

# Parallel Multiphysics Simulation for the Stabilized Optimal Transportation Meshfree (OTM) Method

Sandeep Kumar<sup>1</sup>, Pierre Gosselet<sup>4</sup>, Dengpeng Huang<sup>2</sup>,  
Christian Weïßenfels<sup>3</sup>, Peter Wriggers<sup>1</sup>

<sup>1</sup> Institute of Continuum Mechanics, Leibniz Universität Hannover, 30167 Hannover, Germany

kumar@ikm.uni-hannover.de, wriggers@ikm.uni-hannover.de

<sup>2</sup> Institute of Applied Dynamics, Friedrich-Alexander-Universität

Erlangen-Nürnberg, 91058 Erlangen, Germany

dengpeng.huang@fau.de

<sup>3</sup> Institute of Materials Resource Management,

Data-driven Computational Materials Science and Engineering,

Universität Augsburg, 86159 Augsburg, Germany

christian.weissenfels@mrm.uni-augsburg.de

<sup>4</sup> Université de Lille, CNRS, Centrale Lille, UMR 9013 – LaMcube – F-59000 Lille, France

pierre.gosselet@univ-lille.fr

10.1016/j.jocs.2022.101739

## Abstract

This paper presents a parallel implementation for the Optimal Transportation Meshfree (OTM) method on large CPU clusters. Communications are handled with the Message Passing Interface (MPI). The Recursive Coordinate Bisection (RCB) algorithm is utilized for domain decomposition and for implementing dynamic load-balancing strategy. This work involves three new concepts to reduce the computational efforts: Dynamic halo regions, Efficient data management strategies for ease of addition and deletion of nodes and material points using advanced STL container, and nearest neighborhood communication for detection of neighbors and communication. Also, Linked Cell approach has been implemented to further reduce the computational efforts. Parallel performance analysis is investigated for challenging multiphysics applications like Taylor rod impact and serrated chip formation process. Adequate scalability of parallel implementation for these applications is reported.

**Keywords:** Optimal Transportation Meshfree Method, Parallel Computing, MPI, Dynamic

## 1 Introduction

Processes involving large deformations, such as Additive Manufacturing or cutting, present a challenge while modeling with standard approximation tools like the Finite Element Method. If the Lagrangian description is used, these large deformations can result in severe mesh distortions. In this case adaptive remeshing procedures and mapping of state variables from one configuration to another are required. Inefficient computations and accumulated numerical errors can result. Alternatively, meshfree methods seem quite adapted to such simulations. For instance, the Smoothed Particle Hydrodynamics (SPH) has shown big potential. A more recent solution scheme is the Optimal Transportation Meshfree (OTM) method [Li et al., 2010]. This method is motivated by the Optimal Transportation Theory [Villani, 2013] integrated with local maximum entropy (LME) meshfree interpolation [Arroyo and Ortiz, 2006] and material point sampling method [Sulsky et al., 1994; Wessels et al., 2019, 2018]. A detailed analysis on the LME meshfree interpolation applied to metal forming process is provided in [Cueto and Chinesta, 2013]. The advantage of the OTM method is the similar transition of the FEM to a meshfree method [Li et al., 2010].

Realistic simulations of many engineering applications problems require large-scale computations. Computation on HPC clusters needs efficient and scalable codes. In order to utilize the full potential of the computing power of multi-core architectures, it is necessary to exploit both the intra and inter-node parallelism. Different parallel programming models are developed over the years. An overview can be found in [Prims et al., 2019]. Currently, existing parallel programming models are based on distributed memory and shared memory platforms. Message Passing Interface (MPI) is

most widely used standard paradigm on distributed memory platforms but it can also be applied to shared memory nodes. Other parallelization approaches exist, such as, OpenMP (for shared memory platforms), CUDA and OpenCL (for graphics processing units (GPUs)). Coupling our approach with OpenMP is possible and this will be the subject of future work.

Development of MPI based software requires carefully thought strategies for data-partitioning and data communication. Data-partitioning refers to the process of dividing the problem domain into smaller subdomains. Subdomain geometry affects the scalability since it is associated with equivalent work load through load-balancing. In order to achieve high scalability, subdomains are expected to contain same amount of work (load balancing) while minimizing the need for communications. Several approaches exist in field of non-overlapping domain decomposition methods to solve the challenging mechanical heterogeneous problems on massively parallel architectures. A new parallel mesh generation method has been developed by [Gharbi et al., 2021] which leads to subdomains with shape well-suited for Schur based domain decomposition methods, such as FETI [Farhat and Roux, 1991] and BDD [Mandel, 1993] solvers. Another domain decomposition method, Orthogonal Recursive Bisection (ORB) algorithm has been implemented by [Oger et al., 2016; Yang et al., 2020] which has been shown to lead to scalable results at large processor numbers. Also, due to the distributed nature of the method, duplication of the data is required for communication, resulting in the increased overall memory requirement.

Generally when a parallel program is run on several processes simultaneously, there are data dependencies between the tasks. A process might need intermediate results in order to carry out its computations and this intermediate result could be located on a different process. Hence, bottlenecks occur which slow down the computation. The MPI library provides different communication primitives: point-to-point and collective communication. One way to maximize the performance of parallelization is to reduce the overheads due to communication operations. Overhead is defined as *the length of time that a processor is engaged in the transmission or reception of each message; during this time, the processor cannot perform other operations* [Culler et al., 1993]. Collective communication operations, introduced in the latter versions of MPI, have been a key concept used in large scale parallel applications to minimize the communication overheads [Barigou and Gabriel, 2017; Barigou et al., 2015]. Although they are widely used due to their increased productivity and performance, there are some limitations. Due to the dependencies on all the processes of a communicator, there exist scalability issues and conventional collectives support limited communication patterns, such as, broadcast and all-to-all, see [Ghazimirsaeed et al., 2020]. In order to address these issues, Neighborhood Collectives, introduced by the MPI 3.0, provide an alternative to the users to define arbitrary communication patterns. This can be used to implement nearest neighbor collective operations where each process interacts with only a small neighborhood of processes. Neighborhoods can be described either by Cartesian neighborhoods or by general communication graphs, for more details see [Hoeffler et al., 2011] and [Message Passing Interface Forum, 2015].

When performing computations within each subdomain, all the necessary information should be available in the same process. But, for the subdomain's boundary, some of the required information will be located on other processes. Hence, a communication pattern is required, the most commonly used is halo regions. At every computation step, the halo regions are exchanged with the neighboring processes so that every process can access to the necessary information. The goal of the halo regions is to locally replicate the domain residing in other processes. At every computation step, when processes communicate with their neighbors, performance of any process depends strongly on the performance of its neighbors. This can result in delays [Laoidé-Kemp, 2015]. Additionally, the probability for delays can increase with the number of processes.

Large scale computations requires an adaptive code to run efficiently on distributed memory systems. Good data management and domain decomposition are critical parameters. Parallelizing OTM shares many common issues with discrete element methods [Visseque et al., 2013]. Some of the cumbersome tasks during simulation using the OTM method in a parallel environment involves modifying, deleting and adding particles in a subdomain, adjusting the subdomain partitioning dynamically and performing migration of particles to maintain load balance during the simulations. These tasks require flexible and efficient data management scheme. In mesh-based methods, flexible and efficient data management schemes for parallel systems have been implemented for adaptive hp finite element method [Laszloffy et al., 2000], and for simulation tool for geophysical mass flows [Patra et al., 2005]. In meshfree methods, [Cao et al., 2017] developed data management strategies for a MPI parallel implementation of the SPH method to simulate volcano plumes. [Ferrari et al., 2009] used a flexible way in linked lists using pointers so that particles can be deleted or added during the simulation. Similar approach for modifying pointer-based information has been adopted in the current work.

The following three reasons motivate the work presented in this paper. First, while parallelization approach to OTM method has been implemented by [Li et al., 2014], efficient implementation strategies have been presented by introducing communication for both nodal and material point halo regions for localized updates within every subdomain. Second, with the use of improved data structures for halo regions, flexibility have introduced to handle variable workloads (dynamic halo regions). This is helpful when new nodal or material point quantities are added into the communication. Finally,

with the use of nearest neighborhood communication for neighbor detection and communication, the communication costs are reduced even though total halo particles increases with increase in number of subdomains.

In this work, computational strategies are proposed for parallel processing of OTM Method using MPI (Message Passing Interface) for scalability on large-scale computer clusters. Mechanisms are presented for efficient addition or removal of nodes and material points from their corresponding influence and support domains respectively, thereby reducing computational overheads. Hence, storage issues related to the fixed-size arrays are eliminated. Dynamic halo region is implemented that can handle variable workloads. Both nodes and material points within every subdomain and their corresponding influence and support domains are managed by STL map which can ensure quick and flexible access and modifications. In order to ensure good static and dynamic load balance, Recursive Coordinate Bisection (RCB) algorithm, a Cartesian based decomposition method is used for both static and dynamic decomposition (dynamic load balancing). The RCB decomposition method facilitates good scalability by ensuring minimum interfacial surface area between the sub-domains. Parallel decomposition of spatial domain is carried out in such a way that each subdomain is physically compact and the computations can be performed locally at each process. The flexibility of our data access methodology, data structures, dynamic halo regions enables efficient parallel implementation of OTM method. To reduce global communications, nearest neighbor communication operations are implemented using MPI collectives [Message Passing Interface Forum, 2015].

The outline of this paper is as follows. Section 2 contains a brief description of the OTM Method. Section 3 discusses the data structures, as well as the parallel implementation for domain decomposition and communication. Results from applying the presented method to the Taylor rod impact and the serrated chip formation process are presented in Section 4.

## 2 Optimal Transportation Meshfree Algorithm

The Optimal Transportation Meshfree (OTM) method is an Updated Lagrangian formulation which can be used for both solid and fluid flow simulations based on [Li et al., 2010]. OTM method can be viewed as an evolution of finite element method because the spatial domain under investigation is discretized by two types of points. The material points are used as integration points, where quantities like stress, strain, density, etc., are determined. At the nodal points, the primary variables are computed by solving discretized equations of motion. A search algorithm establishes the connectivity between nodes and material points during the computation: the nodes associated with a material point form its support domain, whose shape is in general arbitrary. The material point values are determined with the help of basis functions. In general, maximum entropy shape functions [Arroyo and Ortiz, 2006] are used. Since OTM method have some shortcomings, the stabilized formulation due to [Weißenfels and Wriggers, 2018] is used. The whole algorithm is sketched in Algorithm 1.

