Benefits of hybrid parallelization in reducing load imbalance in molecular dynamics simulations

Julian Morillo[a,*], Maxime Vassaux[b], Peter V. Coveney[b], Marta Garcia-Gasulla[a]

[a]Barcelona Supercomputing Center, c/Jordi Girona 29, 08034 Barcelona (Spain)
[b]Centre for Computational Sciences, University College London, 20 Gordon Street, London, WC1H 0AJ (United Kingdom)

Abstract

The most widely used technique to allow for parallel simulations in molecular dynamics is spatial decomposition, where the physical geometry is divided in boxes, one per processor. This technique can inherently produce computational load imbalance when either the spatial distribution of particles or the computational cost per particle is not uniform. This paper shows the benefits of using a hybrid MPI+OpenMP model to deal with this load imbalance. We consider the LAMMPS, a molecular dynamics simulator that provides its own balancing mechanism and an OpenMP implementation for many of its modules, allowing for a hybrid setup. In this work we extend the current OpenMP implementation of LAMMPS and optimize it and evaluate three different setups: MPI-only, MPI with the LAMMPS balance mechanism, and hybrid setup using our improved OpenMP version. This comparison is made using the five standard benchmarks included with LAMMPS distribution plus two additional test cases. Results show that the hybrid approach can deal with load balance problems better and straightforwardly than the LAMMPS balance mechanism and improve simulations with issues other than load imbalance.

Keywords: load balance, parallel computing, molecular dynamics, MPI, OpenMP, hybrid programming model

1. Introduction and Related Work

LAMMPS (Large-scale Atomic/Molecular Massively Parallel Simulator, [1, 2]) is a classical molecular dynamics code with a focus on materials modeling. It has potentials for solid-state materials (metals, semiconductors) and soft matter (biomolecules, polymers), and coarse-grained or mesoscopic systems. It can be used to model atoms or, more generically, as a parallel particle simulator at the atomic, meso, or continuum scale.

LAMMPS can be run in parallel using MPI and a spatial-decomposition of the simulation domain. The basic idea of a spatial decomposition method is to divide the physical geometry into small boxes, one per processor. Each processor will compute primarily on atoms within its box. This may induce load imbalance in problems with non uniform atom densities. The problem of load imbalance in MPI programs is well known [3] and in particular in molecular dynamics simulations is widely recognized [4, 5].

The challenge of the MPI load imbalance problem comes from the nature of MPI programming, where each process has its own data that can only be shared by explicit message passing. Simultaneously, the nature of load imbalance is dynamic and affected by many factors, therefore difficult to predict. Traditionally, solutions to load imbalance can be divided into two groups; The ones that are applied before execution; in this group, we can consider different mesh partitioners [6, 7]. These solutions are static and cannot address

*Corresponding author
Email addresses: julian.morillo@bsc.es (Julian Morillo), m.vassaux@ucl.ac.uk (Maxime Vassaux), p.v.coveney@ucl.ac.uk (Peter V. Coveney), marta.garcia@bsc.es (Marta Garcia-Gasulla)

load changes during the execution. Moreover, they need to be tuned for new architectures, algorithms, or simulations.

The approaches applied during the execution can be classified as solutions that "move" data and solutions that change computational resources. The methods that redistribute data [8, 9], usually execute a load balancing algorithm with a given frequency. This algorithm determines if there is a load imbalance problem, when necessary, computes a new partition, and finally redistributes the data as needed. These approaches are not able to deal with very dynamic load imbalance. They also need to be able to measure load and decide how frequently the load balancing algorithm is executed because the cost of redistributing the data is not negligible. Usually, these solutions are implemented within each application; LAMMPS provides its own balancing mechanism [10].

In the category of solutions applied at runtime that change the computational resources, we find different approaches. Adaptive MPI [11], for example, rely on virtualized processes, and the runtime is in charge of scheduling them to achieve a good load balance. They run on top of CHARM++ [12] which implies a change in the programming language and model. Etinski et al. [13] propose to use the Dynamic Voltage and Frequency Scaling (DVFS) reducing the frequency of less loaded processes to save power and use the turbo for more loaded ones. Also, in this category, we find the Dynamic Load Balancing library [14, 15]; this library changes the computational resources assigned to the different MPI processes to help balance their load.

