DEM-based modelling framework for spray-dried powders in ceramic tiles industry. Part II: Solver implementation.

J.M. Tiscar, A. Escrig, G. Mallol, J. Boix, F.A. Gilabert

Abstract

In the preceding Part I, a combined experimental-numerical study to characterize fine spray-dried powder used in the ceramic tile pressing process was presented. In the present Part II, the algorithm proposed by (Mazhar, 2011) to solve the numerical simulation of powder dynamics through the Discrete Element Method (DEM) was extended. In this paper, the original algorithm was adapted for efficient use on multi-core CPUs and a single GPU. In both cases, history-dependent contact models were considered. The efficiency of the algorithms was compared among them and with LIGGGHTS, a reference software package for particle simulation using DEM. The results demonstrated a higher performance of the codes developed compared to LIGGGHTS, particularly in demanding scenarios with a large number of particles (more than 1 million) of small size (median diameter in volume less than 1 mm). In particular, the CPU-based algorithm was suitable for simulating the mould filling in ceramic tiles industry.

Keywords: DEM, Parallel computing, GPU computing, multi-threading
1. Introduction

Numerical simulation has become a useful tool to optimize and design industrial equipment. Concerning powder processing, the Discrete Element Method (DEM) [1] is widely used to simulate the granular behaviour, however it still involves excessive computational efforts [2–5]. In DEM, the Newton equations for translation and rotation are solved for each particle. This approach allows the use of contact models which reproduce the physical behaviour of the particles, helps to investigate granular flow characteristics and analyses the particle tracking. Different models of contact are present in literature, such as linear [6] and Hertz [7, 8] model. DEM is applied to an extensive range of physical and engineering problems, ranging from the study of agglomerates [9] (micro-scale) to rockfall dynamics [10] (macro-scale). In the context of the simulation of industrial processes, a huge number of particles must be modelled to obtain representative results, and despite the development of techniques for reducing the number of particles [11], simulating an industrial scenario used to involve an important computational effort.

One of the most used software packages in DEM simulations is LIGGGHTS [12], which is an open source software and an improved version of LAMMPS. LAMMPS is another open source software used in the field of molecular dynamics (MD) [13], based on a contact detection algorithm that relies on a uniform grid data structure handled by means of the well-known and widely-used linked-list data structure [14]. Regarding parallel computing, LIGGGHTS uses message-passing interface (MPI) to scale across multiple processing cores and take advantage of linked lists. Although it is a good method to obtain parallel code, some researches have proposed other algorithms in the multi-threaded computing approach [15, 16].

In addition to the CPU computing, nowadays, GPU computing is also used for
Numerical simulations in different scientific fields [17–19]. Several reports have shown a good performance of GPUs versus single CPU computing [20, 21], whilst other researchers have noticed a similar performance for both multi-thread CPU and GPU computing [22].

The main differences between both computing platforms are briefly described below (refer to [23] for further description):

- **Number of cores.** Whilst the current CPUs consist of 4 or more cores, the GPUs can have hundreds of independent cores. However, the clock frequency of GPU cores is significantly lower than the CPU ones.

- **Cache management.** The single instruction multiple data (SIMD) model of a GPU sets several cores to run the same code in parallel, so GPUs do not provide a general cache for all memory accesses like CPUs. This design hides the latency of memory accesses, but increases the overhead in the inter-core communication through main memory.

- **Algorithm design.** The development of scientific applications on GPUs used to be more challenging and more complex than on CPUs [17]. The particular architecture of the GPUs implies using all cores and avoiding inter-core communications (limiting concurrency capabilities) to get a high performance.

There are also algorithms based on a CPU-GPU hybrid computation that demonstrate an opportunity to improve, for example, the simulations of gas-solid two-phase flows [24, 25]. In this context, Mazhar et al. [26] developed a remarkable parallel collision detection algorithm for multibody problems suitable to be carried out in two levels using CPU-GPU hybrid computing. This algorithm overcomes some limitations regarding the use of linked list, such as the process of selecting the optimal
cell size. However, it is focused on a hybrid computing platform regardless of the physical model underlying the detection. In this work, Mazhar’s detection algorithm is extended and specially modified to be independently executed in a multi-core CPU or in a single GPU.

Multi-thread computing is usually faster than the sequential one, specially in DEM simulations, but the decision whether to use a CPU or GPU code for multi-thread computing is not straightforward. The small- and medium-sized industrial powder-related sectors are sometimes reluctant to use large calculation clusters for simulation and modelling activities. At least in the ceramic tile manufacturing sector, the affordability and simplicity of the technology are a crucial factor in the modelling process. Hence the interest in being able to perform DEM simulations on non-specialized or low-cost equipment.

After conducting a literature review, it has been found that many studies simplify the computations to demonstrate the speed-up of the GPU computing versus CPU, although these simplifications could call into question the accuracy and reliability of the model. For example, Govender et al. [27] showed an important speed-up in DEM simulations via GPU computing, but using a history-independent contact model with a special mesh treatment [28]. On the other hand, Radeke et al. [21] claimed that their GPU code is able to simulate the behaviour of one million particles during one physical minute in about four days of wall-clock time. However, their code works on single-precision floating-point format, and a special treatment is needed to achieve a good accuracy [29]. Note that the GPU double-precision capability is significantly slower than the single-precision one [30]. Also the GPU code of Zheng et al. [31] exhibited an increment in efficiency of about 73 speed-up ratio compared with the CPU algorithm. However, they use a threshold value to limit the maximum number of particles in a cell and another threshold value to limit the maximum
number of contacts of a particle. These considerations allow the use of the shared memory in GPU computing, which increases the performance, and facilitates reading and writing of contact history. As a consequence, the code is limited to simple geometries and narrow particle size distributions. Recently, Gan et al. [32] used a new algorithm [33, 34] to evaluate its performance on a single-GPU, multi-GPU and CPU configurations. They demonstrated that a single GPU outperforms CPU when simulating mono-sized particles. This also occurs for a multi-GPU computing with elliptical particles, which was industrially validated [35]. However, even though history-dependent models can be used, it requires to limit the maximum number of contacts of each particle. In fact, the use of threshold values that limit the number of contacts of each particle is still used in the latest reference dealing with GPU algorithms [36].

In response, opposing studies have also arisen that exemplify the scepticism that exists regarding the vast superiority of GPUs over CPUs in the massive parallelization of scientific computing algorithms. Lee et al. [23] showed how, with regard to scientific calculation, the enormous differences announced between the two processing units are of dubious credibility, while other studies showed that over-simplistic programming can lead to disappointing results [37]. Regarding DEM, Shigeto et al. [22] emphasized that the comparison between the use of both calculation architectures must be realistic and fair, evidencing that the enormous differences between them revealed by some researchers are far from fulfilling both principles. In addition, Washizawa et al. [38] stated that the computational speed of a practical DEM model is significantly slower to run on a GPU than on a CPU.

As it has been seen, there is a lack of clarity comparing CPU and GPU algorithms in equal conditions. In this work, an extension and modification of the Mazhar’s detection algorithm is introduced in an in-house developed DEM solver. This proposal
enables the use of complex physical models and do not require expensive nor technically complex HPC facilities to simulate large particle assemblies, but it can still exploit the multi-core features of conventional desktop or laptop computers. Furthermore, no limiting parameters are required that have to be defined manually at the beginning of the computation. Additionally, the key ideas of this proposal have been extended and integrated in two different versions of the solver, suitable for running on a multi-core CPU and a single dedicated GPU, respectively. Both solvers have been benchmarked against LIGGGHTS.

The present modified solvers still keeps the advantages of the Mazhar’s algorithm but it also covers very important aspects:

(i) An efficient layout to cope with the particularities of the physics of the contact model (storage and access to the contact history, regardless of the number of contacts).

(ii) A general treatment for the boundaries, whose handling is similar to that of powder particles. Definition of the calculation domain is never needed.

(iii) An efficient layout allowing for running large-scale simulations in low performance equipments with non-abundant RAM memory.

(iv) A simple layout that avoids the use of manually predefined parameters, which facilitates the usage of the solver without compromising the reliability of the results.

This paper is organized as follows: the process flow for DEM simulations is summarized in Section 2; the extension of Mazhar’s algorithm for multi-core CPU and single GPU platforms are explained in Section 3; Section 4 compares the performance of the two algorithms showed in Section 3 with LIGGGHTS, and discusses their applicability; finally, a few conclusions with several final remarks are presented. The
equipment and programming specifications used in this work are included in the Appendix A.