---

### Algorithm 1 Algorithmic implementation of a computation step in OTM

---

**Require:** Initial nodal set and material point set

1. Compute local mass matrix and local nodal force vector
  2. Update primary variables and nodal coordinates
  3. Update material point coordinates
  4. Constitutive updates at material point
  5. Search Algorithm to update support domains
  6. Recompute shape functions
- 

### 2.1 Update of Primary Variables

As shown in [Weißenfels and Wriggers, 2018], the OTM method can also be derived from the weak form and the formulation is made with respect to the current configuration, balancing the virtual work at the boundary with virtual work inside of the body and the inertia term.

$$\int_{\Omega} \delta \mathbf{u} \cdot \rho \ddot{\mathbf{u}} dv + \int_{\Omega} \text{grad } \delta \mathbf{u} : \boldsymbol{\sigma} dv = \int_{\Omega} \delta \mathbf{u} \rho \cdot \hat{\mathbf{b}} dv + \int_{\partial\Omega} \delta \mathbf{u} \cdot \hat{\mathbf{t}} da, \quad (1)$$

where the displacements  $\mathbf{u}$  are the primary variables. The Cauchy stress tensor, specific body force and density correspond to  $\boldsymbol{\sigma}$ ,  $\hat{\mathbf{b}}$  and  $\rho$  respectively. The surface traction  $\hat{\mathbf{t}}$  is prescribed at the Neumann boundary.

The support domain is defined as the domain around each material point, containing nearest nodes in its neighborhood (Fig 1). This domain is updated at every computation step by applying a suitable search algorithm. The shape functions  $N_I(\mathbf{x}_{p_n})$  are continuously updated as the OTM method has the usual structure of updated Lagrangian procedures. At each material point, the test

function and the displacements are approximated through shape functions  $N_I(\mathbf{x}_{pn})$  and nodal values within its support domain

$$\mathbf{u}_p(\mathbf{x}_{pn}) = \sum_{I=1}^{n_{np}} N_I(\mathbf{x}_{pn}) \mathbf{u}_I, \quad \delta \mathbf{u}_p = \sum_{I=1}^{n_{np}} N_I(\mathbf{x}_{pn}) \delta \mathbf{u}_I, \quad \text{grad } \delta \mathbf{u}_p = \sum_{I=1}^{n_{np}} \mathbf{B}_I(\mathbf{x}_{pn}) \delta \mathbf{u}_I, \quad (2)$$

where  $n_{np}$  specifies the number of nodes in the support domain of each material point at current computation step. The matrix  $\mathbf{B}_I(\mathbf{x}_{pn})$  contains the derivatives of shape functions at node  $I$ . In contrast to the FEM, overlapping of support domains is allowed in OTM method. In problems of large deformations, non-admissible nodal distributions can be eliminated by the update of support domains at every time or load step.

Using (2), the approximation of (1) can be transformed into an algebraic equation using an assembly procedure

$$\left[ A_{p=1}^{n_{mp}} \sum_I \sum_J^{n_{np}} N_I(\mathbf{x}_p) \mathbf{1} N_J(\mathbf{x}_p) m_p \right] \cdot \ddot{\mathbf{u}} = A_{p=1}^{n_{mp}} \sum_I \left[ N_I(\mathbf{x}_p) \hat{\mathbf{b}}_p m_p - \mathbf{B}_I(\mathbf{x}_p) \sigma_p v_p \right]. \quad (3)$$

where  $\ddot{\mathbf{u}}$  is the global nodal acceleration vector,  $n_{mp}$  is the total number of material points in the body,  $m_p$  is the mass at the material point  $p$  and  $v_p$  is its volume in the current configuration. In order to guarantee that the conservation of the mass during the computation, the mass of a material point is assumed to be constant.

Using the explicit central difference time integration scheme and the concept of lumped mass matrix, the equilibrium of the body is transformed into a set of independent nodal equilibrium equations. This step is equivalent to the Finite Element Method framework given in [Bathe, 2006], for instance

$$m_{I_n} \frac{\Delta \mathbf{u}_{I_{n+1}} - \Delta \mathbf{u}_{I_n}}{(\Delta t)^2} = \mathbf{p}_{I_n} + \mathbf{r}_{I_n}. \quad (4)$$

The boundary forces,  $\mathbf{p}_{I_n}$ , are prescribed only at the nodes of the Neumann boundary. The nodal residual vector  $\mathbf{r}_{I_n}$  and the nodal mass  $m_{I_n}$  are given as

$$m_{I_n} = \sum_p^{n_{mp}^I} N_{I_n}(\mathbf{x}_{pn}) m_p, \quad \mathbf{r}_{I_n} = \sum_p^{n_{mp}^I} \left[ N_{I_n}(\mathbf{x}_{pn}) \hat{\mathbf{b}}_p m_p - \mathbf{B}_I(\mathbf{x}_{pn}) \sigma_p v_p \right] \quad (5)$$

where  $n_{mp}^I$  is the number of material points in the influence domain of Node  $I$ . The corresponding material points within each influence domain can be determined from the support domain directly, without any need of additional search algorithm, see Fig 1.

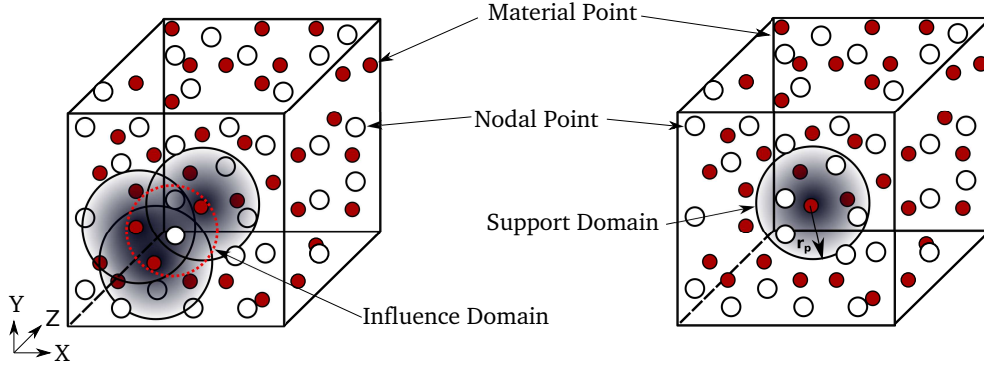


Figure 1: Influence Domain of a node and Support Domain of a material point

By solving (4), the nodal position vector can be updated from the displacement increments of the next computation step

$$\mathbf{x}_{I_{n+1}} = \mathbf{x}_{I_n} + \Delta \mathbf{u}_{I_{n+1}} \quad (6)$$

The integral is evaluated by a single point in each support domain. LME shape functions (see Section 2.3) are rational exponential functions. Hence, the total material points are not enough to accurately integrate the weak form in (1). A stabilization term is added to the nodal residual vector to penalize the inaccurate behavior due to underintegration within every support domain, see [Weifenfels and Wriggers, 2018]:

$$r_{I_n-stab} = r_{I_n} - \varepsilon \sum_p^{n_{mp}^I} N_{I_n}(\mathbf{x}_{pn}) \mathbf{e}_{I,pn} \quad (7)$$

where,  $\varepsilon$  is the penalty parameter,  $\mathbf{e}_{I,pn}$  is the error due to underintegration

$$\mathbf{e}_{I,pn} = \frac{\mathbf{x}_{In} - \mathbf{x}_{pn} - (\tilde{\mathbf{x}}_{In} - \tilde{\mathbf{x}}_{pn})}{\|\mathbf{x}_{In-1} - \mathbf{x}_{pn-1}\|}, \quad \tilde{\mathbf{x}}_{In} - \tilde{\mathbf{x}}_{pn} = \Delta \mathbf{F}_{pn} [\mathbf{x}_{In-1} - \mathbf{x}_{pn-1}] \quad (8)$$

and  $\Delta \mathbf{F}_{pn}$  is the increment of the deformation gradient, as computed in (10).

## 2.2 Update of Kinematic Quantities

The position vector of a material point at the next computation step is updated by multiplying the shape functions at the current time step with the nodal coordinates of the next time step

$$\mathbf{x}_{pn+1} = \sum_I^{n_{np}} N_{In}(\mathbf{x}_{pn}) \mathbf{x}_{In+1} \quad (9)$$

The deformation gradient at the next computation step,  $n+1$ ,

$$\mathbf{F}_{pn+1} = \Delta \mathbf{F}_{pn+1} \mathbf{F}_{pn} \quad (10)$$

is updated in terms of the current value of the deformation gradient,  $\mathbf{F}_{pn}$ , and the increment of the deformation gradient is

$$\Delta \mathbf{F}_{pn+1} = \mathbf{1} + \sum_I^{n_{np}} \frac{\partial N_{In}(\mathbf{x}_{pn})}{\partial \mathbf{x}} \Delta \mathbf{u}_{In+1}, \quad (11)$$

Accordingly, the volume and density at each material point is also updated by  $\Delta \mathbf{F}_{pn+1}$

$$v_{pn+1} = \det(\Delta \mathbf{F}_{pn+1}) v_{pn} \quad (12)$$

$$\rho^{n+1} = \frac{m_p}{v_{pn+1}} \quad (13)$$

## 2.3 Local Max-Ent shape functions

In meshfree methods, the polynomial basis functions, which are normally used within the finite element framework, are not appropriate. In OTM method, local maximum entropy (LME) [Arroyo and Ortiz, 2006] approximation function is used which has to be determined for an arbitrary number of nodes within the support domain. The LME shape functions possess weak Kronecker- $\delta$  property at the boundary and it is fulfilled only on convex boundaries, see [Li et al., 2010]. Also, the LME shape functions does not fulfill either the first order completeness or the partition of unity condition. In order to achieve convergence to the correct solution of the equation of motion, computational algorithms should fulfill these basic conditions, see [Hughes, 1987] and [Belytschko et al., 1998].

The LME shape functions has an exponential ansatz and it belongs to the class of radial basis functions. First order completeness condition is enforced using Lagrangian multiplier method and the partition of unity condition is enforced through normalization

$$N_I(\mathbf{x}_p) = \frac{Z_I(\mathbf{x}_p)}{Z}, \quad Z_I = \exp(-\beta |\mathbf{x}_p - \mathbf{x}_I|^2 + \lambda(\mathbf{x}_p - \mathbf{x}_I)), \quad Z = \sum_I^{n_{np}} Z_I \quad (14)$$

where  $\lambda$  is a Lagrangian multiplier, which is determined by solving  $\sum N_I(x_p)(x_p - x_I) = 0$  using Newton- Raphson algorithm. The parameter  $\beta$  is calculated as  $\beta = \frac{\gamma}{h^2}$ , where  $\gamma$  controls the degree of locality of LME shape functions and it should be in the range of 0.8 to 4, and  $h$  is the characteristic nodal spacing.