We propose to use the hybrid programming model MPI+OpenMP [16, 17] to alleviate the load balance problem. This approach offers the advantages of a hybrid code: improves the load balance, and at the same time reduces the pressure on the communication between MPI processes. In contrast with other approaches that need to be programmed adhoc for each input or architecture, an OpenMP parallelization can be exploited in many situations without needing to tune the code specifically.

Deng et al. [4], for example, describes an adaptive method for achieving load balance in parallel computations that is tested on standard short-ranged parallel molecular dynamics calculations. Our proposal, in contrast, is to use a hybrid (MPI+OpenMP) approach. We argue that the use of OpenMP can help alleviate MPI scaling issues, especially the ones related to load balance, and that this can be done straightforwardly by leveraging on the OpenMP characteristics. Moreover, our evaluation is not limited to short-ranged molecular dynamics calculations: mid-range and long-range simulations are also considered, including all the benchmarks provided by the LAMMPS distribution, together with two extra testcases with quite different characteristics regarding load balance.

Many LAMMPS modules have OpenMP versions for shared-memory parallelism, allowing for hybrid setups in which MPI+OpenMP configurations can be run. Although there are OpenMP versions of many LAMMPS modules, many of them lack an OpenMP implementation. Other ones present a parallelization pattern that is not optimum for performance or programmability. This leaves MPI as the only parallel option for these parts of the code. Nonetheless, the code is designed to be easily modified or extended with new functionality. In this paper, we use such a feature to parallelize with OpenMP some code regions that lack this parallelism and improve the original OpenMP implementation in other sections of code.

The main contributions of this paper are: i) we present some addings/improvements to the LAMMPS OpenMP implementation; ii) we provide an extensive evaluation of this improved LAMMPS hybrid version against the MPI-only version as a baseline case but also against the LAMMPS balance mechanism mentioned previously; iii) we show how the hybrid approach can deal with imbalance issues better than the balance mechanism and, furthermore, it can improve performance in cases where load imbalance is not the main problem.

The document is organized as follows. Section 2 presents a description of LAMMPS and the benchmarks/testcases used for the evaluation together with the performance analyisis tools and efficiency metrics employed. In Section 3 we explain why MPI load imbalance is a problem, the difficulties to address it, and why it is very common in molecular dynamics simulations, together with the two compared approaches to solve it: the LAMMPS balancing mechanism and the use of a hybrid model. Section 4 includes a description of our proposed additions to the LAMMPS OpenMP implementation. Section 5 contains the environment employed for the evaluation together with a characterization of the benchmarks/testcases used. Finally, a complete performance comparison of the three evaluated scenarios is done for all the considered benchmark-

s/testcases. Section 6 concludes our study with comments and remarks.

2. Background

2.1. LAMMPS

The Large-scale Atomic/Molecular Massively Parallel Simulator (LAMMPS, [1, 2]) is a highly parallelised code for the simulation of classical molecular dynamics. LAMMPS is mainly and widely used by the materials science community. In classical molecular dynamics, atoms or molecules are described as particles which dynamics are controled by Newton's equations of motion. Particles interactions are determined by potentials describing pairwise, multi-body and long-range interactions. Last, the thermodynamic properties of the ensemble of particles modelled during the course of a molecular dynamics simulation are controlled by the integration of global constraints on the particles position and velocities. LAMMPS, as its name best describes, is intended for parallelism and runs as well on single processors as in parallel using message-passing interface (MPI) and a spatial-decomposition of the simulation domain. Many modules provided by LAMMPS to integrate global constraints or interatomic potentials have versions that provide accelerated performance on CPUs, GPUs, and Intel Xeon Phis. LAMMPS is distributed by Sandia National Laboratories, a US Department of Energy laboratory as an open source code under the terms of the GPL, and its design is meant to be easily extended with new functionalities, which makes it easy for external developers to contribute to the improvement of the code.