2. The scheme of the DEM simulations

The general DEM scheme consists in updating the kinematics and dynamics of a system constituted by the so-called "discrete elements". These elements are physical objects that can interact with each other and can represent either individual particles or any other basic geometric entity capable of interaction. The physics governing the contact between the discrete elements depends on the selected constitutive models. As presented in Part I, these models are described by a set of mathematical equations representing the material response under different loads. Although simplifications are always assumed, the main physical features must be retained in order to agree with the experimental observations. The equations of motion incorporating these physically-based contact models are numerically integrated, where the contact forces and transferred moments between elements are continuously computed over time. Therefore, DEM is a very suitable simulation method to accurately describe the complex motion of big assemblies of elements like rocks, stones, powders, granules, seeds, etc. Conversely, DEM is not suitable to study the stress and strain fields inside the elements, unless coupling with other continuous-based numerical method is considered, like the finite element method (FEM).

In the present work, the granules of the spray-dried powder are physical objects modelled as spheres. The boundaries of the setup confining the powder (e.g., rotating drum, mixer, mould feeder, conveying belt, etc.) are meshed using triangles.

Fig. 2 shows a general DEM flowchart. First, all variables are initialized, as the domain parameters, the contact model, the integrator scheme and the mesh movements, amongst others. Second, the geometry and particles are indexed in the
simulation scope and the data structure is created. Once all parameters are prepared, if the simulation involves GPU computing, the requested information is copied to the GPU.

Subsequently, contacts between objects are detected, the forces are calculated and finally an integrator scheme is applied in order to update the particle positions. Afterwards, if these stages are executed on the GPU, the relevant information is copied back to the CPU. These operations are repeated iteratively until the final time is achieved. Furthermore, if the confining boundaries experience any type of motion, all their geometrical elements must be updated at the beginning of the next time increment. Generally speaking, the three basic stages in a DEM solver are:

(i) The collision detection algorithm that foresees and registers all the contacts between the elements.

(ii) The calculation of forces.

(iii) The numerical integration scheme that updates all positions.

Likewise, the collision detection process is typically the most demanding computational stage, followed by the force calculation and finally by the integration scheme [39]. The simplest searching algorithms to detect pair of elements under interaction usually lead to a computational complexity of $O(n^2)$, meaning that the amount of time and memory scales as the number of elements is squared [40].

However, the most efficient and sequential collision detection algorithms can lead to a computational complexity of $O(n)$. These algorithms are based on performing the so-called “space subdivision” process, consisting in dividing the physical space in small “cells” or “bins”. This process greatly reduces the amount of evaluations because only the element pairs inside neighbouring cells are checked. This subdivision can be performed by using uniform bins, quadtrees, octrees or by any other topo-
logical arrangement. The evenly distributed bin approach is commonly used due to its good ratio performance to implementation cost. Having said that, the main differences between algorithms rely on the structure and management of the data [41]. Moreover, this subdivision is fully independent of the computational platform.

With the aim of reducing the computational cost even more, all these algorithms can be implemented taking advantage of the parallelism of modern multi-core platforms. Here it is important to remark that computational complexity $O(n)$ is unaltered, as $n$ is an invariant.

This work proposes a set of algorithms focused on reducing the computational cost by means of parallelizing the collision detection and force computation stages. The algorithms are an extension of Mazhar’s algorithm for multi-core CPU and single GPU platforms. The description of how the initialization and the integration stages are parallelized, as well as the boundary movements, is omitted in this paper due to their minor impact on the final performance of the algorithms. Finally, note that this paper presents two independent algorithms for CPU and GPU computing, but not for a hybrid computing like the original proposal [26].

3. Algorithms for DEM computing

Traditionally, contact detection for DEM simulations has been performed using a two-phase approach based on a linked-list structure [14]. A linked list is a linear data structure where the elements are not stored at contiguous memory locations. It is a dynamic data structure which enables a fast insertion and deletion data, and efficient memory utilization. This structure is used by most DEM software, such as LIGGGHTS [12]. Despite its efficiency, this strategy has two major drawbacks. Firstly, all cells of the domain have to be evaluated even though they do not contain
objects. If the domain is big, it can lead to important memory limitations. Secondly, for the contact evaluation, searching is limited to all objects in the given cell and in the neighbouring cells (2 cells in 1D, 8 cells in 2D and 26 in 3D). Therefore, the minimum size of a cell has to be the same as the largest element in the system to ensure a correct evaluation. It involves an important limitation if the elements are very different in size, because it worsens load balance on multi-core platform. Poor load balancing can be avoided by increasing the cell size, however, decreasing the number of cells might imply an increase in the order of complexity of the algorithm. Consequently, a pre-analysis step is usually required to run efficient DEM computations [42].

In order to avoid these disadvantages, in this work a strategy inspired by that used by Mazhar et al. [26] has been developed. This new proposal extends the Mazhar’s strategy, describing how to track the previous contact history, required in some constitutive models, and reducing dramatically the memory usage. The latter is specially relevant to simulate big assemblies of discrete elements (>1 million of elements).

3.1. Boundary treatment

As stated above, in this work the boundaries are meshed using triangles. This triangles are, in turn, decomposed into simpler elements, also considered as discrete elements (or boundary elements). Consequently, the boundary becomes part of the data structure of the model, consisting of granule, face, edge and vertex type elements. Figure 3(a) exemplifies a mesh. Note that while all faces must be part of the data structure of the model, not all the edge and vertex elements need to take part on it. Only the outer edges and vertices are likely to be in single contact with the granules (figure 4). On the one hand, the outer edges are considered to be those that
belong to several faces, as long as one of them is not coplanar with the others, or
those that belong only to one face. On the other hand, outer vertices are only those
that belong to non-collinear edges.

Since a granule can be in contact with several geometric elements at the same
time, it is necessary to establish rules to avoid multiple contacts arising from the
same boundary. Therefore, when detecting a contact with a boundary element, it is
necessary to validate the contact to ensure its uniqueness. The checking rules are
defined as follows:

1. A granule is in contact with a face element. It is validated without any addi-
tional condition.
2. A granule is in contact with an edge. It is validated only if the granule is not
in contact with a face containing that edge.
3. A granule is in contact with a vertex. It is validated only if the granule is not
in contact with any edge or face connected to that vertex.

The proposed methodology is inspired by the procedure established by Su et al. [43] to
perform simulations of particles flow in arbitrary complex geometries. Figure 4 shows
the three possible scenarios of a granule in contact with the boundary. In scenario 1,
the centroid of the granule can be projected into the face and the granule-face contact
is detected. In scenario 2, the centroid of the granule cannot be projected into the
face, so the granule-face contact is not identified. Thus, a granule-edge contact is
detected. In scenario 3, the centroid of the granule cannot be projected neither into
the face nor into the edge, so that only the granule-vertex contact can be detected.

The checking rules stated above must be applied for each detection.
3.2. DEM Algorithm on a CPU

The following algorithm is based on Mazhar’s algorithm [26], whose memory management is excellent. It is focused on the sorting, which is extremely fast in GPU computing, specially using the radix sort. A similar strategy can be addressed to the multi-core CPU computing introducing some variations in the original proposal.

On the other hand, the original strategy only considers sphere-sphere interactions because the domain is spherically decomposed. How to track the contact history is not discussed. In the following, a general algorithm for DEM simulations focused on multi-core CPU computing is explained together with the tracking of the contact history. The latter feature enables the use of realistic contact models, such as those which consider the tangential micro-slip [44] during the contact.

Stage 1. Indexation. This stage begins by identifying all the intersections produced between the elements and the cells surrounding them. Figure 3(b) shows a geometry and the cells surrounding that geometry, where only the cells intersecting those elements are indexed. The size of the cells is defined as twice the mean diameter of all the granular elements.

For each element, the maximum and minimum points that delimit its position in the domain are determined by means of a bounding box and the cells owning those points are identified. Figure 5 shows an example of the bounding box and the bounding points of a granular and edge element.