## 3 Software Design

The method is written for use on multi-CPU architectures. The parallel codes are written in C++ (would also be possible with Fortran) and make use of its object-oriented features. The code utilizes the Message Passing Interface (MPI) for communication and synchronization between processes. MPI is a standard paradigm for implementing parallel software in distributed memory platforms, [Balaji et al., 2010; Plimpton and Devine, 2011]. In order to exchange message and manage processes, MPI provides a collective set of library routines. It is generally used in high end computing applications involving intensive calculations [Notay and Napov, 2015].

The approach to parallelize the OTM method with MPI is to separate the spatial domain into distinct subdomains and allocate nodes and material points to each MPI process, such that each process treats its own subdomain independently. One advantage of this approach is the minimum

impact on the contents and structure of a serial code. Halo regions of nodes and material points are then distributed between the subdomains at every computation step such that the primary nodal variables and constitutive updates at the material points can be computed in parallel.

### 3.1 Domain Decomposition

To decompose the domain, the Recursive Coordinate Bisection (RCB) algorithm is used from Zoltan library [Boman et al., 2007]. The objective of the partitioning library is to provide a initial computational workload which is uniformly distributed. This is accomplished by a distribution of almost equal number of particles (nodes and material points) in each process. Domain decomposition is conducted by cutting along the partition planes in the spatial domain recursively (Fig 2). Each sub-domain is assigned to one process. Hence, the decomposition depends on the number of processes and the domain size [Selvam and Hoffmann, 2015]. The goal of using this domain decomposition algorithm is to ensures geometrical locality of the particles and to simplify the creation of halo regions. Both nodes and material points carry their influence and support domain information respectively during the distribution process.

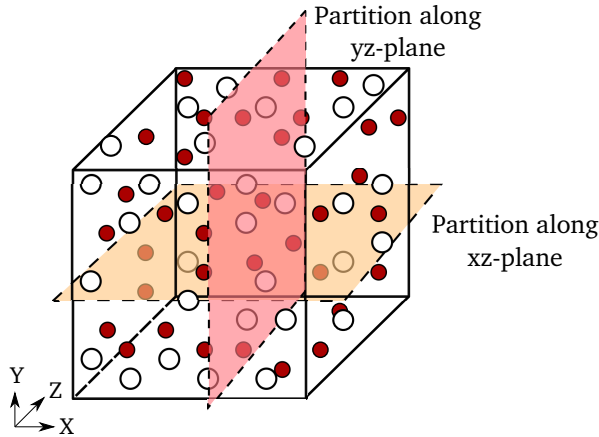


Figure 2: Domain decomposition into four processes using RCB.

Movement of particles and subsequent adjustments in the subdomain will cause load-imbalance among the processes. The computational load at a given time interval is monitored to re-assign the workload evenly among the processes and to minimize the communication. Major constraints are minimizing the computational efforts to compute the new division of the domain and minimizing the number of particles that need to be migrated among the processes. For the purpose of dynamic load balancing, the Recursive Coordinate Bisection algorithm is called. At optimized time intervals, within the mid-increment of the time step, to update the subdomain boundaries if required.

### 3.2 Dynamic Halo Regions

The nodal and material point updates are performed in each subdomain in parallel. For a node and material point, its influence and support domain could be spread across multiple processes (Fig 3). Nodes and material points, which are close to the division boundaries of subdomains need to share information. For this, the halo regions are necessary. These halo regions are copies of nodal and material point data that are sent to neighbor processes via two communication steps.

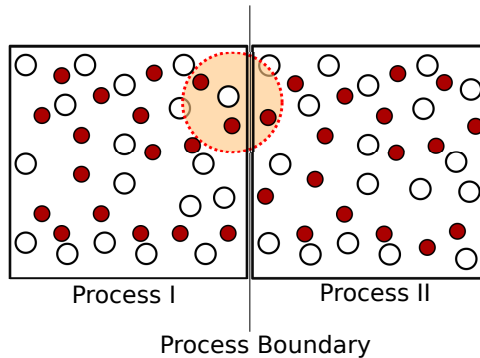


Figure 3: Influence domain of a node spread across multiple processes.

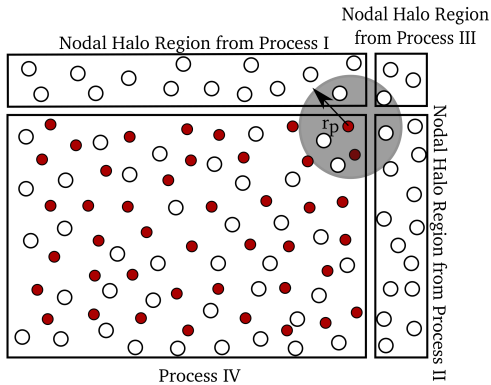


Figure 4: Nodal halo region and support domain with support nodes from halo region.

After nodal and material point updates, the halo regions are constructed dynamically during the computation, depending on the amount of communication. The first round involves nodal halo communication for the material point updates, where position, velocity, influence domain and other nodal data are communicated. Afterward, all data at material points can be computed within each subdomain. For material points whose support nodes are located in neighboring subdomains, its support domain is reconstructed through halo nodes (Fig 4). Hence, support domains are formed using nodes in its own subdomain and nodal halo region. After the update of material point information, the second round of communication involves material point halo communication for the nodal updates. Similarly, influence domain of nodes at the boundary of subdomains are reconstructed through material point halo regions (Fig 5). Nodal updates take place locally at each subdomain using the information from its own subdomain and from material point halo region.

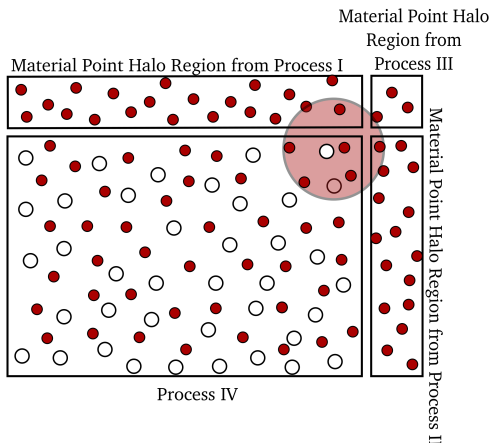


Figure 5: Material point halo region and influence domain using material points from halo region.

At the beginning of the halo communication process, the neighbors of each subdomain need to be detected. This will enable every subdomain to initiate communication locally only with its neighbors and to implement the nearest neighbor communication patterns (*sparse collective operations*). In general, for large scale applications, efficient implementation of sparse collective communication operations is most important [Hoefer and Träff, 2009]. At every time step, identification of nearest neighbors of a subdomain is facilitated through the process of bounding box intersection with its neighbor subdomains. A set of local neighborhoods (*process neighborhood*) is defined for every subdomain (Fig 7). Each process neighborhood consists a list of  $k$  target processes and a list of  $k$  source processes. For each subdomain, halo regions will be sent to target processes and simultaneously, it will receive halo regions from the same target processes. So, the source and target processes are same for each process but the amount of information to be received and sent may differ. Bounding box consists of coordinate information of both nodes and material points located at the lower and upper bounds of each subdomain and it is recomputed at every computation step. Even though the problem never occurred in our simulations, it is always checked that the minimal dimension of the bounding box is larger than the radius of the support domains so that only the nearest neighbors need to be detected, see Figure 6 for an illustration of the situation to be avoided. If such situation occurred, load balancing should be realized by calling Zoltan library in order to adapt the domain decomposition and transfer particles between subdomains. In Fig 8,  $(B_{max}^{II}, B_{min}^{II})$  represents the bounding box of Process II and  $(B_{max}^I, B_{min}^I)$  represents the bounding box of Process I. Before performing intersection, bounding boxes from neighboring processes are extended by a width equivalent

to maximum support radius of the sub-domain. Also, it ensures that there are no missing neighbor detection. This is performed at regular intervals to determine the extent of overlap of bounding boxes, i.e. halo regions. The maximum support radius at each sub-domain is used to extend the bounding boxes gathered from neighboring processes. Width of halo region for each sub-domain is identified as the overlap region between the bounding boxes of each sub-domain.

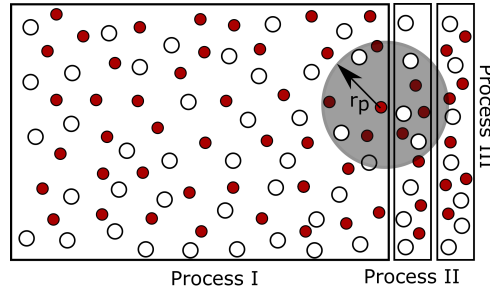


Figure 6: Support domain across multiple processes.

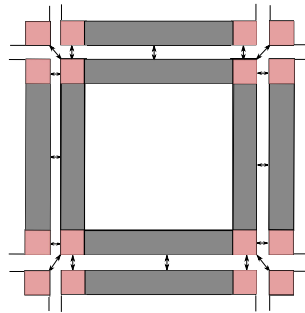


Figure 7: Schematic Representation of Nearest Neighborhood Communication for a typical halo region update operation, showing the halo exchanges for the faces and corners of sub-domains: Grey regions represent the halo communication among the faces of the sub-domains (each with one neighbor) and red regions represent the halo exchange among the corner parts of the sub-domains (each with three neighbors).

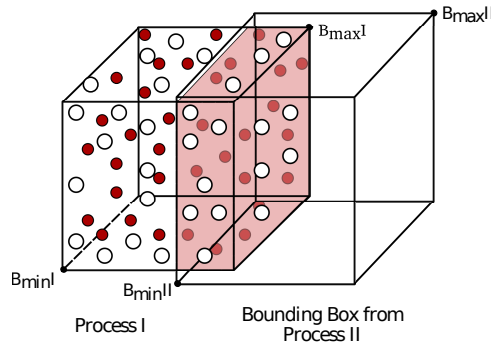


Figure 8: Bounding Box intersection between Process I and Process II and Identification of halo nodes in Process I in the intersection region is depicted in the Figure. Nodes in the intersection region will be sent as halo nodes to Process II.

