly all of its data structures related to device computations are MFEM objects. With the MFEM 4.0 release all MFEM data containers are built on an internally-defined `Memory<T>` memory buffer container. Similar to MFEM’s developer-configurable `Device` policy, the memory policies used are configurable at run-time. Each `Memory` object contains a host pointer and, if a compute device is being used, a device pointer. These `Memory` objects know how to allocate and copy memory between their associated spaces. They also track the validity of the underlying data in each space (host, device, or both), which greatly aids memory synchronization and debugging. As part of our porting effort, we enhanced MFEM’s memory manager to support Umpire allocators. We also introduced finer-grained controls on per-allocation setup to make it easy to use a single permanent allocator by default, but a temporary or scratch allocator when appropriate.

In order to measure memory utilization, we ran a representative 3D shaped charge problem on 12 GPUs with ALE and TMOP. When compared to our initial performant implementation that used 181 GB of device memory, our memory-optimized version used only 97.3 GB, a reduction of nearly 50%, of which 55% (53.6 GB) is now in a shareable temporary allocator. This temporary allocator can be used by other packages in MARBL that, when also configured to use Umpire, further reduce total memory utilization in multi-physics/multi-package simulations.

### 4.3.1 Thread Local Memory

After we started seeing diminishing returns on managing memory spaces and lifetimes, we began to optimize usage of our allocated memory. The CUDA C API makes it easy to query runtime memory usage but it inform on “how” your memory is actually being used. Notably, while much of the device memory is explicitly allocated by MARBL itself, some is allocated by the CUDA runtime, and for large problems we observed that our explicitly allocated device memory was only around 70% of the total allocated device memory. Further investigation revealed that this was due to CUDA runtime allocations for kernels with large amounts of spilled thread-local-memory.

As discussed in Section 4.1, RAJA forall was used as the initial abstraction for device-friendly kernels and while a number of kernels were revisited and ported to RAJA Teams for improved performance, there were still many that existed in their original forall state. A subset of our remaining RAJA forall kernels used large iteration-local stack arrays of fixed size based on our maximum allowable polynomial order. Our maximum order requires much more iteration-local memory than is needed in many simulations and, crucially, even if a simulation is running at a higher order the iteration-local arrays cause a CUDA runtime allocation that is never released; if there is just one kernel that has large thread-local storage requirements, those requirements will effect the memory availability for the entire simulation. Specifically, the CUDA runtime will allocate (bytes per thread)  $\cdot$  (max threads per SM)  $\cdot$  (number of SMs) bytes.

To find the kernels that were causing noticeable increases in device memory usage we intercepted calls to `cudaLaunchKernel` and measured the device memory availability before and after. We then incrementally refactored these kernels to a teams-style shared memory implementation or, for kernels where this wasn’t appropriate, we pre-allocated the thread-local memory in our Umpire temporary pool. Pre-allocating global memory resulted in a roughly 2x slowdown for each kernel modified this way but, for MARBL this was acceptable because the kernels in this group accounted for a negligible amount of runtime both before and after the refactor.

This refactoring removed the use of thread-local arrays owned by the CUDA runtime and freed up more than 2 GB of additional memory per device, a significant improvement considering that our GPUs have only 16 GB of memory.

Figure 3 shows our overall, temporary, permanent, and CUDA runtime allocations from when Umpire was introduced to after we implemented the thread-local-array refactoring.

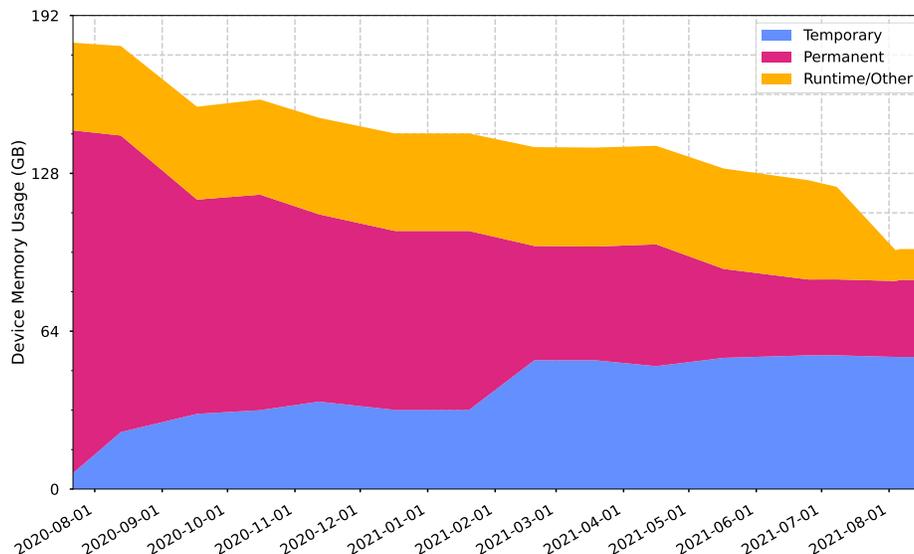


Figure 3: Device memory usage improvements during our GPU port for a 12 GPU 3D Shaped Charge MARBL simulation with ALE+TMOP. Each compute node has four NVIDIA Volta V100 GPUs with 16 GB of memory.