The number of cells ($N_{\text{cells}}$) intercepting any granular element is given by:

$$N_{\text{cells}} = (I_{\text{max}} - I_{\text{min}} + 1)(J_{\text{max}} - J_{\text{min}} + 1)(K_{\text{max}} - K_{\text{min}} + 1) \quad (1)$$

where $\{I_{\text{min}}, J_{\text{min}}, K_{\text{min}}\}$ and $\{I_{\text{max}}, J_{\text{max}}, K_{\text{max}}\}$ are the cell coordinates of the bounding points of that element (this terma is herinafter denoted by Coord). In the 2D case presented in figure 5, the granular element occupies 4 cells and the edge
element occupies 3 cells. Following this construction, it can be noted that a vertex
element only occupies 1 cell. It is important to remark that faces (or edges) are, in
general, arbitrarily oriented, thus generating a bounding box that may span a big
number of cells, as equation 1 states. Obviously, this situation is highly inefficient
because most of the cells in the bounding box may be empty. To overcome this
drawback, this work adopts the algorithms proposed by the references [45] and [46]
to efficiently index the cells surrounding edges and faces, respectively. The algorithm
to index the cells surrounding edges registers only the cells visited by a 3D line
with endpoints on the grid. The algorithm to surround the faces by cells scans the
original bounding that encloses the face and then applies a filter equation. This
filter evaluates the distance between the cell and the face, wherein only the cells
intersecting the face are labelled and stored.

In order to parallelise this indexation stage, the supporting arrays A1 and A2 are
defined. A1 has a size equal to the number of elements, and its function is to store
the number of cells occupied by each element, as figure 6 shows. Array A2 is used to
save the cumulative sum of A1, wherein each entry contains the memory shift of the
corresponding location in A1. Counting the number of cells occupied by each element
and the creation of A2 can be done in parallel without risk of race conditions\(^1\), since
the number of elements in the simulation is known at the beginning of this stage.

**Stage 2. Grouping the neighbouring contacts.** The next step is to put in con-
tiguous cells all those elements that may potentially be in contact. This is done in
the same way as in Stage 1, but instead of counting the number of cells, the cell

\(^1\)A race condition is an undesirable situation that occurs when two or more execution threads
perform operations at the same time over shared resource, and the results may change depending
on the order of execution.
information is stored. For this purpose, a new array A3 is created. A3 includes the 
Coord variables that contains a given element as well as the element ID (ID_Elem).

Figure 7 shows this operation, where it can be seen that thanks to the memory shifts 
saved by A2, this process can be fully parallelised.

Finally, A3 is sorted according to Coord, so that all elements occupying the same 
cell are arranged contiguously. It is worth mentioning that two elements occupying 
the same cell is a necessary condition but not sufficient to set up contact.

Stage 3. Contact identification. The main differences between the proposed 
algorithm and the one proposed by Mazhar start at this stage. After sorting, A3 is 
divided into as many chunks as threads can be simultaneously launched in the CPU. 
Then, each thread identifies the real contacts in the cells of its chunk.

Figure 8 exemplifies the contact identification with 3 parallel threads: all posi-
tions of A3 corresponding to the same cell are coloured in the same colour, and the 
ID_Elem is the number that appears in each position of the array. The different 
line styles correspond to the contact evaluations made by processing threads 1, 2 and 
3, respectively. On the other hand, the colour of the arrows indicates the number 
of iterations performed by a thread for each position in A3. For example, focusing 
on the first chunk (thread 1), a forward evaluation with 4 iterations is performed as 
follows:

1. Checking contacts between elements 1-4, 1-10 and 1-41.
2. Checking contacts between elements 4-10 and 4-41.
3. Checking contact 10-41.
4. Checking contact 3-5, 3-7, 3-32 and 3-33.

It can be noted that this methodology might lead to the identification of a particular 
contact in different cells. Nevertheless, the contact uniqueness in the cells can be
ensured by means of the collision centre, a concept suggested by Mazhar for the
granule-granule contact. This idea is here extended to deal with the granule-vertex,
granule-edge and granule-face contacts. In such cases, the centre of collision coincides
with the projection of the centre of the granule on the corresponding element.

Stage 4. Forces and moments calculation. This stage can be performed efficiently
and without race conditions by incorporating a critical section \(^2\) in the code after the
contact checking.

Algorithm 1 exemplifies this procedure for two granular elements, although it is
also applicable to evaluate the contact between other element types. In Lines 3-5
the forces and moments are computed. These magnitudes are subsequently added to
each element in Lines 14, 15, 18 and 19. The procedure is free from race conditions
by means of a locking mechanism, as showed in Lines 13, 16, 17 and 20. Once the
elements in contact are known, the application of the physical model is direct as
long as the previous contact history is not required. However, most of the physical
models need to retrieve the previous contact history to calculate the current forces
and moments \([47]\).

Algorithm 2 describes how the previous contact history can be extracted if it ex-
ists. Once the contact is identified, the contact history is retrieved from a contact list
that each granule is associated with. First, in Lines 7-12 the lowest indexed granular
element that constitutes the contact is checked to see if it contains information from
a previous contact with the second element. If so, this information is returned in
Line 10. Otherwise, a new contact is created in Line 14 and appended in Line 15 to
the contact list of the lowest indexed granule element. The used contacts are marked

\(^2\)A critical section is a shared resource of a computer that can only be accessed by one execution
thread at a time.
as active in Lines 9 and 14 to track the previous history used in each iteration. Once the forces and moments are computed, the contact history is rewritten with the new information generated so that it can be retrieved in the next iteration. The process can be boosted by using a stable sorting algorithm in A3, which maintains the original relative order of the elements. This ensures that the elements are sorted not only according to the cell coordinates but also by their identifiers (figure 8). Finally, all contacts that are not marked as active are removed from the contact lists.

3.3. DEM Algorithm on a GPU

The code developed to run on a single GPU is inevitably more complex than that for the CPU due to restrictions in programming language, memory and communication management. On the one hand, the maximum amount of RAM available on a GPU is typically less than that available on a CPU. Therefore, developing an effective GPU-oriented code enforces an efficient and wise memory usage. On the other hand, while the GPU has a large number of cores (allowing massive parallelization of code), the performance of these cores is limited to run relatively simple code with a balanced workload to avoid the so-called “branch divergence” [48–50]. This happens when threads executed in a same block (or warp\(^3\)) run through different instructions controlled by conditional statements. This situation must be avoided as non-equally conditioned branch split can unbalance the workflow.

The following describes the algorithm used for DEM computing on GPU architecture, paying attention to the differences with respect to the CPU-based scheme.

Stage 1. Indexation. After transferring all the necessary information (geometry, global values, tags, etc.) from the system to the GPU memory, this step performs

\(^3\)A warp is a set of 32 threads within a thread block such that all the threads in a warp execute the same instruction.
the same actions as Stage 1 in the CPU-based architecture.

**Stage 2. Grouping of neighbouring contacts.** This stage is performed in the same way as Stage 2 for the CPU architecture.

**Stage 3. Contact identification.** This stage starts after grouping array A3 (obtained in Stage 2) using a stable sorting algorithm to reduce the branch divergence in the next steps. In order to improve the understanding of the following parts of the algorithm, Figure 9 contains a flowchart of the complementary arrays created from A3 and their usage. Arrays A4, A5 and A6 facilitate the efficient GPU storage and kernel executions. Arrays A7 and A7p store the contacts, while A8, A9 and A10 make it possible to store the force and moments of each contact without the use of atomic operations. Note that atomic operations, depending on its implementation mode, might lead to significant bottlenecks in the performance of a GPU algorithm [51].

To analyse A3 in parallel, as many execution threads as cells are executed. This strategy, different from the one performed in the section 3.2, reduces the branch divergence and helps to balance the workload on the GPU cores. It requires creating an array A4, which contains the cells offsets of A3.

Figure 10 sketches the layout of A4 and algorithm 3 shows its construction via pseudocode. The construction of A4 proceeds as follows. In Line 3, A4 is defined and allocated using the same size as A3 with all values initialized to $0xfff$. This hexadecimal number is the largest unsigned integer value of 32 bits possible on x86 architectures. In Lines 9 and 10, every position of A3 is analysed by separated threads where two conditions are checked:

(i) Whether that element is the first one in its cell.

(ii) Whether that cell contains more elements.

If (i) and (ii) are fulfilled, the index of its position is assigned to the same location in
A4 as line 11 shows. Finally, in Line 14 array A4 is sorted. This allows for keeping on the left side all values other than 0xfff and whose value is the position of the first element of each cell in array A3 that may contain contacts. The finalized array A4 can be seen in figure 10.