Nodes within every intersecting bounding boxes are identified as halo nodes, see Fig 8. Halo nodes, which are to be sent, are serialized, i.e., the nodal data is represented as a large array of chars and stored in a buffer. Halo nodes sharing boundary with each neighbor is sent to the specific neighboring sub-domain (Fig 9). Here, MPI virtual topology functionality is used for sparse collective operations, it uses the set of local neighborhoods, i.e., source and target lists. Graph topology interface is used as it provides full flexibility in describing neighborhoods and the communication graphs are not limited to symmetric exchange patterns, which is in contrast to the Cartesian topology mechanism [Message Passing Interface Forum, 2012]. Before sending the halo nodes, the processes, at first, communicate how many nodes are to be exchanged along with total size of nodal data. After determining the total size of each buffer, memory allocation is made in the target process where the halo region is to be



received (receive buffer) from its neighbor ([Fig 9]. The overall nodal halo communication step is sketched in Algorithm 2.

---

**Algorithm 2** Nodal Halo Communication Step

---

**Require:** Bounding box computation at every process

**Require:** Detection of neighbor processes

1. Exchange bounding box with all neighbor processes using `MPI_Allgather`
2. Identify intersecting bounding boxes
3. Identify nodes at the intersection (or Halo nodes) to be sent

**Require:** Create local process neighborhood using `MPI_Dist_graph_create_adjacent`

1. Determine the nodal size for halo.
  2. Exchange the nodal sizes with nearest neighbors using `MPI_Neighbor_alltoallv`
  3. After receiving the nodal sizes at destination process, allocate memory for receive buffer
  4. Pack the halo nodes to be sent in a send buffer.
  5. Exchange the halo node information using `MPI_Neighbor_alltoallv`.
- 

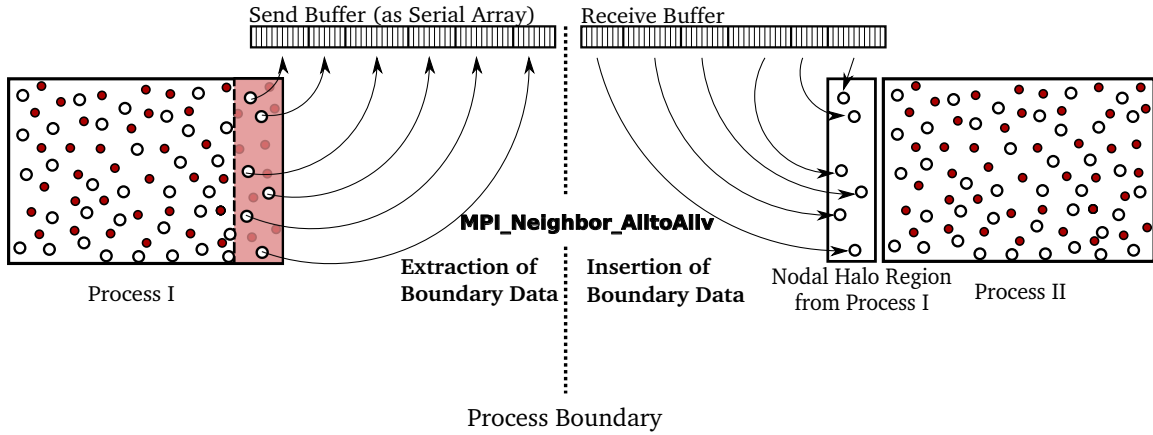


Figure 9: MPI Communication from Process I to Process II

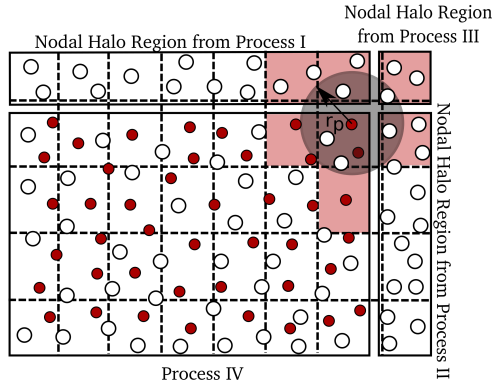


Figure 10: Linked Cell Method

After the nodal halo communication, material points perform the search process using nodal information from its own subdomain and from nodal halo region. In order to improve the computational efficiency of search algorithms, the linked cell method [Griebel et al., 2007] has been implemented. Linked cell method within OTM method is a new feature in this work. In solids undergoing large deformations or in fluid simulations, both the nodes and material points may change its position with time. An efficient search algorithm is needed to dynamically update the support domain while solving the equations without incurring excessive computational costs. Linked cell method significantly reduces the computational efforts, when the number of particles is large. The main idea of the linked list is to map the nodal positions on a grid. All the nodes have a unique particle ID and a data structure stores all the information of each grid. For each cell, a list of nodal IDs and pointer to those nodes are stored. After the nodal updates and subsequent formation of nodal halo regions, both the subdomain and nodal halo region is subdivided into static cells, (Fig 10). Only nodes in the vicinity of a material point are checked during the search process. This is done by identifying the cells which

intersect with the search radius and nodes in those intersecting cells are considered within the search process. A brief sketch of the steps involved is shown in Algorithm 3.

---

**Algorithm 3** Algorithmic scheme for Linked Cell Method

---

**Require:** Nodal Halo Region

**Require:** New support radius

1. Divide the subdomain and nodal halo region into cells, see Fig 10
  2. Identify intersection of cells with the support domain
  3. Find the nodes which belong to the material point
- 

After material point updates, communication of material point halo region follow in the same line as of nodal halo communication, starting with the bounding box intersection, communicating total number of material points to be exchanged, and communication of material point information.

---

**Algorithm 4** Parallel OTM Time Step

---

For Process  $\mathbb{P}^I, I = 1, \dots, P$  :

**Require:** Reading of Input information and Process  $\mathbb{P}^I$  storing its own set of nodes and material points.

- Initial nodal set and material point set
- Initial support domain of material points

**Require:** Domain Decomposition by Zoltan, see Fig 2.

**Require:** Initial material point halo regions(steps are similar to Algorithm 2)

For computation step  $t_k \rightarrow t_{k+1}$

1. Complete the influence domain with halo material points.
2. Compute the local mass matrix and local nodal force vector.
3. Update primary variables and nodal coordinates

**Require:** Nodal halo regions (for details, see Algorithm 2)

**Require:** Load balancing at optimized intervals, let's say at every time increments of  $t_{k+500}$

- Clear both nodal and material point halo regions
  - Call Zoltan functions for load balancing (steps are similar to domain decomposition as in Section 3.1)
4. Complete the support domain with halo nodes, see Fig 4.
  5. Update material point coordinates.
  6. Constitutive updates at material point.
  7. Division of subdomain and nodal halo region into cells (Linked Cell Method, see Fig 10 )
  8. Search algorithm to update the support domains
  9. Recompute shape functions

**Require:** Material Point halo regions (steps are similar to Algorithm 2)

---

### 3.3 Data Management Strategies

The core of parallel OTM method is formed by data structures and algorithms implemented as C++ templates (Not limited to C++ and could be also implemented in Fortran). To store all the data of the nodes and material points, C++ classes have been defined. For the management of nodal and material point data (removal or addition), STL maps to store pointers to objects of nodal and material point data are preferred. This gives us flexibility for quicker removal and addition of nodal and material point data during load-balancing and for the formation of support and influence domains.

Information that is contained in a particle (node or material point) are its identifier (Global particle ID), coordinates, flags (indicators, such as, the particle is a node or material point and if the node is on the physical boundary of the problem domain) and its affiliate (rank of the process that the node or material point belongs to). Additional information that is contained in a node and material point is its pointer-based influence and support domain information respectively. With the help of this data structure, every subdomain handles pointers to objects of nodes and material points, bounding box information (maximum and minimum coordinates), and neighbor information (halo regions for nodes and material points). Choosing a proper way to handle this STL container depends on the problem itself. For instance, there is continuous update of support and influence

domain in the OTM method, whose sizes can vary dynamically at every time step. After optimized intervals of dynamic load-balancing, the pointers to new particles (nodes and material points) are handled effectively by this container.

For nodal and material point halo communication, the data is packed into a serial array, whose size is varying. So, flexible data structures have been designed to pack all the information in the buffer. The message size for each node or material point is maintained as number of nodal or material point variables multiplied with the (size of double precision floating number), in order to prevent any kind of memory misalignment issues while packing information of mixed data types in a serial array.

The size of each nodal information is of arbitrary number of bytes due to the varying size of its influence domain. In [Li et al., 2014], MPI data structure was used to pack the nodal information. This restricts the information to be packed since only fixed-size information could be used to communicate. Here, the influence domain information of a node is packed more efficiently using a flexible size for every node. Similarly, for every halo material point, its support domain information is also included in the halo region. Packing support and influence information in halo region assists in localised updates within a subdomain. For instance, whenever the support domains of boundary material points are updated (Step 8 of Algorithm 4) and those material points are exchanged through halo communication, the updated support domain information of halo material points will assist in updating the influence domain of the nodes locally at each sub-domain. This flexibility feature for packing any amount of information for halo communication is necessary for localized updates of nodes and material points.

Another advantage of using STL map for support and influence domain is that pointers to the support nodes or influence material points can be released while preserving their IDs. This proved to be helpful in situations where the support domains need to be constructed again using halo nodes after nodal updates. For instance, at time step  $t_k$ , support domains are updated (Step 8 of Algorithm 4). Subsequently, for the material point updates (Steps 4-7 of Algorithm 4) at time step  $t_{k+1}$ , the support domain computed at previous time step  $t_k$  will be used.

Object-oriented implementation, robustness and flexibility of the parallel method to include additional physical phenomena are taken into consideration. With the use of *Eigen* templated library, all the vector and matrix information are stored in contiguous memory locations and matrix operations are optimized.

## 4 Parallel Performance

The objective of parallelism is to perform simulation of larger and complex problems. To evaluate the ability of the parallel implementation, strong scaling tests are conducted which measure the performance with increasing number of processes, keeping the problem size constant.