## 5 Cross Platform Scaling Studies

Scaling studies help characterize how well performance scales across large problems and/or compute resources. This section describes a set of cross platform scaling studies comparing MARBL’s performance across three different architectures, including a Commodity Technology Systems (CTS) cluster at LLNL, which we refer to as CTS-1 and two Advanced Technology Systems (ATS) – Sierra, a GPU-based system at LLNL and Astra, an ARM-based system at SNL. These systems have the following characteristics:

- **Sierra** is LLNL’s 125 petaflop ATS system. It has a Mellanox EDR InfiniBand interconnect and 4,320 compute nodes, each of which has two sockets containing 20-core POWER9 processors, 4 NVIDIA Volta V100 GPUs, and 256 GB of memory. Our scaling studies used up to 2,048 nodes, comprising half of the machine, and ran with 4 MPI ranks per node, with 1 GPU per rank.
- **Astra** is a 2.3 petaflop system deployed under the Sandia Vanguard program. It has a Mellanox EDR InfiniBand interconnect and is composed of 2,592 compute nodes, of which 2,520 are user accessible. Each node contains two sockets containing 28-core Cavium ThunderX2 64-bit Arm-v8 processors and 128 GB of memory. Our scaling studies used up to the full machine (2,496 nodes) and utilized RAJA’s OpenMP policy, with 2 MPI ranks per node and 28 threads per rank.
- LLNL’s **CTS-1** clusters are commodity systems with Intel OmniPath interconnect, dual 18-core Intel Xeon E5-2695 2.1GHz processors and 128 gigabytes of memory per node. Our scaling studies used up to 256 nodes, with 36 MPI ranks per node on a system with 1,302 compute nodes and a peak of 3.2 petaflops.

Since we are attempting to characterize and compare performance of our code across different computer architectures, the unit of comparison is an entire compute node of a given architecture. Such studies, which we refer to as *node-to-node* scaling studies, differ from typical scaling studies which measure the baseline performance on a single processor. Node-to-node scaling studies allow natural comparison between a code’s performance on different architectures and its granularity matches how our users typically view their allocations.

Our study includes three types of comparisons: Strong scaling studies use a fixed *global* problem size and polynomial order and vary the number of compute nodes. Weak scaling studies use a fixed *local* problem size and polynomial order per compute node and vary the number of compute nodes. Throughput scaling studies use a fixed compute resource (e.g. 1 compute node) and vary the problem size and polynomial order.

Strong scaling demonstrates the ability of a code to solve a fixed problem faster as one adds additional compute resources, while weak scaling demonstrates the ability of a code to solve a larger problem faster as one adds more resources. In contrast, throughput scaling characterizes the machine performance and demonstrates the balance between parallelism and memory capacity; that is, it measures the work required to saturate memory bandwidth on the compute

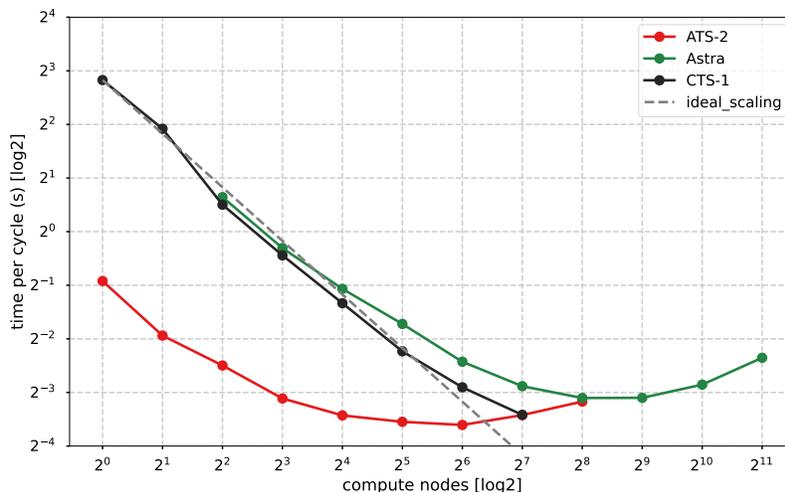


Figure 4: Node-to-node strong scaling for MARBL’s Triple-Pt 3D problem with two levels of uniform refinement and TMOP. Each data point represents a 500 cycle Lagrange hydro simulation, with an ALE remap every 50 cycles. Data is plotted on a log2-log2 scale where the independent axis measure the number of compute nodes while the dependent axis measures the time per cycle. We also plot ideal strong scaling performance for the CTS-1 series: a line with slope  $-1$ .

resources. Commodity systems have modest memory bandwidth and compute resources. As such, they can only accommodate a relatively small amount of parallelism. They are easy to saturate, yielding overall lower performance. On the other hand, GPU-based systems, like Sierra, have very high memory bandwidth and compute resources. As such, they require large amounts of parallel work to saturate, but can yield very high performance.

## 5.1 Strong Scaling

Our strong scaling study ran the 3D triple-point shock interaction (Triple-Pt) benchmark, see for example [7], in ALE mode with TMOP with a total of 112,896 quadratic elements and 7,225,344 quadrature points distributed among varying numbers of compute nodes. Our study starts with one compute node and iteratively doubles the node count in successive runs. For ideal strong scaling, the time per cycle would decrease linearly as we increase the node count. Specifically, as we double the node count, the time per cycle would reduce by a factor of two.