The next step is to identify the contacts in parallel. First, the number of contacts in each cell is identified and counted to perform the memory allocation of the contacts history. Secondly, the number of contacts in each cell is identified again, and their history is stored in the previously allocated memory space. It is noticeable that in the CPU algorithm the contacts are only identified once, since a dynamic memory allocation can be done on-the-fly.

The counting procedure is outlined in figure 11 and described via pseudocode in algorithm 4 as follows. In Line 4 an intermediate array (A5) is defined and allocated with a size equal to the number of cells with different values of 0xfff in A4 (hereinafter called active cells). In Line 6, as many processing threads as number of active cells in A4 are executed. Each thread counts the number of contacts of its assigned cell by brute force, as described in Lines 9-20. Finally, in Line 21 the total number of contacts of the cell is stored in A5.

Similar to the construction process of A2 from A1 (see figure 6), the array A6 is created to store the accumulated sum of A5. Once the number of contacts is known, a new array A7 is allocated, where the contact history can be written without race conditions. A7 is filled by re-identifying the contacts and using A5 and A6, similarly as performed with A3 (see figure 7).

Previous contact history can be efficiently obtained by an extra step after writing the array A7. To do that, two arrays are kept in memory: A7 for the current instant (t) and A7p for the previous instant (t-Δt). The only requirement is to know the position of every contact at both instants, as long as both exist.
Figure 12 and algorithm 5 show the process of linking A7 to A7p with a sketch and pseudocode, respectively. In Line 5, a parallel thread is executed for each component of array A7. In Line 7, every thread has to identify whether the current contact existed at t-Δt. If so, the cell coordinates containing this contact history (Old_Coord) are retrieved, otherwise, a null value is returned. This checking is needed because it might happen that the cell that evaluates the contact changes with the movement of the elements. Line 8 checks if the current contact is new, in which case its history is initialized in Line 9.

If the current contact exists in A7p, Old_Coord is now known and the cell where the contact occurred is searched in A7p using a binary search, as shown in Line 12. The binary search, of $O(\log n)$, is possible because A7 is naturally ordered by cell coordinates. After finding the cell, the contact is found using a linear search. In Lines 13-22 a backward linear search is described. Line 15 checks whether the elements in contact are those desired. If so, Line 16 updates the history in A7. If not, Line 19 reduces the array index. If after completing the backward linear search the contact history has not been found, Lines 22-30 perform a forward linear search in the same way.

Finally, note that the branch divergence can be reduced by evaluating the contacts in a staggered manner (first the granule-granule contacts, second the granule-face contacts...).

Stage 3. Forces and momentums assignment. This part also requires extra stages compared to the CPU code, due to the peculiarities of the GPU, which imply maximizing code parallelization while minimizing code concurrency. Figure 13 sketches this assignment and algorithm 6 describes the procedure via pseudocode. Arrays A8, A9 and A10 are created and allocated with a size equal to twice the size of A7 and initialized to 0xfff as shown in Lines 4-7. In Line 8, as many processing threads as
number of contacts are executed. In Line 9, the forces and moments of each contact are computed by the corresponding thread. For the first element in contact, its identifier, force and moment are set in the corresponding positions of A8, A9 and A10 as observed in Lines 10-12. For the second element in contact, Line 13 checks whether it is a granule element. If so, its identifier, force and moment are set in Lines 14-16.

In Line 19, A9 and A10 are sorted by key, using as key A8. Then, A9 and A10 are reduced in Line 20, and the position of the first geometrical element in A8 is checked in Line 21. After these operations, A9 and A10 contain the total forces and moments for each granule element. Finally, in Lines 22-25 the new forces and moments are transferred to the granule elements without race conditions and in parallel.

4. Performance Analysis

The algorithms previously described were implemented on both multi-core CPU and single GPU platforms, and were analysed together with the software LIGGGHTS. The equipment and programming specifications used for the analysis are described in Appendix A. The three DEM codes were compared among them to analyse the effect of number of granules and its polydispersity in their performance. The capabilities of each code were evaluated simulating a mould filling, which is a crucial stage in the ceramic tile forming process. Basically, it involves pouring in a mould, via a transportation system, a certain amount of powder (usually a spray-dried powder).

The study consisted in filling a small mould (figure 14). The system was composed of three objects: a hopper, a feeder and a mould. In total, eight different Grain Size Distributions (GSD) were tested with a median granule size in volume ($d_{50}$) of 3 and 0.5 mm and a geometric standard deviation ($\sigma_{geo}$) of 1, 1.2, 1.4 and 2, respectively. In the ceramic tile industry, tiles manufactured in a mould of this size are considered to be small-sized tiles, so the results obtained here can be relatively extrapolated to the
industrial mould filling. Table 1 summarises all simulations performed in this section. Each simulation was labelled depending on its group, polydispersity and the code used. The simulations were split in two groups: group A includes the simulations with a $d_{50}$ of 3 mm, and group B those with a $d_{50}$ of 0.5 mm. Three DEM codes were tested: the code developed to be used in a CPU platform, the one to be used in a GPU platform and the code of LIGGGHTS executed in a CPU platform. In each group, and for each code, four simulations were executed with different levels of polydispersity. Table 1 also includes the number of granules and the timestep used in each simulation.

4.1. Simulation conditions

Figure 15 depicts the three different phases of the mould filling simulations, taking simulation A4 as example. Initially the powder is poured into the feeder while the hopper is moving backwards (figure 15(a)). Following, the feeder moves forward depositing the powder into the mould (figure 15(b)). Finally, the feeder returns to its original position (figure 15(c)).

The number of granules varied between $8.7 \times 10^4$ and $1.3 \times 10^6$, depending on the simulation. The physical model used was the Linear Spring-Dashpot (LSD) model [52]. Imposing properties that require very low time steps in integration does not make sense given the comparative nature of the analysis. However, similar values for granule density and stiffness to those calibrated in [52] were used. All properties and parameters were kept constant, independently of the code used. However, depending on the granular distribution used, the time step was revised. On the other hand, LIGGGHTS incorporates the LSD model in a different way from the codes developed [12], so in order to impose a similar behaviour in all codes, it was decided to eliminate the rolling resistance and the viscous damping. Table 2 shows the
properties of the granular material used in all cases in this study. In addition, the
default cell size of LIGGGHTS was used, which coincides with the maximum size of
the largest granule.

Powder distribution was generated previously in order to fill the hopper by gravity. As a consequence, all initial conditions in the mould filling simulations contained
the particles deposited on the hooper, which reduced the physical time of the simu-
lations.

4.2. Results

Performance results conducted are described below. The performance was evalu-
ated by measuring the total time of each simulation using the time command provided
by Linux. The results are detailed in two separate groups as given in table 1, and
discussed later.

4.2.1. Group A

Figure 16 compares the average elapsed time per iteration as function of the
number of contacts in different simulations with powder distributions of a $d_{50} = 3$
mm. Note that large differences are observed in the number of granules used for each
experiment due to the different degrees of dispersion. While for a GSD with $\sigma_{geo} = 2$
the hopper was completely filled with 88000 granules, for $\sigma_{geo} = 1$ a total number of
504000 granules was required. All plots of figure 16 are equivalent and the number of
granules and its dispersion does not change the fact that the LIGGGHTS code is the
one that works best, followed by the CPU and GPU code. The order of complexity
of the algorithms is, for each case, linear with respect to the number of granules.

The differences observed in calculation speed of each simulation are mainly at-
tributed to the large differences in the number of granules in each experiment. Like-
wise, the differences between the three algorithms are reduced with the decrease in the number of granules.

Regarding to the GPU code, in figure 16(a) the maximum difference observed between the GPU code and the LIGGGHTS one is approximately 300 ms, while in figure 16(d) the difference is only about 20 ms. As discussed in Section 3.3, issues such as code divergence, memory allocation, and too complex threads penalize the GPU code.

4.2.2. Group B

Figure 17 illustrates the average elapsed time per iteration for a given number of contacts in different simulations with powder distributions of $d_{50} = 0.5$ mm. These GSD are similar to the true distribution of spray-dried powder used in the ceramic tile industry [53]. The results shown here are hence of great relevance.

In these cases, the $d_{50}$ is so small that in none of the simulations of group B it was computationally possible to fill the hopper completely, because of the time that would be required to complete the calculations (more than 10 million granules would be involved). Consequently, a number of granules around 1 million was considered sufficient to evaluate the code. This enabled all the calculations to be completed in a reasonable time. It is important to mention that the differences in the number of granules used in these simulations were negligible. This statement was confirmed by carrying out complementary performance analyses using exactly the same number of granules (1 million).