The computation being explicit in time with lumped mass, we expect the nodal and material point updates to scale perfectly with the number of subdomains. The use of linked cells method makes the search for support domains independent on the number of processes. The number of neighbors is at most 8 (4 edges, 4 vertices) in 2D and 26 in 3D (6 faces, 12 edges, 8 vertices), so the neighbor-communications do not depend on the total number of processes. The number of particles in the halos decreases sublinearly with the number of processors, and the ratio between inner and halo particles strongly depends on the dimensionality of the physical space. So that at some point, the cost of forming the halo should increase in comparison with the cost of the update of inner particles, for increasing number of subdomains. To sum up, we can expect our implementation to scale well if we take a low-rank parallel reference, up to a sufficiently large amount of subdomains when the formation of halo regions dominates the computational cost.

Every simulation is run for 2000 time steps. Variation of the computational efforts could occur between simulations due to fluctuations in cluster load and differences in configurations of the cluster nodes. Hence, each simulation is run for 3 times and the average CPU time is used in the studies. Output files are written in binary format of *vtk* for every process. Time taken to write the data files is also taken into consideration. The computational time is the maximum wallclock time for a single time step in Algorithm 4. Speedup is measured as

$$\text{Speedup} = \frac{t_n}{t_p} \quad (15)$$

For the baseline calculation, the sequential time  $n = 1$  is used and  $t_p$  is the maximum wallclock time for a single time step with  $p \geq n$ . Efficiency is measured as

$$\text{Efficiency} = \frac{n \times t_n}{p \times t_p} = \text{Speedup} \times \frac{n}{p} \quad (16)$$

In this section, we will assess the strong scalability characteristics of our parallel approach. The studies are performed on the LUIS Cluster of Leibniz Universität Hannover using only Haswell-based

nodes. Each Haswell-based node consists of two 8-core Intel Xeon E5-2630 processors. All nodes are interconnected with the Infiniband technology. The cluster nodes are based on Torque/ Maui (qsub, qstat) workload manager. Each sub-domain is assigned to one process (core).

## 4.1 Application to Taylor rod impact

The Taylor rod impact test is a widely accepted benchmark where a copper rod hits a rigid frictionless wall. The three-dimensional bar has a length of  $L = 32.4$  mm and a circular cross-section with radius  $R_0 = 3.2$  mm. The initial velocity is 227 m/s.

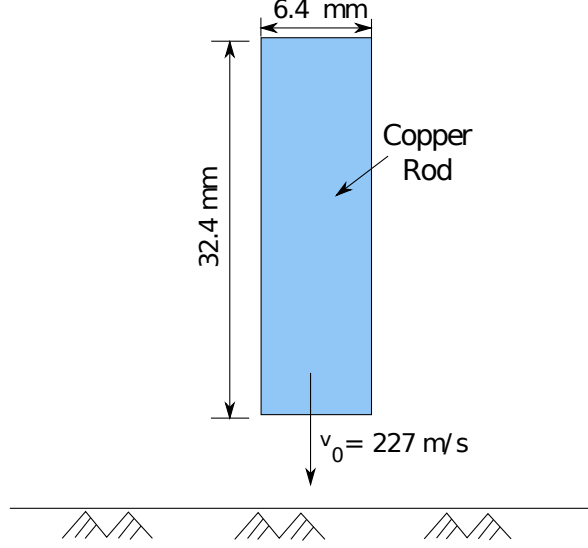


Figure 11: Geometrical setup of the Taylor rod test.

### 4.1.1 Material Model

In this benchmark problem, a finite plasticity material model with linear isotropic hardening is used to model the behavior of the rod. The formulation is based on the multiplicative split of the deformation gradient into an elastic and plastic part

$$\mathbf{F}_{pn} = \mathbf{F}_{pn}^e \mathbf{F}_{pn}^p \quad (17)$$

Assuming the strain behavior as Hencky strain measure, it can be expressed in terms of the left elastic Cauchy-Green strain tensor  $\mathbf{b}_{pn}^e$  as

$$\boldsymbol{\varepsilon}_{pn}^e = \ln \mathbf{V}_{pn}^e = \ln \left( \sqrt{\mathbf{b}_{pn}^e} \right) \quad (18)$$

where  $\mathbf{V}_{pn}^e$  is the elastic left stretch tensor.

Using the exponential map integrator, the Equation 18 can be expressed as

$$\boldsymbol{\varepsilon}_{pn}^e = \ln \left( \sqrt{\bar{\mathbf{b}}_{pn}^{etr}} \right) - \frac{\gamma_{pn} - \gamma_{pn-1}}{\Delta t} \frac{\partial f_{pn}}{\partial \boldsymbol{\tau}_{pn}}, \bar{\mathbf{b}}_{pn}^{etr} = \mathbf{Q}_{pn} \mathbf{b}_{pn}^{etr} \mathbf{Q}_{pn}^T \quad (19)$$

The elastic left Cauchy Green tensor is transformed in the principal stress space using the rotation tensor  $\mathbf{Q}_{pn}$ .

The onset of plastic yielding is defined by the yield function  $f$ . The yield surface divides the elastic domain from the plastic domain and the Kirchoff stresses  $\boldsymbol{\tau}_{pn}$  must lie within the elastic domain or on the yield surface. von Mises plasticity model is used and its deviatoric part leads to plastic deformations

$$\boldsymbol{\tau}_{pn} = p\mathbf{I} + \mathbf{s}_{pn} = K \operatorname{tr}(\boldsymbol{\varepsilon}_{pn}^e) \mathbf{1} + 2\mu \left( \boldsymbol{\varepsilon}_{pn}^e - \frac{1}{3} \boldsymbol{\varepsilon}_{pn}^e \cdot \mathbf{1} \otimes \mathbf{1} \right) \quad (20)$$

where, the constants  $K$  and  $\mu$  are the compression modulus and the second Lamé constant.

The plastic flow (or evolution of the plastic deformation gradient) can be defined in terms of the plastic strain as

$$\mathbf{d}_{pn}^p = \dot{\gamma}_{pn} \frac{\partial f_{pn}}{\partial \boldsymbol{\tau}_{pn}} \quad (21)$$

where  $f_{pn}$  is the yield function and it is expressed in terms of norm of the deviatoric stress  $\|s\|$  as  $f_{pn} = \|s_{pn}\| - \sqrt{\frac{2}{3}}\sigma_Y$ .

The accumulated plastic strain  $\bar{\varepsilon}_{pn}^p$  can be expressed in terms of the evolution equation of the hardening variable as

$$\dot{\bar{\varepsilon}}_{pn}^p = \sqrt{\frac{2}{3}} \|\dot{\varepsilon}_{pn}^p\| = \dot{\gamma}_{pn} \quad (22)$$

where  $\dot{\gamma}_{pn}$  is the rate of the plastic variable.

The evolution equation for the plastic strain in case of isotropic associated plasticity can be expressed in terms of Lie derivative of the elastic left Cauchy Green tensor

$$\mathcal{L}_v \mathbf{b}_{pn}^e = -2\mathbf{d}_{pn}^p \mathbf{b}_{pn}^e = -2\dot{\gamma}_{pn} \frac{\partial f_{pn}}{\partial \boldsymbol{\tau}_{pn}} \mathbf{b}_{pn}^e \quad (23)$$

where  $\mathbf{d}_{pn}^p$  is the plastic rate of deformation tensor and the plastic isotropy is modeled as  $\mathbf{W}^p = 0$  with  $\mathbf{W}^p$  as the skew symmetric part of the plastic velocity gradient.

To model large plastic deformations, the von-Mises yield criteria is applied alongwith linear isotropic hardening behavior (hardening modulus  $H$ )

$$f_{pn} = \|2\mu \boldsymbol{\varepsilon}_{pn}^{etr}\| - 2\mu \Delta \gamma_{pn} - \sqrt{\frac{2}{3}} \left[ \sigma_{Y_0} + H \left( \bar{\varepsilon}_{pn-1} + \sqrt{\frac{2}{3}} \Delta \gamma_{pn} \right) \right] \leq 0 \quad (24)$$

where  $\sigma_{Y_0}$  corresponds to yield stress and  $\bar{\varepsilon}_{pn-1}$  corresponds to isotropic hardening variable computed at previous computation step. For  $f < 0$ , the Kirchoff stresses lie in the elastic domain. But, when  $f > 0$  for  $\Delta \gamma_{pn} = 0$ , the yield criteria is violated and the plastic increment has to fulfill the constraint  $f = 0$  for Kirchoff stresses to lie on the yield surface. This can be corrected using Equation (24). Through back transformation using the rotational tensors as in Equation (19), the Cauchy stress tensor in Equation (3) can be written as

$$\boldsymbol{\sigma}_{pn} = \mathbf{Q}_{pn} J [K \text{tr} \boldsymbol{\varepsilon}_{pn}^e \mathbf{1} + 2\mu (\boldsymbol{\varepsilon}_{pn}^e - \text{tr} \boldsymbol{\varepsilon}_{pn}^e \mathbf{1})] \mathbf{Q}_{pn}^T \quad (25)$$

#### 4.1.2 Contact Formulation

Additionally, a contact algorithm is needed to model the copper rod striking a rigid wall. A simple contact algorithm is used assuming that the wall is rigid and the tangential movement is frictionless. The normal gap  $g_{I_{n+1}}$  of each node at the next time step can be computed as

$$g_{I_{n+1}} = (\mathbf{x}_{I_{n+1}} - \bar{\mathbf{x}}) \cdot \mathbf{n} \quad (26)$$

where  $\bar{\mathbf{x}}$  is the coordinate of the rigid plane and  $\mathbf{n}$  is the normal vector on that rigid plane. To enforce the non-penetration condition, a Dirichlet boundary condition is applied on the corresponding node with prescribed displacements at the next time step

$$\mathbf{u}_{I_{n+1}} = \mathbf{x}_{I_n} - g_{I_{n+1}} \mathbf{n} \quad (27)$$

The above condition is only applied when the non-penetration condition is violated  $g_{I_{n+1}} < 0$ . More details about formulations on two contacting deformable bodies can be found in [Wriggers, 2006].

#### 4.1.3 Numerical Evaluation

The material parameters are chosen as  $\nu = 0.35$  for the Poisson ratio,  $E = 117.10^9 N/m^2$  for the Young's modulus,  $\rho_0 = 8.93 \cdot 10^3 kg/m^3$  for the density,  $H = 100.10^6 N/m^2$  for the hardening modulus and  $Y_0 = 400.10^6 N/m^2$  for the initial yield stress. For a stable explicit time integration scheme, a computation step size of  $\Delta t = 4.10^{-9}$  s is selected.