Figure 4 depicts the code’s node-to-node strong scaling performance on Sierra (red), Astra (green) and CTS-1 (black). As can be seen in the figure, MARBL exhibits very good strong scaling on the CPU-based architectures up to 128 CTS-1 and Astra nodes, when communication costs begin to dominate the runtime. As expected, the code exhibits worse strong scaling for higher node counts on the GPU, where, as the node count increases, there is no longer sufficient work to feed each GPU. As such, for this problem, the strong scaling begins to level off when using more than 32 nodes (128 GPU’s).

## 5.2 Weak Scaling

Our weak scaling study was performed on the same Triple-Pt 3D problem as Section 5.1 starting with 2 levels of uniform refinement for our first data point. Each subsequent run in the series maintained the same average number of quadrature points per rank by applying an extra level of uniform refinement (multiplying the global problem size by a factor of eight and increasing the node count by a factor of 8). As such, we have up to four runs in each of our weak scaling series, using 4, 32, 256 and 2048 nodes. We note that 2048 nodes is half of Sierra and 80% of Astra.

Figure 5 depicts the code’s weak scaling performance on Sierra (red), Astra (green) and CTS-1 (black). Since each rank has the same local problem size, ideal weak scaling would maintain the same compute time as we scale out to more nodes. This is not generally achievable due to the increased inter-node communication as we scale out to more nodes. In addition to plotting the compute time per cycle, we also annotate each data point with the weak scaling *efficiency*, measured as the time per cycle of the current data point divided by that of the first data point in the series.

As can be seen in the figure, our code achieves good weak scaling performance as we scale out to the maximum number of nodes in the study. Comparing Sierra performance to the other platforms, we observe 8–16 $\times$  node-to-node speedup for Sierra runs compared to those for the CPU-based platforms.

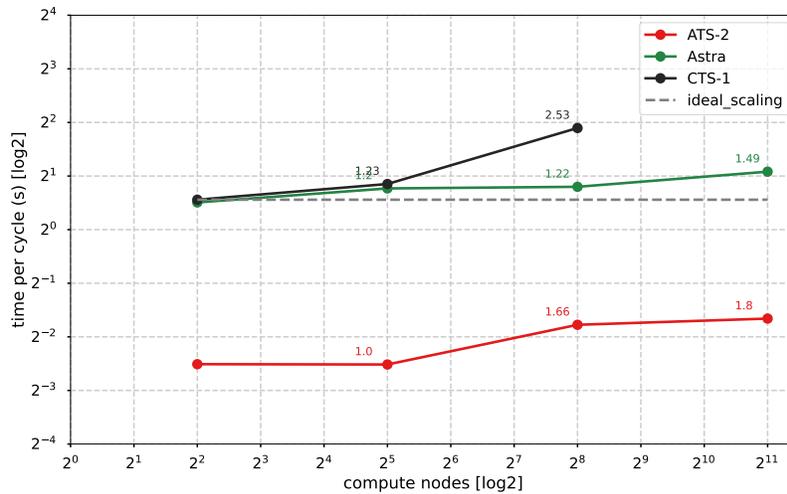


Figure 5: Node-to-node weak scaling for MARBL’s Triple-Pt 3D problem on 4, 32, 256 and 2048 compute nodes. Each data point represents a 500 cycle Lagrange hydro simulation with an ALE remap step every 50 cycles. Data is plotted on a log2-log2 scale where the independent axis is the number of compute nodes and the dependent axis is the time per cycle (in seconds). The annotations indicate the weak scaling efficiency with respect to the 4 node run on that platform. Ideal weak scaling would correspond to an efficiency of 1.0 and all data points in that series would lie on a horizontal line.

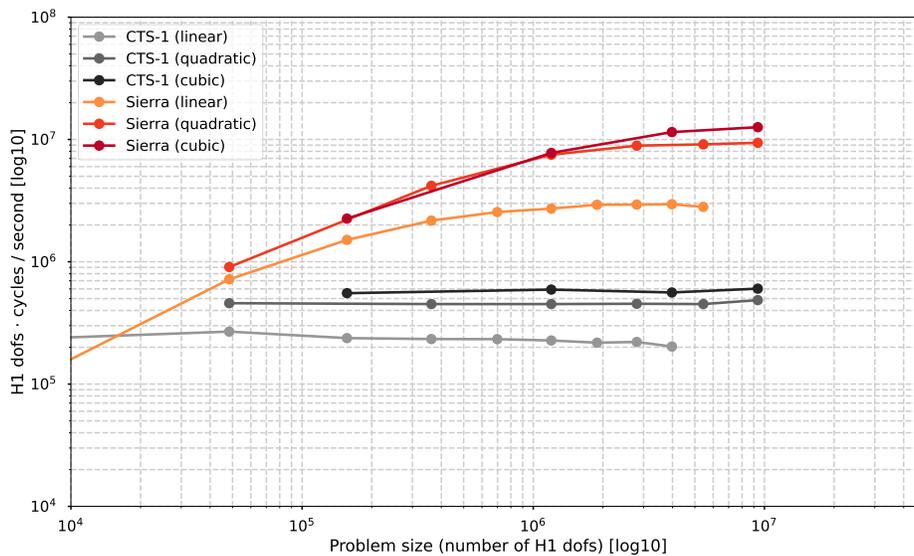
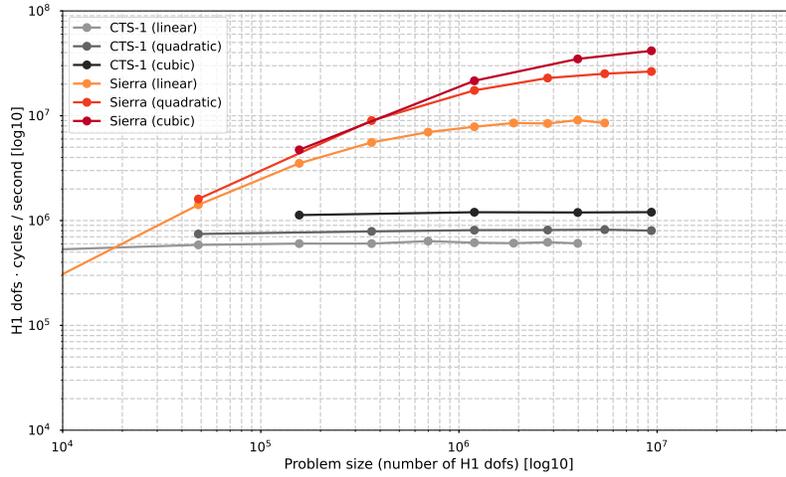