As in group A, the order of complexity of the algorithms is, for each case, linear with respect to the number of granules. Figure 17 shows that code developed for CPUs outperforms the code developed for GPUs, even more so than the LIGGGHTS code. These differences are reduced by increasing $\sigma_{geo}$. It is remarkable that the
simulation with a value of $\sigma_{\text{geo}} = 1$ (figure 17(a)) could only be done with the CPU code, since the RAM memory needed to run the simulation on the other two codes was larger than that available. Table 3 provides the maximum RAM consumption of each simulation shown in figure 17. It highlights the limitation of using a single GPU in DEM simulation, due to the amount of RAM available. It also demonstrates the bigger amount of RAM required by the LIGGGHTS code, since equipment with which the simulations have been performed has 32GB of RAM (table A.1). Regarding the other simulations (figures 17(b), 17(c), 17(d)), it was possible to perform them without any trouble.

Memory availability is specially important in the LIGGGHTS code, since it is required to explicitly define a calculation domain and store information of all the cells present in it. This implies that the smaller the size of the cells and the larger the domain, the more memory is required. In contrast, in the CPU and GPU codes, the impact is much smaller, as the domain is not fully indexed, and the size of the cells only influences the construction of the contact identification array $A_3$ (figure 7).

On the other hand, as $\sigma_{\text{geo}}$ increases, the default cell size of LIGGGHTS increases. This involves two phenomena: (i) RAM consumption decreases. (ii) The number of contacts per cell increases, so the resolution tends to $O(n^2)$. Here, the increment leads to a drastic reduction in memory consumption, and a substantial improvement in the speed of the algorithm is observed. Therefore, the impact of memory management on LIGGGHTS performance is greater than the impact of increasing $\sigma_{\text{geo}}$, even if it means an increment of the number of evaluations of possible contacts per cell.

This pattern is shown in figure 18, where the slope of the trend lines drawn in figure 17 is plotted as a function of the value of $\sigma_{\text{geo}}$. The impact of $\sigma_{\text{geo}}$ on the scalability of the codes developed is very low. However, the impact on LIGGGHTS is very important.
4.2.3. Discussion

Simulations performance exhibits large differences depending on the number of contacts involved. While in the simulations of the group A the LIGGGHTS code is superior, in the simulations of the group B the CPU and GPU codes overperform LIGGGHTS. RAM consumption seems to be decisive for LIGGGHTS code performance, as observed in table 3, where the enormous difference in memory consumption of the algorithms is evidenced. Although this indicator alone does not determine the quality of the algorithm, excessive memory use can expose problems in its management, such as cache conflicts [54] and fragmentation [55].

With respect to GPU code, the results generally show much lower memory consumption than LIGGGHTS code, but on the order of 4 times higher than CPU code. This is due to the singular data structure used in GPU code, which needs to store and replicate information in memory due to the limitations of the architecture. Note here the huge advantage of the simulation algorithm developed over the one used by LIGGGHTS, since with the available RAM in the GPU (5Gb), the possibilities of the LIGGGHTS resolution scheme in this architecture would be really restrictive. In both developed codes, the reduction of memory with the increased dispersion is attributable to the reduction in the size of the array A3.

Finally, most of the results shown in figure 17 have hysteresis, specially the LIGGGHTS and GPU code. This indicates that for the same number of contacts there are parts of the calculation where the algorithm performance is different. The explanation for this hysteresis is shown in figure 19. This figure shows the typical evolution of the number of contacts in the simulations of the group B. In all the sub-figures there are two differentiated areas where the number of contacts is extremely high, corresponding to the filling of the feeder and to the moment when all the gran-
ules are on the mould. In figure 19(a) the different regions of the curve are associated with the instant of filling shown in figure 15.

The first zone (between iteration $80 \cdot 10^3$ and $300 \cdot 10^3$ for figure 19(a)) corresponds to the values of the upper part of the hysteresis in figures 17(b), 17(c) and 17(d). The second zone (between iteration $400 \cdot 10^3$ and $600 \cdot 10^3$ for the figure 19(a)) corresponds to the lower part of the hysteresis. The explanation of hysteresis may be related to memory management. The continuous increase of the number of contacts in the first zone necessarily involves a continuous memory allocation, while in the second zone the information can simply be reallocated. Therefore, in the second part of the simulation there would be no such over-cost and there is a greater probability of memory access optimization. This behaviour is more difficult to observe in the CPU code. In the first place, in this code the contacts are not stored in a single container, so the optimization in the memory access is not so evident. Secondly, the CPU code has a much lower RAM consumption than the other two, so the problems associated with memory management are less visible.

As a general summary, figure 20 shows the average calculation time per iteration as a function of the number of granules. Each algorithm and all the simulations carried out have been included with the exception of figure 19(a), whose execution was, considering our hardware, only feasible with the CPU code.

Considering all simulations, the order of complexity of the CPU and GPU codes remains linear, while the LIGGGHTS code shows a fairly exponential one. For scenarios with a low number of granules ($< 0.5 \cdot 10^6$) the LIGGGHTS code performs better than those developed. In scenarios with a higher number of granules ($> 0.5 \cdot 10^6$), the CPU code outperforms all others. GPU code is also faster than LIGGGHTS when the number of granules is greater than $0.9 \cdot 10^6$ and small ($d_{50} = 0.5$ mm).

Finally, it is worth mentioning that the performance of the GPU code is not
as expected. On the one hand, although the parallelization capacity of the GPU
architecture is excellent, the data has to be processed carefully. A lot of intermediate stages are required to perform efficiently DEM simulations on GPU, hence the code results less clean and comprehensible than the CPU one. On the other hand, although it can be minimized, DEM simulations involve some divergence and may cause memory bank conflicts in its execution. This is the case during the heterogeneous contact evaluation (different element types in contact have different mathematical treatment, the elements to evaluate are not necessarily adjacent in memory, etc.) and the searching of the previous contact history, amongst others. Unfortunately, GPU computing has an important performance penalty in those situations [56, 57].

On the other hand, in all cases the CPU code is faster than the GPU code. This differences are more noticeable as the number of granules increases. As a last remark, it is important to note that the conclusions drawn in figure 20 are subject to the cases tested in this work, and the software and hardware conditions used. Therefore, the conclusions are not necessarily extrapolated to other scenarios (for example, to scenarios where, despite the number of granules, the number of cells in the domain is not large enough to significantly increase RAM consumption) and other computer equipment.

5. Conclusions

This paper presents two algorithms that overcome the restrictions of the traditional DEM resolution algorithm (based on a linked-list approach). The former is aimed to be used on a multi-core CPU, and the latter in a single GPU. Both algorithms make an efficient use of the available RAM on the system, and are not limited by the domain calculation size. The proposed algorithms are a modified and extended version of the algorithm proposed by Mazhar et al. [26]. Both codes have
been compared with the reference software LIGGGHTS by means of simulating a mould filling as performed in ceramic tile manufacturing.

Regarding to the LIGGGHTS code, a direct relationship between the memory consumption of LIGGGHTS (table 3) and its performance has been observed. Indeed, the high computational cost of LIGGGHTS code in some situations is attributed to management issues that can occur when RAM memory consumption is so high. Although it has not been studied, it would be possible to increase the cell size in the LIGGGHTS code in those simulations with higher memory consumption. This would mean decreasing RAM memory consumption, although it would also imply an increase in the number of granules in each cell (with the increase in the number of evaluations that this involves). A previous study should be perform to identify the optimal cell size.

With respect to the GPU code, the following conclusions have been reached:

• RAM limitation is a significant handicap of using a single GPU for DEM simulations. This drawback might be overcome by using a multi-GPU platform or a CPU-GPU hybrid computation.

• Contact identification is much more complex to do on a GPU than on a CPU, involving extra steps and making the code much less maintainable and readable.

• Despite using a high-performance graphics card (Kepler K20c), the developed code for a single GPU performs significantly worse than the two CPU-based codes analysed. It must be highlighted that the calculations always used double-precision floating-point format and the granular system was formed by a polydisperse size assembly of particles, where contact history is accurately tracked for an arbitrary number of contacts. These aspects penalize tremendously the GPU performance.
The conclusions reached concerning the CPU code are as follows:

- CPU code has proven to be the most versatile and balanced of the three codes.
- It is efficient for simulating both large and small quantities of granules.
- Its performance is markedly superior to the other codes when the number of granules to be simulated is high, as in industrial scenarios.
- Its low memory consumption allows to run demanding simulations in conventional computers without a high memory capacity.