LAMMPS is used to simulate the dynamics and the properties of a wide range of systems including amorphous and crystalised materials, proteins and much more. The need of computational chemists to simulate larger systems for longer periods of time has continuously pushed the improvement of LAMMPS scalability. Besides, LAMMPS is no also frequently coupled with other tools such as machine-learning or continuum model simulators for scale-bridging purposes. As a result, large ensembles (up to thousands) of molecular dynamics simulations can be simulated simultaneously. LAMMPS has already been used to simulate the dynamics of tens of billions of atoms. On what is referred to as the "Lenard-Jones" benchmark, the highest throughput recorded was 4.34 TFlops in 2005 on a $40x10^9$ atoms simulation. In a more recent attempt LAMMPS was shown to reach $2.35x10^{-8}$ s/atom/timestep.

2.2. Benchmarks

The evaluation reported in Section 5 is performed executing a combination of LAMMPS standard benchmarking scenarios and a couple of additional scenarios triggering more specifically load-balancing issues. LAMMPS features a set of five standard benchmark representative of the diversity of systems that can be simulated. We assume that parallel efficiency is highly impacted by the range of interatomic potential interactions. We therefore can classify the five scenarios into one of the following three classes of problems: (i) short-range, (ii) mid-range and (iii) long-range interactions. Each scenario simulates the dynamics of 32,000 atoms. (i) The so-called "Granular chute" scenario simulates the convective flow of falling particles interacting via a frictional history potential. The "Polymer chain" melt scenario simulates the thermal fluctuations of hundred monomers long chains. The two scenarios are representative of short-range interacting systems, each particle interacting on average with respectively 7 and 5 neighbours. (ii) The "EAM" scenario simulates the thermodynamic fluctuations of a metallic copper bulk solid which atoms interact via the embedded atom method (EAM) potential. The "Lennard-Jonnes" scenario simulates the thermodynamics of an atomic fluid. The two scenarios are representative of a mid-range interacting system, each particle interacting on average with respectively 45 and 55 neighbours. (iii) The "Rhodopsin" scenario simulates the conformation changes of the rhodopsin protein in a solvated lipid bilayer, the CHARMM force-field is used to described atoms pairwise and multi-body interactions. The "Rhodopsin" scenario also integrates long-range Coulomb interactions, resulting in each particle interacting on average with 440 neighbours. The data required for the simulation of these benchmarks is included in the distribution of LAMMPS. Further details on the constraints applied during the simulation of the scenarios can be found on the Benchmark's page on LAMMPS website (https://lammps.sandia.gov/bench.html).

In addition to LAMMPS standard benchmark scenarios, we introduce the simulation of an epoxy resin and the simulation of a graphene-based nanocomposite. The simulation of the epoxy resin is referred to as the "Epoxy" scenario, it simulates the non-equilibirum dynamics of highly-crosslinked epoxy polymer chains under applied stretching. The epoxy resin is constrained with fixed number of atoms and temperature. Meanwhile, the volume is controlled throughout the simulation and varied at fixed strain rate. The second scenario features the simulation of graphene-oxide (GO) sheet embedded in a polymer precursor, it is referred to as the "CG-GO" scenario. The GO sheet is a dense, two-dimensional packing of carbon atoms, while polymer precursors consist in a disordered phase of poly(methyl methacrylate) (PMMA) precursors. The "CG-GO" scenario simulates the dynamics of GO during annealing, the number of atoms and the volume of the system are fixed and the temperature is increasing from 300K to 500K.

These two custom systems face specific computational efficiency issues which justified the improvement of the current load balancing methods available in LAMMPS. We will perform efficiency measurements which highlight existing bottlenecks and propose load balancing improvement based on the measurements analysis.

2.3. Performance Tools and Efficiency Metrics

In this paper we go one step further and we aim at gaining in-sight on the reasons for the performance achieved in the different situations at study. For this we rely on performance analysis tools and the performance methodology promoted by Center of Excellence (CoE) for Performance Optimization and Productivity (POP) [1].

The performance analysis tools used in this work are the following:

Extrae: To obtain traces of the different executions, it supports MPI and OpenMP among other parallel programming models[18, 19].

Paraver: To visualize the traces obtained with Extrae. It allows us to analyze in detail the behaviour of the program and also to compute the performance metrics[20, 21].