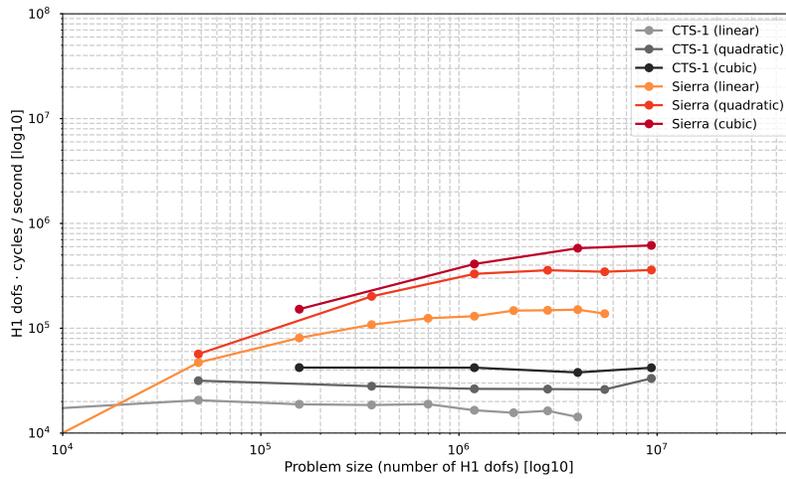
Figure 6: Node to node throughput study for MARBL’s Triple-Pt 3D problem for varying problem size and polynomial order on a single compute node of Sierra (red series) and CTS-1 (gray series). Each data point represents a 500 cycle Lagrange hydro simulation with an ALE remap step every 50 cycles.

### 5.3 Throughput Scaling

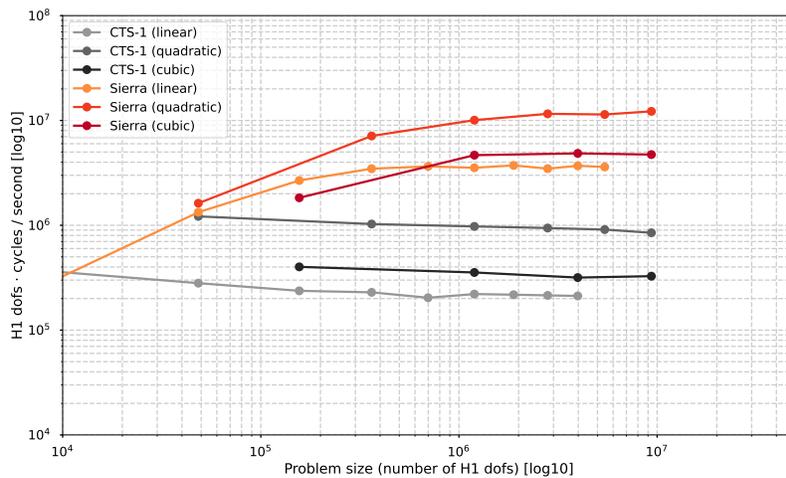
Throughput studies compare the rate an application can process units of work over time at varying problem sizes. In this section we present throughput studies for linear, quadratic, and cubic elements by considering the total number of degrees of freedom for the kinematic variables (H1 basis functions). As can be seen in Figure 6, MARBL’s throughput is significantly higher on GPU-based platforms than on CPU-based ones. Further, throughput increases both with problem size and with increasing order. This is in agreement with results shown for Finite Element benchmark problems in [18] by the Center for Efficient Exascale Discretizations (CEED).



(a) Node-to-node throughput: Lagrange phase



(b) Node-to-node throughput: Mesh Optimization phase



(c) Node-to-node throughput: Remap phase

Figure 7: Break-down of the node-to-node throughput study results from Figure 6 into its three phases: (a) Lagrange (b) Mesh optimization (TMOP) and (c) Remap.