According to the findings of this work, the CPU-based solver developed in this research will be used in Part III of this collection to validate and study the spray-dried powder model in a real filling system commonly used by the ceramic industry. The validation process will be performed by comparing the model predictions with the experimental results obtained from a prototype.

As final remark, it has been proven that the GPU architecture does not necessarily outperforms the CPU one when implementing DEM. This findings agree with the results of Lee et al. [23], Shigeto et al. [22] and Washizawa et al. [38]. However, Gan et al.[32] also showed that the use of a single GPU does seem adequate to simulate a maximum of about 1 million mono-sized and/or ellipsoidal particles.

In the author’s opinion, the computational cost of simulating fine powders by using spheres in DEM is not necessarily lower on a single GPU than on a multi-core CPU. The extension of Mazhar’s algorithm presented in this paper is a general-purpose algorithm that handles efficiently typical spray-dried powders used in ceramic tile industry. This extended algorithm enables the use of complex models that require double precision calculations and tracking of the contact history. Furthermore, the purposed code architecture does not rely on predefined threshold values.
to setup the optimum functioning of the solver. If a high-performance equipment is not available and a solver that requires predefined values to adjust its functioning is not desired, this research suggests that there are not substantial reasons to use a single GPU computing platform to run powder simulations with DEM.

6. Acknowledgements

The authors of this paper wish to thank MACER S.L and the CDTI Ministry of Science and Innovation of Spain, as part of the project with the reference number: IDI20140989. J.M. Tiscar wishes to thank the Conselleria de Cultura, Educació i Ciència de la Generalitat Valenciana for its financial support in the conducted study with a PhD scholarship, through the programme VALi+D for researchers training.

Appendix A. Equipment and programming specifications

Table A.1 shows the hardware and software specifications for the simulations. The total numbers of CPU parallel threads used in the simulations performed with the CPU and LIGGGHTS code was 8. Furthermore, the domain in LIGGGHTS was divided along the z-axis into 8 portions. This ensured that no thread was idle. Regarding the GPU code, the simulations were performed launching parallel thread blocks of 256 threads each of them. The number of blocks used depended on the parallel threads required. All calculations were performed in double-precision floating-point format.

The codes were written taking advantage of an object-oriented programming (OOP) approach and mechanisms of Run-Time Type Information (RTTI) when possible. OOP is a programming language model that organizes software design around data, or objects, rather than functions and logic. On the other hand, RTTI is a
mechanism that exposes information about an object’s data type at runtime. It provides a way to write cleaner code at the expense of an overhead. Unfortunately, GPU programming still has much work to do with respect to the implementation capabilities for OOP and RTTI. Although a procedural programming approach may be more appropriate to perform High Performance Computing (HPC) [58], OOP involves a natural implementation structure (easy to debug and maintain) which does not compromise excessively the efficiency and feasibility of the code.

References


N. Govender, D. Wilke, S. Kok, A large scale discrete element framework for NVIDIA GPUs, in: NVIDIA GTC 2015, California.


Table 1. Total of simulations in the performance analysis.

<table>
<thead>
<tr>
<th>Label</th>
<th>GSD</th>
<th>Number of granules</th>
<th>$\Delta t$ (µs)</th>
<th>Group</th>
</tr>
</thead>
<tbody>
<tr>
<td>CPU</td>
<td>GPU</td>
<td>LIGGGHTS</td>
<td>$d_{50}$ (mm)</td>
<td>$\sigma_{geo}$</td>
</tr>
<tr>
<td>A2-CPU</td>
<td>A2-GPU</td>
<td>A2-LIGGGHTS</td>
<td>3</td>
<td>1.2</td>
</tr>
<tr>
<td>A3-CPU</td>
<td>A3-GPU</td>
<td>A3-LIGGGHTS</td>
<td>3</td>
<td>1.4</td>
</tr>
<tr>
<td>A4-CPU</td>
<td>A4-GPU</td>
<td>A4-LIGGGHTS</td>
<td>3</td>
<td>2</td>
</tr>
<tr>
<td>B1-CPU</td>
<td>B1-GPU</td>
<td>B1-LIGGGHTS</td>
<td>0.5</td>
<td>1</td>
</tr>
<tr>
<td>B2-CPU</td>
<td>B2-GPU</td>
<td>B2-LIGGGHTS</td>
<td>0.5</td>
<td>1.2</td>
</tr>
<tr>
<td>B3-CPU</td>
<td>B3-GPU</td>
<td>B3-LIGGGHTS</td>
<td>0.5</td>
<td>1.4</td>
</tr>
<tr>
<td>B4-CPU</td>
<td>B4-GPU</td>
<td>B4-LIGGGHTS</td>
<td>0.5</td>
<td>2</td>
</tr>
</tbody>
</table>
Table 2. Simulation parameters in the performance analysis.

<table>
<thead>
<tr>
<th>Contact Granule-Granule</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Granule density, $\rho_g$ (kg/m$^3$)</td>
<td>2000</td>
</tr>
<tr>
<td>Normal stiffness, $k_n$ (N/m)</td>
<td>100</td>
</tr>
<tr>
<td>Tangential stiffness, $k_s$ (N/m)</td>
<td>100</td>
</tr>
<tr>
<td>Rolling stiffness, $k_r$ (N/m)</td>
<td>0</td>
</tr>
<tr>
<td>Damping normal, $\gamma_n$ (kg·m/s)</td>
<td>0</td>
</tr>
<tr>
<td>Damping tangential, $\gamma_s$ (kg·m/s)</td>
<td>0</td>
</tr>
<tr>
<td>Sliding friction, $\mu_{s,g-g}$</td>
<td>0.5</td>
</tr>
<tr>
<td>Rolling friction, $\mu_{r,g-g}$</td>
<td>0</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Contact Granule-Boundary</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Granule density, $\rho_g$ (kg/m$^3$)</td>
<td>2000</td>
</tr>
<tr>
<td>Normal stiffness, $k_n$ (N/m)</td>
<td>200</td>
</tr>
<tr>
<td>Tangential stiffness, $k_s$ (N/m)</td>
<td>200</td>
</tr>
<tr>
<td>Rolling stiffness, $k_r$ (N/m)</td>
<td>0</td>
</tr>
<tr>
<td>Damping normal, $\gamma_n$ (kg·m/s)</td>
<td>0</td>
</tr>
<tr>
<td>Damping tangential, $\gamma_s$ (kg·m/s)</td>
<td>0</td>
</tr>
<tr>
<td>Sliding friction, $\mu_{s,g-s}$</td>
<td>0.5</td>
</tr>
<tr>
<td>Rolling friction, $\mu_{r,g-s}$</td>
<td>0</td>
</tr>
<tr>
<td>Hopper velocity (m/s)</td>
<td>0.1</td>
</tr>
<tr>
<td>Feeder velocity (m/s)</td>
<td>0.5</td>
</tr>
</tbody>
</table>
Table 3. Maximum RAM consumption (MB), in figure 17.

<table>
<thead>
<tr>
<th>Group B</th>
<th>CPU</th>
<th>LIGGGHTS</th>
<th>GPU</th>
</tr>
</thead>
<tbody>
<tr>
<td>$d_{50} = 0.5$ mm</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>(a)</td>
<td>1070</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>(b)</td>
<td>1029</td>
<td>28534</td>
<td>4275</td>
</tr>
<tr>
<td>(c)</td>
<td>882</td>
<td>12175</td>
<td>3356</td>
</tr>
<tr>
<td>(d)</td>
<td>664</td>
<td>2835</td>
<td>2922</td>
</tr>
</tbody>
</table>
Table A.1. Hardware and software specifications for the simulations.