The POP CoE has defined a methodology for performance analysis. This methodology is independent of the tool being used for the analysis and defines a set of performance metrics. This metrics are well defined, accepted by the community and meaningful, pointing the analysts to the main factors affecting the performance and scalability of the code. In this paper we use some of this metrics as they allow us to compare the different LAMMPS benchmarks using a common ground.

The POP metrics are hierarchical and multiplicative, meaning that the parent metric can always be computed as the product of its childs. Each metric can get values between 0 and 100, and the metric indicates how well that indicator is performing. For example, a load balance of 70% indicates that 30% of the cpu time used is lost due to load imbalance, and also that addressing the load imbalance problem we will enable to improve the execution by at most 30%. Specifically we are going to use 3 metrics from the POP methodology: Parallel Efficiency, Load Balance and Communication efficiency.

These efficiency metrics are based on the simplification of a process into two states: the state in which it is performing computation, which is called Useful (blue), and the state in which it is not performing computation, e.g., communicating to other processes, which is called Not useful (red).

An example of this simplification can be seen in Figure 1 were we can see two processes, named p_1 and p_2 running from left to right, the time being represented in the x axis. We can see how their execution changes between the two states from Useful to Not useful and vice versa during their execution.

We call $P = \{p_1, \ldots, p_n\}$ the set of MPI processes and n the number of MPI processes. For each MPI process p we define the set $U_p = \{u_1^p, u_2^p, \ldots, u_{|U|}^p\}$ of the time intervals where the application is performing useful computation (the set of the blue intervals). We define the sum of the durations of all useful time intervals in a process p as shown in Equation 1 and we call it the useful duration of a process.

[1] https://pop-coe.eu/

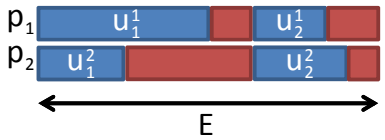


Figure 1: State evolution of two processes

$$D_{U_p} = \sum_{U_p} \blacksquare = \sum_{j=1}^{|U_p|} u_j^p \tag{1}$$

Similarly we can define $\overline{U}_p$ and $D_{\overline{U}_p}$ for the red intervals.

We also define the elapsed time E as $E = max_{p=0}^{n} D_{U_p} + D_{\overline{U}_p}$, the elapsed time is the total duration of the execution.

Parallel Efficiency (PE). The Parallel efficiency indicates the amount of time that is being lost due to the parallelization of the code. Or, equivalently, the ratio between the time used for useful computation and the total consumed CPU time. As we said the Parallel efficiency PE can be computed as the product of its childs, in this case the Load balance LB and Communication efficiency CE and is defined as shown in Equation 2.

$$PE = \frac{\sum_{U_p} \blacksquare}{E * n}; PE = LB * CE \tag{2}$$

Load Balance. Load balance measures the efficiency loss due to different loads (useful computation) for each process. Its definition can be seen in Equation 3.

$$LB = \frac{\sum_{i=1}^{n} D_{U_i}}{n * \max_{i=1}^{n} D_{U_i}} \tag{3}$$

Communication efficiency. Finally, the Communication efficiency is the efficiency loss for communicating data, it can be divided into two child metrics Serialization efficiency and Transfer efficiency. Serialization corresponds to time lost due to synchronizations between different processes, i.e., when one process needs to wait for another one. Transfer is the time lost in any kind of MPI overhead, it includes different factors such as: network bandwidth, communication latency or implementation overheads. The definition of Communication efficiency can be found in Equation 4.

$$CE = \frac{\max_{i=1}^{n} D_{U_i}}{E} \tag{4}$$

3. Challenges and Proposed Approaches to Load Imbalance

3.1. Imbalance in Molecular Dynamics Simulations

Molecular dynamics (MD) is a commonly used tool for simulation of the structural, thermodynamic, and transport properties of biological and polymeric systems on the picosecond to nanosecond timescale. During a timestep of the MD simulation, forces are computed on each atom due to its interaction with other atoms, and atoms move by integrating simple Newtonian equations of motion [5].