The initial domain is set up by triangulation with the material points located at the barycenters of the tetrahedral elements. Subsequently, the initial mesh is jettisoned and the computations proceed in a meshfree manner. The model contains 5,966 nodes and 28,423 material points. The domain decomposition is performed by distributing nodes and material points across all processes with the help of Zoltan library, see Section 3.1. Fig 12 shows a sequence of snapshots of the Taylor rod impacting axially against a rigid boundary. Here, MPI Process Rank refers to the rank in order to identify a process, which is an integer in the range  $[0, N - 1]$  where  $N$  is total number of MPI processes. Table 1 presents the average size of the subdomains in terms of own particles and halo region.

The strong scaling studies are performed for up to 239 processes. For these studies, the code is run on the same nodes but the allocation of the cores may vary. In practice, the standard deviation for the computational time for the 3 runs is less than 5% of the average. The Parallel Performance Analysis (Fig. 13 and Table 2) shows that the efficiency is about 55% up to 150 process then it slowly decreases.

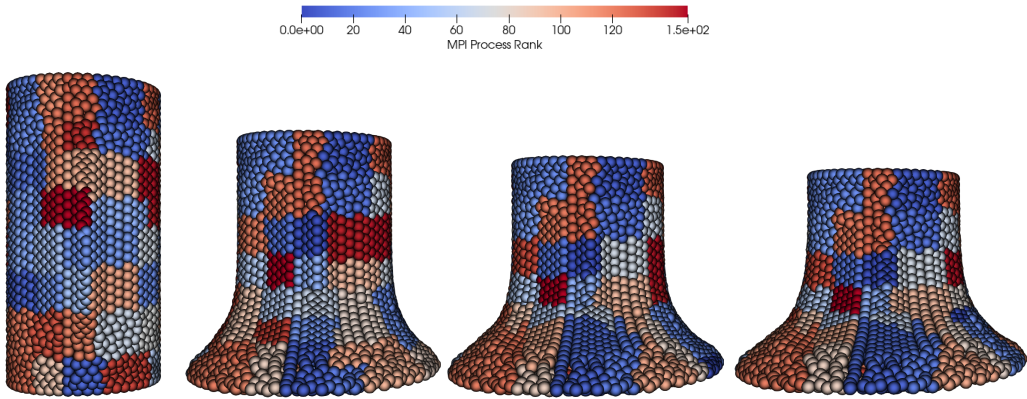


Figure 12: Snapshots of Taylor Rod deformation alongwith nodal distribution in different sub-domains

# MPI processes	# nodes	# MP	Av.# halo nodes	Av.# halo MP
1	5966	28,423	0	0
5	1190	5683	434	2233
50	120	569	209	1100
100	60	285	145	772
150	40	190	116	620
199	30	143	100	544
239	25	119	92	496

Table 1: Taylor rod impact: subdomain and halo average sizes (MP = material point) depending on the decomposition (halo is for the initial time step).

Fig 14 tries to analyze the situation more precisely. In fact it appears that the particles computations are relatively fast, making the overtime due to data distribution significant. Then, in a first regime (number of processes larger than 2 and less than 150), adding subdomains decreases the amount of particles to treat per subdomain as well as the size of the halo particles to be exchanged, see Table 1, making the method scale well when the reference is a parallel computation with few processes. For large number of processes, the exchange time and other incompressible stages (e.g. identification of halo particles) are dominating and the performance tends to deteriorate. A perspective of this work is to make a better implementation for the prediction of halo particles which are to be sent to the neighbors.

Number of MPI processes	Wallclock time (s)	Speedup	Efficiency (%)
1	2697.26	1	100
5	816.67	3.13	62.6
10	517.82	5.208	52.08
50	86.19	31.29	62.58
100	46.69	57.76	57.76
150	32.88	82.03	54.68
199	26.23	102.83	51.97
239	23.78	113.42	47.45

Table 2: Performance of the parallel implementation of the OTM method for the simulation of Taylor rod impact test.

## 4.2 Application to Serrated Chip Formation Process

In the second test case, numerical modeling of chip formation is discussed. Beside physical mechanisms such as plastic deformations additionally adiabatic shear band formation and ductile fracture are involved. Only basic equations are introduced and a more detailed explanation can be found in

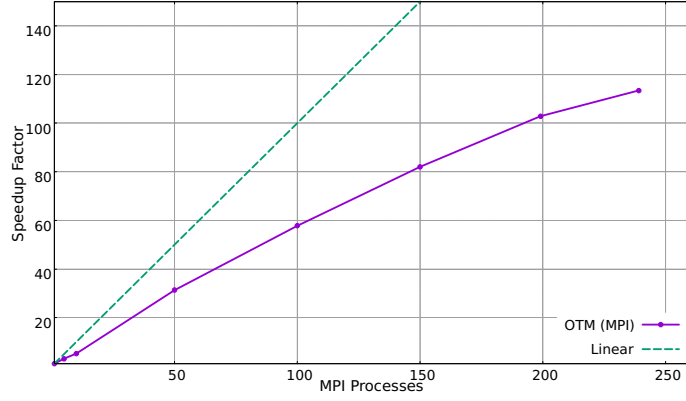


Figure 13: Parallel performance analysis: Strong scaling for Taylor rod impact.

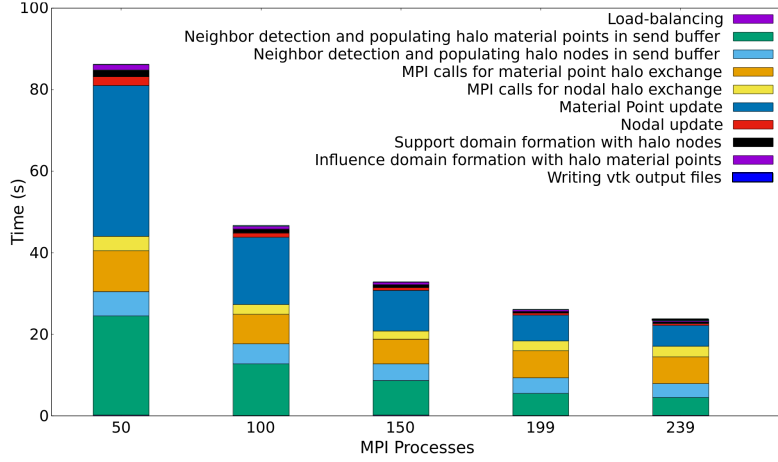


Figure 14: Taylor rod impact: Comparisons of growth in computational and communication overhead time in strong scaling tests.

[Huang et al., 2019].

#### 4.2.1 Material Model

The plastic deformation and the ductile fracture of the workpiece are described by the Johnson-Cook flow stress model and Johnson-Cook fracture model respectively. The evolution equations have the forms as shown in Equation (23). Euler backward time integration scheme is used to solve the evolution equations based on the elastic predictor corrector return mapping algorithm, for more details, see [de Souza Neto et al., 2008].

Using von Mises plasticity, the yield function is expressed as

$$f^{p, flow}(\boldsymbol{\tau}_{pn}) = \sqrt{\frac{3}{2}} \|\text{dev}(\boldsymbol{\tau}_{pn})\| - \sigma_Y(\varepsilon_{eq}^{pn}, \dot{\varepsilon}_{eq}^{pn}, T) \quad (28)$$

where,  $\boldsymbol{\tau}_{pn}$  is the Kirchoff stress,  $\sigma_Y$  the flow stress which is assumed to be a function of equivalent plastic strain rate  $\dot{\varepsilon}_{eq}^{pn}$ , equivalent plastic strain  $\varepsilon_{eq}^{pn}$  and the temperature  $T$ .

Temperature increase occurs due to adiabatic heating from plastic deformation and the temperature evolution can be formulated as

$$\dot{T} = \beta \frac{\sigma_v \dot{\gamma}}{\rho C_p}, \quad \sigma_v = \sqrt{\frac{3}{2}} \|\text{dev}(\boldsymbol{\sigma})\| \quad (29)$$

where,  $\sigma_v$  is vonMises equivalent stress,  $C_p$  is the heat capacity and  $\beta$  is the Taylor-Quinney coefficient.

Multiplicative decomposed power form of the flow stress has been applied to consider the effects of strain hardening, strain rate hardening and thermal softening. The Johnson-Cook hardening law [Johnson and Cook, 1983] is used to capture these effects

$$\sigma_Y = [A + B(\varepsilon_{eq}^{pn})^n] \left[ 1 + C \ln \left( \frac{\dot{\varepsilon}_{eq}^{pn}}{\dot{\varepsilon}_{e0}^{pn}} \right) \right] \left[ 1 - \left( \frac{T - T_r}{T_m - T_r} \right)^m \right] \quad (30)$$

where  $A$  defines the initial yield stress,  $\dot{\varepsilon}_{e0}^{pn}$  is the reference plain strain rate,  $T_m$  is the melting

temperature,  $T_r$  is the room temperature and,  $B$ ,  $C$ ,  $m$  and  $n$  are additional material parameters.

Johnson-Cook fracture model describes the separation of the chip from the workpiece and the serrated morphology on the chip upper surface. At the vicinity of the tool tip, high compression in the material and high concentration of strain occurs. Ductile fracture leads to separation of the material from the workpiece at the vicinity of the tooltip. Also, ductile fracture at the chip upper surface can lead to the formation of serrated chips. Johnson-Cook fracture model is used to model the ductile fracture and to predicts the fracture locations. When the accumulated equivalent plastic strain,  $\varepsilon_{eq}^{pn}$  reaches the critical value,  $\varepsilon_{eqf}^{pn}$ , ductile fracture occurs

$$\varepsilon_{eq}^{pn} \geq \varepsilon_{eqf}^{pn} = [d_1 + d_2 \text{Exp}(d_3 \eta)] \left[ 1 + d_4 \ln \left( \frac{\varepsilon_{eq}^{pn}}{\varepsilon_0^{pn}} \right) \right] \left[ 1 + d_5 \frac{T - T_r}{T_m - T_r} \right] \quad (31)$$

where  $d_1$ ,  $d_2$ ,  $d_3$ ,  $d_4$  and  $d_5$  are the material parameters,  $\eta$  is the stress triaxiality which is defined as

$$\eta = \frac{p}{\sigma_v}, \quad p = \lambda \text{tr}(\varepsilon^e) \quad (32)$$

where  $p$  is the hydrostatic pressure and  $\sigma_v$  is the von Mises stress.