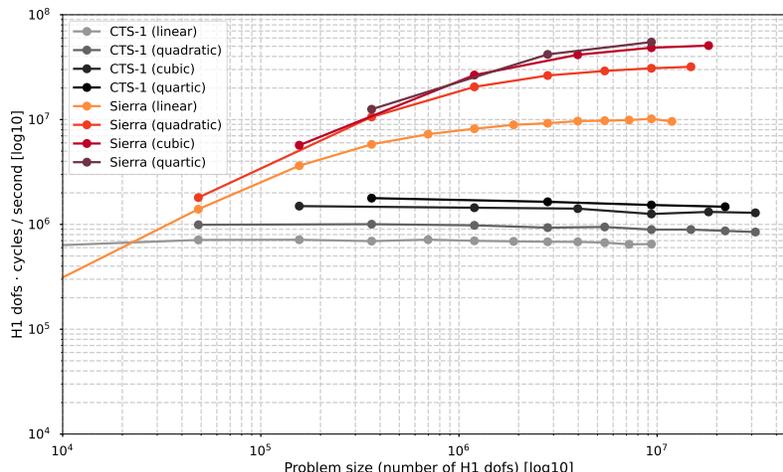


Figure 8: Lagrange throughput study on Sierra for linear, quadratic, cubic and quartic runs.

Breaking out the throughput into the three large phases of MARBL’s ALE hydrodynamics reveals that both the Lagrange phase (Figure 7a) and mesh optimization phase (Figure 7b) achieve even better throughput with increasing order. As can be seen in Figure 8, which is restricted to the Lagrange phase on Sierra, this trend continues for quartic runs. The remap phase in Figure 7c sees a decrease in cubic throughput when compared to quadratic but that is expected in our current implementation which requires full assembly. We are currently investigating a matrix-free remap implementation, to improve our throughput for cubic and higher orders.

## 6 Mini-App Driven Development

Because MARBL was originally developed as a CPU-only code based on full matrix assembly, the introduction of matrix-free methods required significant restructuring and rewriting of its major components. To devise a strategy to refactor MARBL, we developed an MFEM-based miniapp [19], which serves as a proxy application of MARBL’s Lagrangian phase. Laghos builds on the approach described in [7] modeling the principle computational kernels of explicit time-dependent shock-capturing compressible flow in a simplified setting, without the additional complication of physics-specific code. Laghos follows a simplified Lagrangian phase of MARBL which consist of four major computations, namely, inversion of a global CG mass matrix, computation of physics and material specific quadrature point data, computation of artificial viscosity coefficients, and application of a force operator.

As an initial step for Laghos we derived a code structure that separates the finite element assembly operations from the quadrature point-based computations. Such separation is important in the context of matrix-free methods, as it allows the operator splitting that was discussed in Section 3.1. Based on the new code structure, Laghos was used to introduce the first implementation of MARBL’s partially assembled force and mass operators that produced equivalent results as the fully assembled versions, but with improved complexity, see Table 1. Once the matrix-free CPU algorithms were in place, support was added for hardware devices, such as GPUs, and programming models, such as CUDA, OCCA, RAJA and OpenMP, based on MFEM 4.0. Having these different options allowed to perform detailed comparisons of performance and scaling, e.g., see Figure 9, and select the most appropriate options for MARBL, based on the available hardware and the application’s preferences.

The work in the Laghos proxy app was of great benefit to MARBL, as it provided the foundational structure of the code and a baseline version of the major kernels. Consequently, Laghos served as a model for designing more complex kernels for MARBL. A key benefit provided to MARBL from Laghos and MFEM is the ability to drop in replacement kernels for standard finite element operations, for example, the action of the mass operator and diffusion integrator kernels as used in the artificial viscosity calculations [10].

## 7 High-Performance Applications

We now consider two MARBL examples which illustrate the effectiveness of the algorithm and code improvements previously discussed in the context of 3D high-order multi-material ALE simulations on GPUs. In each example, we exercise all three phases of ALE using partial assembly techniques on the GPU including:

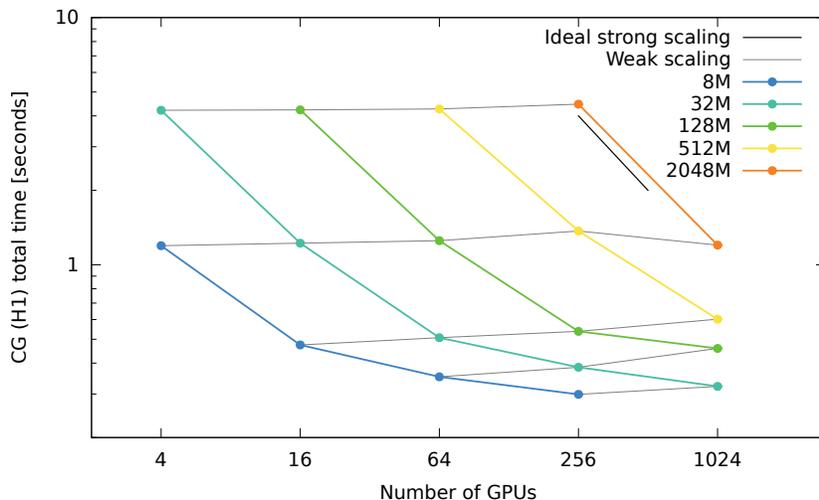


Figure 9: Weak and strong scaling results with Laghos and MFEM-4.1: 2D problem on Lassen, using up to 1024 GPUs.