<table>
<thead>
<tr>
<th>Specifications</th>
</tr>
</thead>
<tbody>
<tr>
<td>Operating System</td>
</tr>
<tr>
<td>Ubuntu 14.04</td>
</tr>
<tr>
<td>Processor</td>
</tr>
<tr>
<td>Intel® Xeon® E3-1234 v3</td>
</tr>
<tr>
<td>• Frequency: 3.4Ghz</td>
</tr>
<tr>
<td>• Bandwidth: 25.6Gb/s</td>
</tr>
<tr>
<td>• Cores: 4</td>
</tr>
<tr>
<td>• Threads: 8</td>
</tr>
<tr>
<td>RAM</td>
</tr>
<tr>
<td>32 Gb</td>
</tr>
<tr>
<td>Graphic Card</td>
</tr>
<tr>
<td>NVIDIA® Kepler K20c</td>
</tr>
<tr>
<td>• Connection: PCI Express x16</td>
</tr>
<tr>
<td>• RAM: 5Gb</td>
</tr>
<tr>
<td>• Bandwidth: 208Gb/s</td>
</tr>
<tr>
<td>• Processor cores: 2496</td>
</tr>
<tr>
<td>CPU code</td>
</tr>
<tr>
<td>Shared memory parallel programming</td>
</tr>
<tr>
<td>parallelization</td>
</tr>
<tr>
<td>• Library: Intel® TBB</td>
</tr>
<tr>
<td>Compiling tools</td>
</tr>
<tr>
<td>• Programming Language: C++</td>
</tr>
<tr>
<td>• CPU: g++-4.8.4 with flag -Ofast</td>
</tr>
<tr>
<td>• GPU: nvcc-7.5 with flag -Ofast</td>
</tr>
<tr>
<td>• CUDA Toolkit 7.5</td>
</tr>
</tbody>
</table>
Fig. 1. Basic architecture differences between CPU and GPU.
Fig. 2. The basic scheme of DEM simulations
Fig. 3. Example of a geometry composed of a triangular mesh. Although all faces are considered, only the outer edges and vertices (green) are introduced into the model. (a) without grid. (b) with grid.

Fig. 4. Types of contact scenarios between the granules and the geometric elements.
Fig. 5. Bounding box of different objects. A uniform spatial subdivision is used to normalize the coordinates.

\[ \text{Coord}\{i, j, k\} : \]
\[ I = \text{int}(x/\text{BinSize}) \]
\[ J = \text{int}(y/\text{BinSize}) \]
\[ K = \text{int}(z/\text{BinSize}) \]

Fig. 6. Examples of Array A1 and A2. A1 results of intersection counting. A2 results of the partial summation performed on A1.
Fig. 7. Construction of array A3 in parallel, from arrays A1 and A2.

Fig. 8. (Color online) Schematic of how the contact evaluation works in array A3 sorted.
Fig. 9. Flowchart of the arrays created from A3, during the DEM algorithm on a GPU.

Array A3 sorted: Potential contacts identified

**Each bin color indicates a different cell**

Objects in the same cell:

1 4 10 41 3 5 7 32 33 2 3 12 6 8 20 30

- Auxiliary arrays for efficient GPU storage and kernel executions

A4: Identify first element of each cell
A5: Count the number of contacts of each cell
A6: Save memory shifts of A5

- Arrays containing each active contact history

A7: Store the history of each active contact (t)
A7p: Store the previous history of each active contact (t-Δt)

- Arrays containing the force and moments of each active contact

A8: Store the element index of each element in contact
A9: Store the force of each element in contact
A10: Store the moments of each element in contact
Fig. 10. Array A4 for the parallel evaluation of contacts.
Fig. 11. Contact counting in the GPU algorithm.

Fig. 12. Searching for a previous contact history.
Fig. 13. Construction diagram of arrays A8, A9 and A10.

Fig. 14. Mould filling system used in the performance analysis.
Fig. 15. Sequence of mould filling in the ceramic tile forming process. Simulation A4.
(a) $\sigma_{\text{geo}} = 1$, 504000 granules

(b) $\sigma_{\text{geo}} = 1.2$, 356000 granules

(c) $\sigma_{\text{geo}} = 1.4$, 224000 granules

(d) $\sigma_{\text{geo}} = 2$, 88000 granules

Fig. 16. Performance analysis of the group A ($d_{50} = 3\, \text{mm}$).
Fig. 17. Performance analysis of the group B ($d_{50} = 0.5$ mm).

(a) $\sigma_{\text{geo}} = 1$, 1316000 granules

(b) $\sigma_{\text{geo}} = 1.2$, 1122000 granules

(c) $\sigma_{\text{geo}} = 1.4$, 946000 granules

(d) $\sigma_{\text{geo}} = 2$, 1071100 granules
Fig. 18. The impact of $\sigma_{\text{geo}}$ on the scalability based on figure 17.
(a) $\sigma_{\text{geo}} = 1$, 1316000 granules

(b) $\sigma_{\text{geo}} = 1.2$, 1122000 granules

(c) $\sigma_{\text{geo}} = 1.4$, 946000 granules

(d) $\sigma_{\text{geo}} = 2$, 1071100 granules

**Fig. 19.** Evolution of the number of contacts in the group B simulations.
Fig. 20. Representation of the average calculation time per iteration as a function of the number of granules.
Algorithms
**Algorithm 1** Compute and assign forces and moments in parallel

**Require:**
1: $G_i$ \hspace{1cm} \triangleright$Granule i abstraction$
2: $G_j$ \hspace{1cm} \triangleright$Granule j abstraction$

**Initialize:**
3: $F = \{F_x, F_y, F_z\}$ \hspace{1cm} \triangleright$3D Force$
4: $M_i = \{M_x, M_y, M_z\}$ \hspace{1cm} \triangleright$3D Moment on $G_i$
5: $M_j = \{M_x, M_y, M_z\}$ \hspace{1cm} \triangleright$3D Moment on $G_j$

6: if not contact between granules then
7: return
8: end if

\triangleright Computation depending on the constitutive model
9: $F \leftarrow$ Compute force between $G_i$ and $G_j$
10: $M_i \leftarrow$ Compute moment on $G_i$
11: $M_j \leftarrow$ Compute moment on $G_j$

12: \{Start critical section\}
13: Lock $G_i$ \hspace{1cm} \triangleright$Mutual exclusion event$
14: Add $F$ to $G_i$
15: Add $M_i$ to $G_i$
16: Unlock $G_i$
17: Lock $G_j$ \hspace{1cm} \triangleright$Mutual exclusion event$
18: Add $F$ to $G_j$
19: Add $M_j$ to $G_j$
20: Unlock $G_j$
21: \{End critical section\}
Algorithm 2 Return contact history

Require:
1: $G_i$ \Comment{Granule $i$ abstraction}
2: $G_j$ \Comment{Granule $j$ abstraction}
3: $CList_i$ \Comment{$G_i$ contact list}
4: $i, j \in \mathbb{N} \mid i < j$

5: \{Start critical section\}
6: Lock $G_i$ \Comment{Mutual exclusion event}
7: for each Contact in $CList_i$ do
8: \hspace{1em} if Contact contains $G_j$ then
9: \hspace{2em} Set Contact as active
10: \hspace{1em} return Contact
11: \hspace{1em} end if
12: end for
13: Initialize $C_{new}$ \Comment{New contact history}
14: Set $C_{new}$ as active
15: Append $C_{new}$ to $CList_i$
16: return $CList_i$[end] \Comment{Unlock is auto-called}
17: \{End critical section\}
Algorithm 3 Construction of array A4

Require:
1: A3 \quad \triangleright\quad \text{Array A3}
2: n \quad \triangleright\quad \text{size of A3}

Initialize:
3: A4 \leftarrow \text{Array(size : } n\text{)}
4: Initialize all A4 values to 0xffff \quad \triangleright\quad 0xffff \text{ is the largest unsigned 32-bit integer on x86 architecture}
5: if A3[0].Coord == A3[1].Coord then
6: \quad A4[0] \leftarrow 0
7: end if
8: for in parallel: i \leftarrow 0 \text{ to } n - 2 \text{ do}
9: \quad if !(A3[i-1].Coord == A3[i].Coord) &&
10: \quad \quad (A3[i].Coord == A3[i+1].Coord) \text{ then}
11: \quad \quad A4[i] \leftarrow i
12: \quad end if
13: end for
14: Sort(A4) \quad \triangleright\quad \text{Stable sorting}
Algorithm 4 Counting Contacts

Require:
1: $A_3$ ▷ Array A3
2: $A_4$ ▷ Array A4
3: $n$ ▷ number of active cells in $A_4$

Initialize:
4: $A_5 \leftarrow \text{Array(size : } n\text{)}$
5: Initialize all $A_4$ values to 0