The deformations in the chip and the workpiece during metal cutting are driven by the cutting tool directly which moves in horizontal direction with a specific cutting depth and cutting speed. The non-penetration condition is defined by a projection of the slave node positions from the workpiece onto the cutting tool surface

$$g^N = (\mathbf{x}^s - \mathbf{x}^m) \cdot \mathbf{n}^m \geq 0 \quad (33)$$

The abbreviations  $g^N$  the normal gap,  $\mathbf{x}^s$  the slave node from the workpiece  $\mathbf{x}^m$  and  $\mathbf{n}^m$  are the orthogonal projection of  $\mathbf{x}^s$  on the tool surface and  $\mathbf{n}^m$  is the normal vector associated to the tool body.

The normal contact force and the stick tangential contact force can be determined by using the penalty method as

$$\mathbf{t}^N = c_N \mathbf{g}_N, \quad \mathbf{t}^T = c_T \mathbf{g}_T \quad (34)$$

where  $c_N$  and  $c_T$  are the penalty parameters.

The tangential contact force in the slip state is determined from the Coulomb friction law as

$$\mathbf{t}^T = -\mu \|\mathbf{t}_N\| \frac{\dot{\mathbf{g}}_T}{\|\dot{\mathbf{g}}_T\|} \quad (35)$$

where  $\mu$  is the frictional coefficient. Further details can be found in [Huang et al., 2019].

## 4.2.2 Numerical Evaluation

Ti6Al4V alloy is used as the workpiece material. The material parameters of the constitutive equations (30) and (31) can be found in [Huang et al., 2019]. The workpiece has a length and height of 300  $\mu\text{m}$  and 120  $\mu\text{m}$  respectively, see Fig 15. The cutting depth is 100  $\mu\text{m}$ . The cutting tool is treated as rigid body with tool radius of 2  $\mu\text{m}$  and rake angle of 0°. For the tool-chip contact modeling, the friction coefficient is set as 0.8. For the workpiece, the melting temperature  $T_m$  and the initial temperature  $T_r$  is set as 1630°C and 25°C respectively. For a stable explicit time integration scheme, a time step size of  $\Delta t = 10^{-10}$  s is selected.

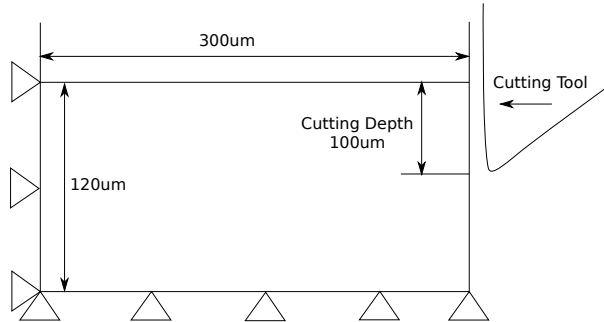


Figure 15: Geometrical model for metal cutting.

The model consists of 27,417 nodes and 107,975 material points. Their distribution in subdomains is presented in Table 3. In this example, the scalability performance of the multiprocessing approach in the numerical solutions of large deformation problem is investigated. In Fig 16, the sequence of the serrated chip formation process is shown together with the corresponding nodal distribution across the sub-domains. Strong scaling studies are conducted for up to 549 processes, see Figure 17 and Table 4, as well as Figure 18 for a more detailed analysis.



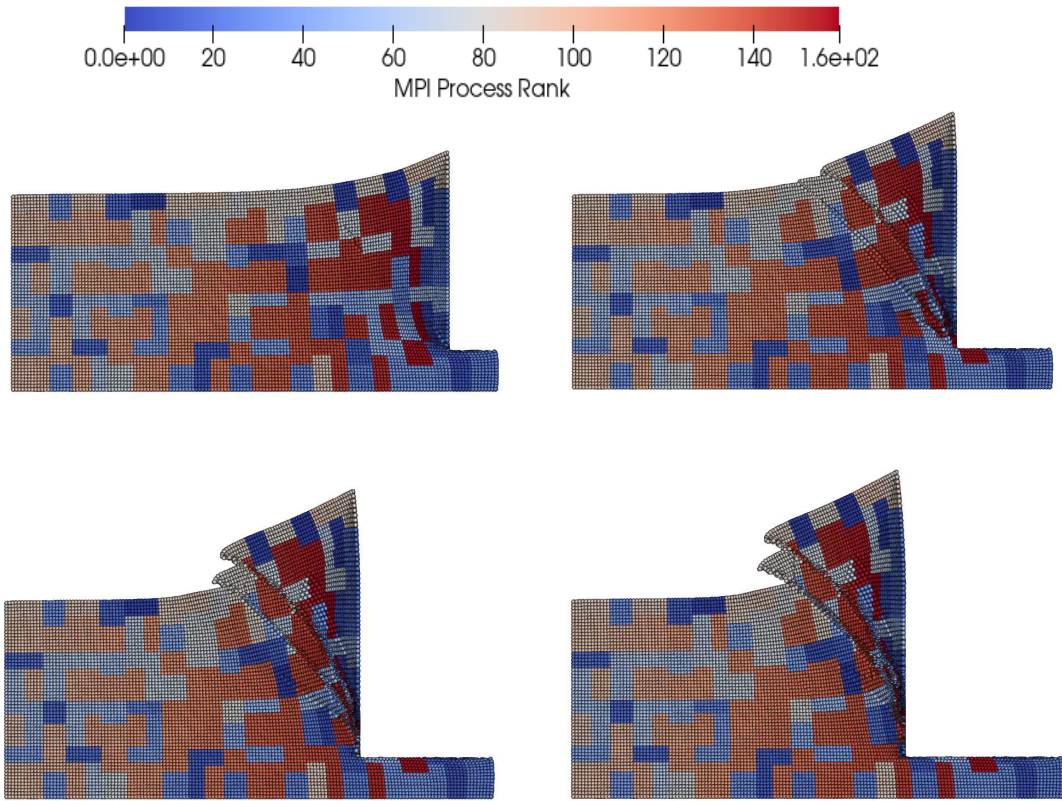


Figure 16: Snapshots of serrated chip formation process alongwith nodal distribution in different subdomains.

We observe that the parallel efficiency is better than from previous test case. The speedup is excellent up to 400 processes. This difference can be explained by the fact that the problem is mostly 2D and each subdomain possesses at most 8 neighbors, which limits the communications. Note that the 199-process case exceeds linear behavior, but this remains in range of measurement variability.

# MPI processes	# nodes	# MP	Av.# halo nodes	Av.# halo MP
1	27,417	107,975	0	0
4	6852	26,993	288	1047
8	3428	13,497	214	864
50	549	2160	98	417
100	275	1080	76	293
150	183	720	63	250
199	138	543	57	225
239	115	452	51	204
299	92	362	47	185
348	79	311	43	170
450	61	240	39	151
549	50	197	36	137

Table 3: Serrated chip formation process :subdomain and halo average sizes (MP = material point) depending on the decomposition (halo is for the initial time step).

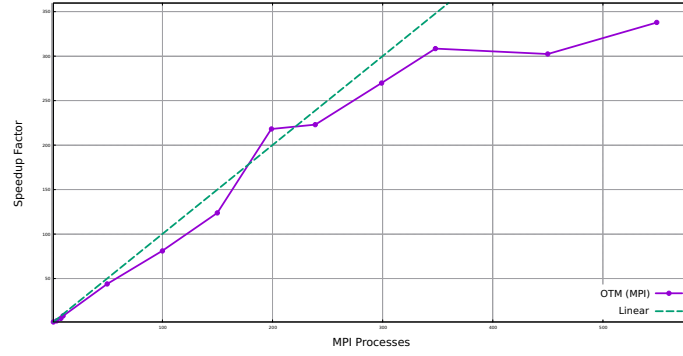


Figure 17: Parallel performance analysis: Strong scaling for serrated chip formation process.

Number of MPI processes	Wallclock time (s)	Speedup	Efficiency (%)
1	11348.8932	1.0	100
4	4514.898	2.51	62.84
8	2224.957	5.1	63.75
10	1432.27	7.923	79.23
50	258.27	43.94	87.88
100	139.88	81.13	81.13
150	91.66	123.81	82.54
199	52.017	218.17	109.63
239	50.877	223.06	93.33
299	42.063	269.80	90.23
348	36.780	308.56	88.66
450	37.520	302.47	67.21
549	33.589	337.87	61.54

Table 4: Performance of the parallel implementation of the OTM method for the simulation of the serrated chip formation process.

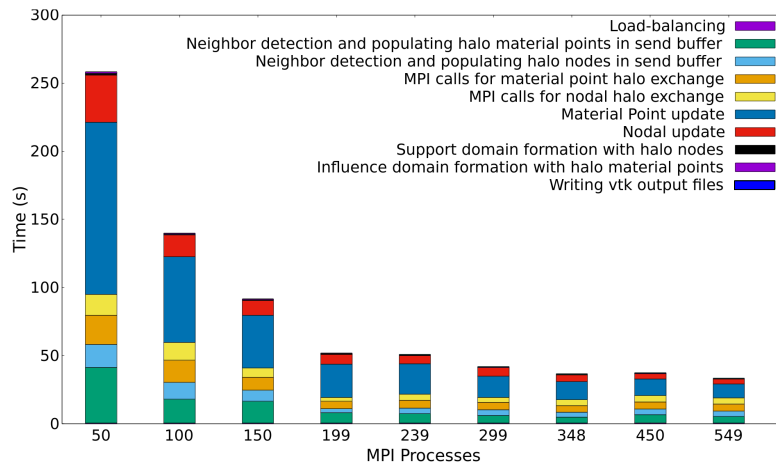


Figure 18: Serrated chip formation process: Comparisons of growth in computational and communication overhead time in strong scaling tests.

## 5 Conclusions

An OTM algorithm for large deformations, parallelized using MPI with an objective for scalability on large scale CPU clusters has been presented. The consistency and robustness of this algorithm is demonstrated by two examples showing large deformation. Strong scaling studies were conducted. Implementation of dynamic halo regions have shown to improve the scalability by its ability to handle variable workloads and eliminating the storage issues related to fixed-size arrays. With the increase in number of processes, good scalability is observed for the 2D Serrated Chip Formation Process example. The communication costs decreases significantly and asymptotically even though more subdomain interfaces are present leading to increase in number of halo particles. The second advantage is the efficient data management strategy using advanced STL container adapted to fulfill various functionalities of data structure modifications. Flexible handling of data structures for two types of particles (nodes and material points) resulted in reduction of computational costs. Together with localized computation within each sub-domain by using nearest neighborhood collectives for both nodes and material points this approach leads to scalable results. Since the method is explicit in time, computations are very simple and with low granularity. We may study new ordering of operations (starting from the innermost particles) so that computations could be started while the halo are being exchanged, in the spirit of [Cornelis et al., 2018]. We could also consider larger halo with several time steps being computed without synchronizations, or even fully asynchronous computations, see for example [Magoulès and Gbikpi-Benissan, 2018]. Anyhow, the first improvement to be implemented is a mixed MPI/OpenMP approach to parallelism.