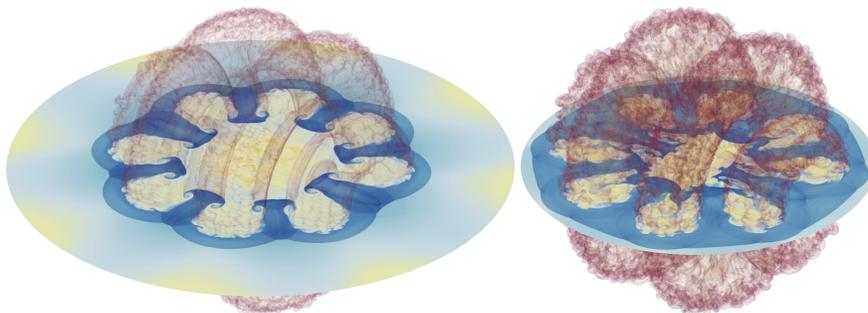


Figure 10: Material density and volume rendering of material interface at two different temporal snapshots for the 3D idealized ICF implosion test run on 256 GPUs of the Lassen machine.

1. High-order Lagrange phase multi-material hydrodynamics on a moving, unstructured, high-order (NURBS) mesh [7], including use of GPU accelerated 3rd party libraries for material equations of state (EOS) evaluation and material constitutive models
2. Non-linear, material adaptive, high-order mesh optimization using the TMOP method
3. High-order continuous (kinematic) and discontinuous (thermodynamic) Galerkin based remap using flux corrected transport (FCT)

## 7.1 Idealized ICF Implosion

Here we consider a 3D two material idealized ICF implosion test problem which is based on the simplified ICF benchmark originally described in [20]. In this variant, we consider a 1/4 symmetry 3D simulation domain using a NURBS mesh and apply a constant velocity drive to the outer radius of the problem (as described in [21]). In addition, to generate vorticity during the implosion, we apply a single mode cosine wave perturbation to the material interface with amplitude  $A = 0.02$  and wavelength  $\omega = 8\pi$ , specifically we modify the material interface radius as  $r_1(\theta) \mapsto r_1(\theta) + A \cos(\omega\theta)$  as described in [10]. This idealized test problem illustrates the impact of both the Rayleigh-Taylor and Richtmyer-Meshkov fluid instabilities in three dimensions at the interface between the two materials as shown in Figure 10. This simulation consists of  $\sim 500 M$  quadrature points and was exercised on 256 GPUs (64 nodes, 4 NVIDIA V100's per node) of LLNL's Lassen machine. To enhance the high-order mesh resolution near the unstable material interface, we employ the material adaptive capabilities of the TMOP mesh optimization phase at the material interface with a 2:1 size ratio as shown in Figure 11.

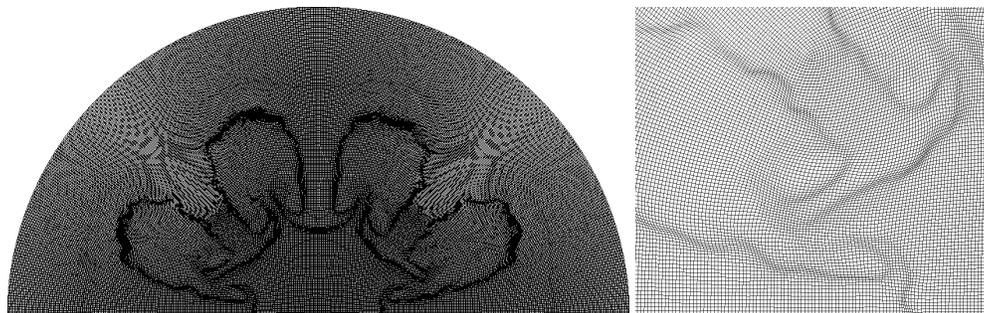


Figure 11: Slice of computational mesh (*left*) with zoomed in view (*right*) highlighting the benefit of material interface adaptive TMOP mesh optimization.

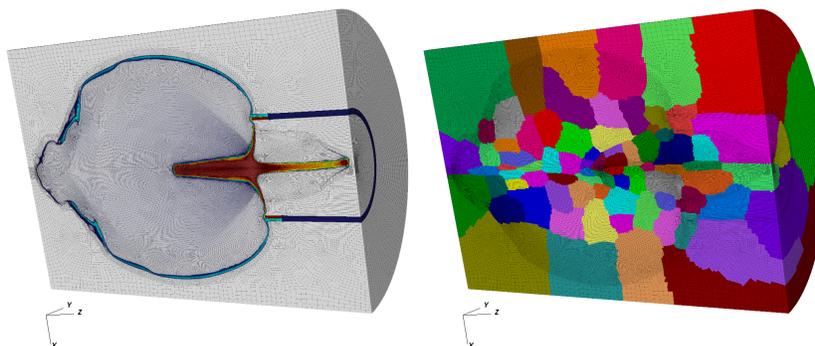


Figure 12: Material density/mesh (*left*) and parallel MPI domains (*right*) for the 3D BRL81a shaped charge run on 96 GPUs of the Lassen machine.