6: for in parallel: $i \leftarrow 0$ to $n - 1$ do

7: NContacts $\leftarrow 0$ ▷ Number of contacts
8: $j \leftarrow 0$
9: do
10: $k \leftarrow j + 1$
11: while $A_3[j].Coord == A_3[k].Coord$ do
12: E1 $\leftarrow A_3[j].Elem$ ▷ Element in $A_3[j]$
13: E2 $\leftarrow A_3[k].Elem$ ▷ Element in $A_3[k]$
14: if E1 and E2 are in contact then
15: NContacts $\leftarrow$ NContacts + 1
16: $k \leftarrow k + 1$
17: end if
18: end while
19: $j \leftarrow j + 1$
20: while $A_3[j].Coord == A_3[j + 1].Coord$ do
21: $A_5[i] \leftarrow$ NContacts
22: end for
Algorithm 5 Update array \( A_7 \) from \( A_{7p} \)

Require:

1: \( A_7 \) \hspace{1cm} \triangleright \text{Current contact Array } A_7 \text{ (t)}
2: \( A_{7p} \) \hspace{1cm} \triangleright \text{Previous contact Array } A_7 \text{ (t-Δt)}
3: \( n \) \hspace{1cm} \triangleright \text{size of } A_7
4: \( np \) \hspace{1cm} \triangleright \text{size of } A_{7p}

5: \textbf{for} \text{ in parallel: } i \leftarrow 0 \text{ to } n - 1 \text{ \textbf{do}}
6: \hspace{1cm} \text{Contact } \leftarrow A_7[i]
7: \hspace{1cm} \text{Old_Coord } \leftarrow \text{Cell coordinates where Contact occurs in the previous step}
8: \hspace{1cm} \textbf{if} \text{ Old_Coord } == \text{ null} \text{ \textbf{then}}
9: \hspace{1cm} \hspace{1cm} \text{Initialize } A_7[i] \hspace{1cm} \triangleright \text{No contact occurred}
10: \hspace{1cm} \hspace{1cm} \text{return}
11: \hspace{1cm} \hspace{1cm} \textbf{end if}
12: \hspace{1cm} \hspace{1cm} \text{pos } \leftarrow \text{BinarySearch}(A_{7p}, \text{Old_Coord}) \hspace{1cm} \triangleright \text{Backward linear search}
13: \hspace{1cm} \hspace{1cm} \text{l } \leftarrow \text{pos}
14: \hspace{1cm} \hspace{1cm} \textbf{do}
15: \hspace{1cm} \hspace{1cm} \hspace{1cm} \textbf{if} \text{ } A_7[i] \text{ and } A_{7p}[l] \text{ are the same contact} \textbf{then}
16: \hspace{1cm} \hspace{1cm} \hspace{1cm} \hspace{1cm} \text{Update } A_7[i] \text{ with } A_{7p}[l]
17: \hspace{1cm} \hspace{1cm} \hspace{1cm} \hspace{1cm} \text{return}
18: \hspace{1cm} \hspace{1cm} \hspace{1cm} \textbf{end if}
19: \hspace{1cm} \hspace{1cm} \hspace{1cm} \text{l } \leftarrow \text{l } - \text{1}
20: \hspace{1cm} \hspace{1cm} \hspace{1cm} \textbf{if} \text{ lft } < \text{ 0} \text{ \textbf{then} \textbf{break} \textbf{end if}}
21: \hspace{1cm} \hspace{1cm} \hspace{1cm} \textbf{while} \text{ } A_{7p}[\text{pos},.Coord ] \text{ } == \text{ } A_{7p}[l].Coord \textbf{end while}
22: \hspace{1cm} \hspace{1cm} \text{r } \leftarrow \text{pos+1}
23: \hspace{1cm} \hspace{1cm} \textbf{do}
24: \hspace{1cm} \hspace{1cm} \hspace{1cm} \textbf{if} \text{ } A_7[i] \text{ and } A_{7p}[r] \text{ are the same contact} \textbf{then}
25: \hspace{1cm} \hspace{1cm} \hspace{1cm} \hspace{1cm} \text{Update } A_7[i] \text{ with } A_{7p}[r]
26: \hspace{1cm} \hspace{1cm} \hspace{1cm} \hspace{1cm} \text{return}
27: \hspace{1cm} \hspace{1cm} \hspace{1cm} \textbf{end if}
28: \hspace{1cm} \hspace{1cm} \hspace{1cm} \text{r } \leftarrow \text{r } + \text{1}
29: \hspace{1cm} \hspace{1cm} \hspace{1cm} \textbf{if} \text{ r } > \text{ np } - \text{1} \text{ \textbf{then} \textbf{break} \textbf{end if}}
30: \hspace{1cm} \hspace{1cm} \hspace{1cm} \textbf{while} \text{ } A_{7p}[\text{pos},.Coord ] \text{ } == \text{ } A_{7p}[\text{r},.Coord ] \textbf{end while}
31: \hspace{1cm} \hspace{1cm} \textbf{end for}
Algorithm 6 Assign forces and moments to the elements

Require:
1: $A7$ \hspace{1cm} \triangleright \text{Contact Array } A7
2: $ElemLst$ \hspace{1cm} \triangleright \text{Elements list}
3: $n$ \hspace{1cm} \triangleright \text{size of } A7

Initialize:
4: $A8 \leftarrow \text{Array}(\text{size : } 2n)$ \hspace{1cm} \triangleright \text{Array of IDs}
5: $A9 \leftarrow \text{Array}(\text{size : } 2n)$ \hspace{1cm} \triangleright \text{Array of Forces}
6: $A10 \leftarrow \text{Array}(\text{size : } 2n)$ \hspace{1cm} \triangleright \text{Array of Moments}
7: Initialize all $A8, A9, A10$ values to $0xfff$

8: for in parallel: $i \leftarrow 0$ to $n - 1$ do
9: \hspace{1cm} Compute force and moments generated in $A7[i]$
10: $A8[2i] \leftarrow A7[i].ID_{\text{Elem}1}$
11: $A9[2i] \leftarrow A7[i].F1$ \hspace{1cm} \triangleright \text{Force in El.1}
12: $A10[2i] \leftarrow A7[i].M1$ \hspace{1cm} \triangleright \text{Moment in El.1}
13: \hspace{1cm} if $A7[i].ID_{\text{Elem}2}$ is a granule then
14: \hspace{2cm} $A8[2i + 1] \leftarrow A7[i].ID_{\text{Elem}2}$
15: \hspace{2cm} $A9[2i + 1] \leftarrow A7[i].F2$ \hspace{1cm} \triangleright \text{Force in El.2}
16: \hspace{2cm} $A10[2i + 1] \leftarrow A7[i].M2$ \hspace{1cm} \triangleright \text{Moment in El.2}
17: \hspace{1cm} end if
18: end for

19: SortByKey($A8,\{A9, A10\}$) \hspace{1cm} \triangleright \text{Sort an associative array } \{A9, A10\},
    \hspace{1cm} \text{according to the key } A8
20: ReduceByKey($A8,\{A9, A10\}$) \hspace{1cm} \triangleright \text{Aggregate the values of } A9 \text{ and } A10
    \hspace{1cm} \text{according to the reduction function applied on } A8

21: $fpos \leftarrow $ first position in $A9$ containing $0xfff$
22: for in parallel: $i \leftarrow 0$ to $fpos - 1$ do
23: \hspace{1cm} $ElemLst[A8[i]].Force \leftarrow A9[i]$
24: \hspace{1cm} $ElemLst[A8[i]].Moment \leftarrow A10[i]$
25: end for

\hspace{1cm} \triangleright \text{Abbreviation: El., Element}
Research highlights

► A novel GPU and CPU-based DEM algorithm is developed.

► Any general-purpose contact model can be used without restrictions in the contact searching.

► The developed algorithms are fairly compared with LIGGGHTS.

► Developed CPU-based DEM algorithm outperforms LIGGGHTS in demanding filling scenarios.

► GPU-based DEM algorithm executed on a single GPU does not enhance the CPU-based algorithm.
Credit Author Statement

**J. M. Tiscar**: Conceptualization, Methodology, Software, Investigation, Writing – Original Draft;

**A. Escrig**: Software, Validation.

**G. Mallol**: Funding acquisition, Resources.

**J. Boix**: Project administration, Supervision, Funding acquisition.

**F. A. Gilabert**: Conceptualization, Software, Formal analysis, Supervision, Funding acquisition, Writing – Review & Editing