## 6 Acknowledgements

Funding supports from Deutsche Forschungsgemeinschaft DFG within the research training centre ViVaCE (IRTG 1627), French-German doctoral college 'Sophisticated Numerical and Testing Approaches' (SNTA) and Graduierten Akademie-Leibniz Universität Hannover is gratefully acknowledged.

## References

- M. Arroyo and M. Ortiz. Local maximum-entropy approximation schemes: a seamless bridge between finite elements and meshfree methods. *International Journal for Numerical Methods in Engineering*, 65:2167–2202, 2006.
- P. Balaji, D. Buntinas, D. Goodell, W. Gropp, and R. Thakur. Fine-Grained Multithreading Support for Hybrid Threaded MPI Programming. *International Journal of High Performance Computing Applications (IJHPCA)*, 24(1):49–57, 2010.
- Y. Barigou and E. Gabriel. Maximizing Communication-Computation Overlap Through Automatic Parallelization and Run-time Tuning of Non-blocking Collective Operations. *International Journal of Parallel Programming*, 45:1390–1416, 2017.
- Y. Barigou, V. Venkatesan, and E. Gabriel. Auto-tuning non-blocking collective communication operations. *IEEE International Parallel and Distributed Processing Symposium Workshop*, pages 1204–1213, 2015.
- K. J. Bathe. *Finite Element Procedures*. Prentice Hall, 2006.
- T. Belytschko, Y. Krongauz, J. Dolbow, and C. Gerlach. On the completeness of meshfree particle methods. *International Journal for Numerical Methods in Engineering*, 43(5):785–819, 1998.
- E. Boman, K. Devine, L. A. Fisk, R. Heaphy, B. Hendrickson, C. Vaughan, U. Catalyurek, D. Bozdag, W. Mitchell, and J. Teresco. *Zoltan 3.0: Parallel Partitioning, Load-balancing, and Data Management Services; Developer's Guide*. Sandia National Laboratories, Albuquerque, NM, 2007. Tech. Report SAND2007-4749W.
- Z. Cao, A.K. Patra, and M. Jones. Data Management and Volcano Plume Simulation with Parallel SPH Method and Dynamic Halo Domains. *Procedia Computer Science*, 108:786–795, 2017.
- J. Cornelis, S. Cools, and W. Vanroose. The Communication-Hiding Conjugate Gradient Method with Deep Pipelines. *ArXiv*, abs/1801.04728, 2018.
- E. Cueto and F. Chinesta. Meshless methods for the simulation of material forming. *International Journal for Material Forming*, 8(1):25–43, 2013.

- D. Culler, R. Karp, D. Patterson, A. Sahay, K. E. Schauer, E. Santos, R. Subramonian, and T. von Eicken. LogP: Towards a realistic model of parallel computation. In *Proceedings of the Fourth ACM SIGPLAN Symposium on Principles and Practice of Parallel Programming*, pages 1–12, 1993.
- E. A. de Souza Neto, D. Perić, and D. R. J. Owen. *Computational Methods for Plasticity : Theory and Applications*. John Wiley & Sons, Chichester, 2008.
- C. Farhat and F.-X. Roux. A method of finite element tearing and interconnecting and its parallel solution algorithm. *International Journal for Numerical Methods in Engineering*, 32(6):1205–1227, 1991.
- A. Ferrari, M. Dumbser, E. F. Toro, and A. Armanini. A new 3D parallel SPH scheme for free surface flows. *Computers & Fluids*, 38(6):1203–1217, 2009.
- Y. E. Gharbi, A. Parret-Fréaud, C. Bovet, and P. Gosselet. Two-level substructuring and parallel mesh generation for domain decomposition methods. *Finite Elements in Analysis and Design*, 192:103484, 2021. doi: 10.1016/j.finel.2020.103484. URL <https://hal.archives-ouvertes.fr/hal-02881249>.
- S. M. Ghazimirsaeed, Q. Zhou, A. Ruhela, and M. Bayatpour. A hierarchical and load-aware design for large message neighborhood collectives. In *Proceedings of the International Conference for High Performance Computing, Networking, Storage and Analysis*, pages 1–13, 2020.
- M. Griebel, S. Knapek, and G. Zumbusch. *Numerical Simulation in Molecular Dynamics: Numerics, Algorithms, Parallelization, Applications*. Springer Science & Business Media, 2007.
- T. Hoefer and J. L. Träff. Sparse collective operations for MPI. In *23<sup>rd</sup> IEEE International Symposium on Parallel and Distributed Processing, IPDPS 2009*, Rome, 2009. IEEE.
- T. Hoefer, R. Rabenseifner, H. Ritzdorf, B. R. de Supinski, R. Thakur, and J. L. Träff. The scalable process topology interface of MPI 2.2. *Concurrency and Computation: Practice & Experience*, 23(4):293–310, 2011.
- D. Huang, C. Weißenfels, and P. Wriggers. Modelling of serrated chip formation processes using the stabilized optimal transportation meshfree method. *International Journal of Mechanical Sciences*, 155:323–333, 2019.
- T.R.J. Hughes. *The Finite Element Method*. Prentice Hall, Englewood Cliffs, New Jersey, 1987.
- G. R. Johnson and W. H. Cook. A constitutive model and data for metals subjected to large strains, high strain rates and high temperatures. In *Seventh International Symposium on Ballistics*, The Hague, The Netherlands, 1983.
- C. Laoide-Kemp. *Investigating MPI streams as an alternative to halo exchange*. PhD thesis, The University of Edinburgh, Edinburgh, U.K., 2015.
- A. Laszloffy, J. Long, and A.K. Patra. Simple data management, scheduling and solution strategies for managing the irregularities in parallel adaptive hp finite element simulations. *Parallel Computing*, 26(13-14):1765–1788, 2000.
- B. Li, F. Habbal, and M. Ortiz. Optimal transportation meshfree approximation schemes for fluid and plastic flows. *International Journal for Numerical Methods in Engineering*, 83(12):1541–1579, 2010.
- B. Li, M. Stalzer, and M. Ortiz. A massively parallel implementation of the Optimal Transportation Meshfree method for explicit solid dynamics. *International Journal for Numerical Methods in Engineering*, 100(1):40–61, 2014.
- Frédéric Magoulès and Guillaume Gbikpi-Benissan. Asynchronous parareal time discretization for partial differential equations. *SIAM Journal on Scientific Computing*, 40(6):C704–C725, 2018. doi: 10.1137/17M1149225.
- J. Mandel. Balancing domain decomposition. *Communications in Numerical Methods in Engineering*, 9(3):233–241, 1993.
- Message Passing Interface Forum. *MPI: A Message-Passing Interface Standard, Version 3.0*. Message Passing Interface Forum, 2012.
- Message Passing Interface Forum. *MPI: A Message-Passing Interface Standard, Version 3.1*. Message Passing Interface Forum, 2015.
- Y. Notay and A. Napov. A massively parallel solver for discrete poisson-like problems. *Journal of Computational Physics*, 281:237–250, 2015.

- G. Oger, D. Le Touzé, D. Guibert, M. de Lefle, J. Biddiscombe, J. Soumagne, and J.-G. Piccinalli. On distributed memory MPI-based parallelization of SPH codes in massive HPC context. *Computer Physics Communications*, 200:1–14, 2016.
- A. K. Patra, A. C. Bauer, C. C. Nichita, E. B. Pitman, M. F. Sheridan, M. Bursik, B. Rupp, A. Webber, A. J. Stinton, L. M. Namikawa, and C. S. Renschler. Parallel adaptive numerical simulation of dry avalanches over natural terrain. *Journal of Volcanology and Geothermal Research*, 139(1-2):1–21, 2005.
- S. J. Plimpton and K. D. Devine. MapReduce in MPI for Large-scale graph algorithms. *Parallel Computing*, 37(9):610–632, 2011.
- O. T. Prims, M. Castrillo, M. C. Acosta, O. Mula-Valls, A. S. Lorente, K. Serradell, A. Cortés, and F. J. Doblas-Reyes. Finding, analysing and solving MPI communication bottlenecks in Earth System models. *Journal of Computational Science*, 36, 2019.
- M. Selvam and K. A. Hoffmann. MPI/Open-MP hybridization of higher order WENO scheme for the incompressible Navier-Stokes equations. In *AIAA SciTech*, 2015.
- D. Sulsky, Z. Chen, and H. L. Schreyer. A particle method for history- dependent materials. *Computer Methods in Applied Mechanics and Engineering*, 118(1-2):179–196, 1994.
- C. Villani. *Topics in Optimal Transportation Theory*, volume 58. American Mathematical Society, Providence, Rhode Island, 2013.
- V. Visseq, P. Alart, and D. Dureisseix. High performance computing of discrete nonsmooth contact dynamics with domain decomposition. *International Journal for Numerical Methods in Engineering*, 96(9):584–598, 2013.
- C. Weïßenfels and P. Wriggers. Stabilization algorithm for the optimal transportation meshfree approximation scheme. *Computer Methods in Applied Mechanics and Engineering*, 329:421–443, 2018.
- H. Wessels, C. Weïßenfels, and P. Wriggers. Metal particle fusion analysis for additive manufacturing using the stabilized optimal transportation meshfree method. *Computer Methods in Applied Mechanics and Engineering*, 339:91–114, 2018.
- H. Wessels, T. Bode, C. Weïßenfels, P. Wriggers, and T. I. Zohdi. Investigation of heat source modeling for selective laser melting. *Computational Mechanics*, 63:949–970, 2019.
- P. Wriggers. *Computational Contact Mechanics*. Springer-Verlag, Berlin, Heidelberg, 2<sup>nd</sup> edition, 2006.
- E. Yang, H. H. Bui, H. D. Sterck, G. D. Nguyen, and A. Bouzza. A scalable parallel computing SPH framework for predictions of geophysical granular flows. *Computers and Geotechnics*, 121, 2020.