## 7.2 Shaped Charge

The BRL81a shaped charge is a device which focuses the pressures of a high explosive onto a metal “liner” to form a hyper-velocity jet which can be used in many applications, including armor penetration, metal cutting and perforation of wells for the oil/gas industry [22]. Modeling such a device requires multi-material compressible hydrodynamics with general equations of state for the various materials, high-explosive burn and elasto-plasticity.

In Figure 12, we show results of a MARBL calculation of the BRL81a with an off-axis high-explosive detonation point to illustrate 3D effects. We use the Jones Wilkins Lee (JWL) equation of state for the high explosive, tabular equations of state (provided by the LLNL LEOS library) for the rest of the materials and a combination of elastic-perfectly plastic and Steinberg-Guinan strength models for the solid materials in the problem. All of the material constitutive models are running on the GPU either directly in MARBL or as a GPU accelerated 3<sup>rd</sup> party library call. The material model libraries perform their operations in batches of element quadrature points in parallel on the GPU during the Lagrange phase.

This problem is a good stress test of the ALE hydrodynamics in MARBL and running this problem at very high resolutions in 3D is important for studying the hydrodynamics of hyper-velocity jet formation. There are still outstanding questions as to the cause of jet instabilities which are experimentally observed. This simulation consists of  $\sim 115$  M quadrature points and ran on 96 GPUs (24 nodes) of the Lassen machine. To enhance the high-order mesh resolution near the hyper-velocity copper jet, we employ the material adaptive capabilities of the TMOP mesh optimization phase at the copper material with a 2:1 size ratio as shown in Figure 13.

## 8 Conclusion

In this paper we presented a two pronged approach based on improved algorithms and abstractions to refactoring MARBL, a next-gen code at LLNL, for performance and platform portability across commodity and advanced supercomputer architectures such as the GPU-based Sierra cluster at LLNL and the ARM-based Astra cluster at SNL. Recognizing changing trends in computing hardware, we began our endeavor by replacing our matrix based algorithms

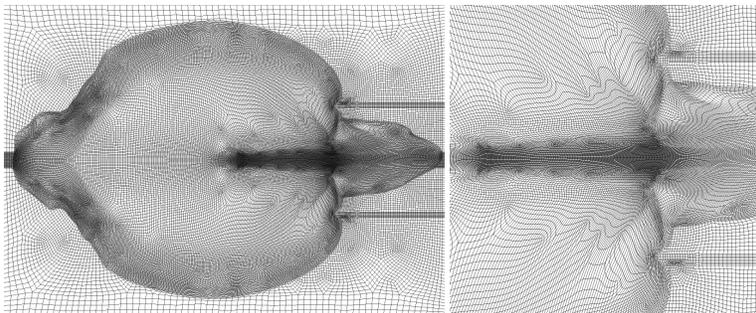


Figure 13: Slice of computational mesh (*left*) with zoomed in view (*right*) highlighting the benefit of material adaptive TMOP mesh optimization.

with matrix-free algorithms wherever possible. The matrix-free algorithms provided improved algorithmic complexity and reduced our memory footprint. To design a single source code which can run on different platforms we integrated the RAJA kernel abstraction layer, Umpire resource memory manager, and MFEM accelerator capabilities. In our efforts we found that extensions and customization of the third-party libraries was crucial for developer productivity and application performance, as was having a proxy app to quickly iterate with different ideas.

To reach improved levels of performance on the MARBL code, we co-designed the RAJA Teams API. The RAJA Teams API provided a simplified interface for expressing and reasoning about hierarchical parallelism and portable access to GPU shared memory. Allocating and managing GPU device memory comes with increased complexity due to the high cost of memory allocations and limited GPU memory. By combining MFEM’s memory manager with Umpire, we were able to have MFEM objects quickly draw and release memory from memory pools. The use of memory pools enabled different physics algorithms of MARBL to share memory thereby reducing the total amount of permanent memory required across the whole simulation.

We also demonstrated the code’s performance on several example problems and with large scale “node-to-node” throughput, weak, and strong scaling studies across different supercomputer architectures. Our abstraction model, based on RAJA, Umpire and MFEM, enabled us to run simulations using the same codebase on GPU-based architectures (with MPI+CUDA) as we ran on CPU-based advanced architectures (using MPI+OpenMP) and on commodity platforms using only MPI.

In summary, by incorporating new algorithms and powerful abstraction layers we were able to achieve a high degree of portability and performance with the MARBL code.

## Acknowledgments

This work performed under the auspices of the U.S. Department of Energy by Lawrence Livermore National Laboratory under Contract DE-AC52-07NA27344. LLNL-JRNL-829593. We would like to thank Veselin Dobrev for support on the integration of MFEM GPU capabilities into MARBL, and the RAJA & Umpire teams for their feedback in software integration and extensions.

## References

- [1] CEED. Center for Efficient Exascale Discretizations, Exascale Computing Project. <https://ceed.exascaleproject.org>, 2021.
- [2] R. Anderson, J. Andrej, A. Barker, J. Bramwell, J.-S. Camier, J. Cerveny, V. Dobrev, Y. Dudouit, A. Fisher, Tz. Kolev, W. Pazner, M. Stowell, V. Tomov, I. Akkerman, J. Dahm, D. Medina, and S. Zampini. MFEM: A modular finite element methods library. *Computers & Mathematics with Applications*, 81:42–74, 2021.
- [3] D. Beckingsale, J. Burmark, R. Hornung, H. Jones, W. Killian, A. Kunen, O. Pearce, P. Robinson, B. Ryujin, and T. Scogland. RAJA: Portable performance for large-scale scientific applications. In *2019 IEEE/ACM international workshop on performance, portability and productivity in HPC (p3hpc)*, pages 71–81. IEEE, 2019.
- [4] D. Beckingsale, M. Mcfadden, J. Dahm, R. Pankajakshan, and R. Hornung. Umpire: Application-focused management and coordination of complex hierarchical memory. *IBM Journal of Research and Development*, 64(3/4):00–1, 2019.

- [5] R. Anderson, A. Black, B. Blakeley, R. Bleile, J.-S. Camier, J. Ciurej, A. Cook, V. Dobrev, N. Elliott, J. Grondalski, C. Harrison, R. Hornung, Tz. Kolev, M. Legendre, W. Liu, W. Nissen, B. Olson, M. Osawe, G. Papadimitriou, O. Pearce, R. Pember, A. Skinner, D. Stevens, T. Stitt, L. Taylor, V. Tomov, R. Rieben, A. Vargas, K. Weiss, D. White, and L. Busby. The Multiphysics on Advanced Platforms Project. Technical report, LLNL-TR-815869, LLNL, 2020.
- [6] R. Anderson, V. Dobrev, Tz. Kolev, R. Rieben, and V. Tomov. High-order multi-material ALE hydrodynamics. *SIAM Journal on Scientific Computing*, 40(1):B32–B58, 2018.
- [7] V.A. Dobrev, T.V. Kolev, and R.N. Rieben. High-order curvilinear finite element methods for Lagrangian hydrodynamics. *SIAM J. Sci. Comp.*, 34(5):B606–B641, 2012.
- [8] Veselin A. Dobrev, Tz. Kolev, Robert N. Rieben, and Vladimir Z. Tomov. Multi-material closure model for high-order finite element Lagrangian hydrodynamics. *Int. J. Numer. Methods Fluids*, 82(10):689–706, 2016.
- [9] A. J. Barlow, P-H. Maire, W. J. Rider, R. N. Rieben, and M. J. Shashkov. Arbitrary Lagrangian-Eulerian methods for modeling high-speed compressible multimaterial flows. *Journal of Computational Physics*, 322(C):603–665, 2016.
- [10] P. Bello-Maldonado, V. Tomov, Tz. Kolev, and R. Rieben. A matrix-free hyperviscosity formulation for high-order ALE hydrodynamics. *Computers & Fluids*, 205:104577, 2020.
- [11] V. Dobrev, P. Knupp, Tz. Kolev, K. Mittal, and V. Tomov. The Target-Matrix Optimization Paradigm for high-order meshes. *SIAM J. Sci. Comp.*, 41(1):B50–B68, 2019.
- [12] P. Knupp. Introducing the target-matrix paradigm for mesh optimization via node-movement. *Engineering with Computers*, 28(4):419–429, 2012.
- [13] Veselin A. Dobrev, Patrick Knupp, Tz. Kolev, Ketan Mittal, Robert N. Rieben, and Vladimir Z. Tomov. Simulation-driven optimization of high-order meshes in ALE hydrodynamics. *Comput. Fluids*, 208:104602, 2020.
- [14] R. Anderson, V. Dobrev, Tz. Kolev, and R. Rieben. Monotonicity in high-order curvilinear finite element Arbitrary Lagrangian–Eulerian remap. *International Journal for Numerical Methods in Fluids*, 77(5):249–273, 2015.
- [15] D. Kuzmin, R. Löhner, and S. Turek. *Flux-corrected transport: Principles, algorithms, and applications*. Springer, 2012.
- [16] MFEM. MFEM: Modular finite element methods [Software]. [mfem.org](http://mfem.org), 2021.
- [17] D. S. Medina, A. St-Cyr, and T. Warburton. OCCA: A unified approach to multi-threading languages. *arXiv preprint arXiv:1403.0968*, 2014.
- [18] V. Dobrev, J. Dongarra, J. Brown, P. Fischer, A. Haidar, I. Karlin, T. Kolev, M. Min, T. Warburton, T. Ratnayaka, S. Tomov, and V. Tomov. Identify initial kernels, bake-off benchmark problems and miniapps. Technical report, Tech. Rep. CEED-MS6, Exascale Computing Project, June 2017.
- [19] Laghos: High-order Lagrangian hydrodynamics miniapp. <https://github.com/CEED/laghos>, 2021.
- [20] S. Galera, P.-H. Maire, and J. Breil. A two-dimensional unstructured cell-centered multi-material ALE scheme using VOF interface reconstruction. *Journal of Computational Physics*, 229(16):5755–5787, 2010.
- [21] V. A. Dobrev, T. E. Ellis, Tz. V. Kolev, and R. N. Rieben. High order curvilinear finite elements for axisymmetric Lagrangian hydrodynamics. *Computers & Fluids*, 83:58–69, 2013.
- [22] W. Walters. A brief history of shaped charges. volume 1, pages 3–10. 24th International Symposium on Ballistics, New Orleans, LA, 2008.