The parallel nature of MD simulations has long been recognized [5, 22]. The overall calculation on P processors should scale as N/P, being N the total number of atoms in the simulated system. For general molecular systems simulated on message-passing machines, most parallel implementations have used the $replicated - data$ technique [23] where a copy of all N atomic positions is stored on each of P processors. This enables easy computation and load-balancing. However at each timestep, the interprocessor

communication needed to globally update a copy of the N-vector of atom positions scales as N, independent of P. Thus replicated-data methods do not scale to large numbers of processors. An alternative known as $force-decomposition$ scales as $N/\sqrt{P}$ but is still sub-optimal [24]. For large N/P ratios, spatial-decomposition methods are clearly the best algorithmic choice. By subdividing the physical volume among processors, most computations become local and communication is minimized so that optimal N/P scaling can be achieved. Such method is the one used by LAMMPS.

The basic idea of a spatial decomposition method for MD is to divide the physical geometry into small boxes, one per processor. Each processor will compute primarily on atoms within its box. This may induce load imbalance in problems with non uniform atom densities.

3.2. Balancing

To alleviate the balancing problem, LAMMPS provides the balance command [10]. This command adjusts the size and shape of processor sub-domains within the simulation box, to attempt to balance the number of atoms or particles and thus indirectly the computational cost (load) more evenly across processors. The load balancing is "static" in the sense that this command performs the balancing once, before or between simulations. The processor sub-domains will then remain static during the subsequent run. To perform "dynamic" balancing, LAMMPS provides the fix balance command, which can adjust processor sub-domain sizes and shapes on-the-fly during a run.

Load-balancing is typically most useful if the particles in the simulation box have a spatially-varying density distribution or when the computational cost varies significantly between different particles. For example, a model of a vapor/liquid interface, or a solid with an irregular geometry containing void regions. In these cases, LAMMPS default of dividing the simulation box volume into a regular-spaced grid of 3d bricks, with one equal-volume sub-domain per processor, may assign numbers of particles per processor in a way that the computational effort varies significantly. This can lead to poor performance when the simulation is run in parallel.

The balancing can be performed with or without per-particle weighting. With no weighting, the balancing attempts to assign an equal number of particles to each processor. With weighting, the balancing attempts to assign an equal aggregate computational weight to each processor, which typically induces a different number of atoms assigned to each processor. The weight assigned to a particle is defined a priori by the user based on his knowledge of the particle, for example the expected number of neighbours and interactions.

3.3. Hybridization

It is not a trivial task to determine the optimal model (pure MPI vs MPI+OpenMP) to use for some specific application. Although pure MPI can sometimes outperform hybrid, it is not less true that lots of counterexamples do exist and results tend to vary with input data, problem size, etc. even for a given code. In order to get optimal scalability one should in any case try to implement the following strategies:

- Reduce synchronization overhead
- Reduce load imbalance
- Reduce computational overhead and memory consumption
- Minimize MPI communication

Works like [25] pinpoint cases where hybrid programming model (MPI+OpenMP) can indeed be the superior solution because of reduced communication needs and memory consumption, or improved load balance.

Hybridizing the code can help alleviate MPI scaling issues, especially the ones related to load balance as the load balance within OpenMP is addressed straightforwardly when using a dynamic schedule with work-sharing or the tasking model (i.e. generating explicit tasks that will be dynamically executed by threads when they become idle).

To perform the tests with the hybrid versions of the benchmarks we have made use of OpenMP features already provided by LAMMPS library. Version lammps-20Nov19 has been used. However, and guided by the Epoxy testcase, some modifications have been added to the OpenMP implementation that are described in the following section.

4. Implementation

The use of the performance tools described in Section 2.3 to trace and analyze the execution of the Epoxy testcase allowed us to find a source of load imbalance in void NPairHalfBinNewtonTri::build(NeighList *list) function (blue regions in Figure 2). The bottom part of Figure 2 shows a timeline with the execution of two LAMMPS iterations for 48 OpenMP threads. The upper timeline represents the execution with 48 MPI ranks (included for reference, both timelines are at the same timescale). It can be seen how the OpenMP parallelization of void NPairHalfBinNewtonTri::build(NeighList *list) alleviates the load imbalance. Note, however, that the MPI execution is almost two times faster than the OpenMP one due to the sequential parts (parts not OpenMP parallelized) in the later.

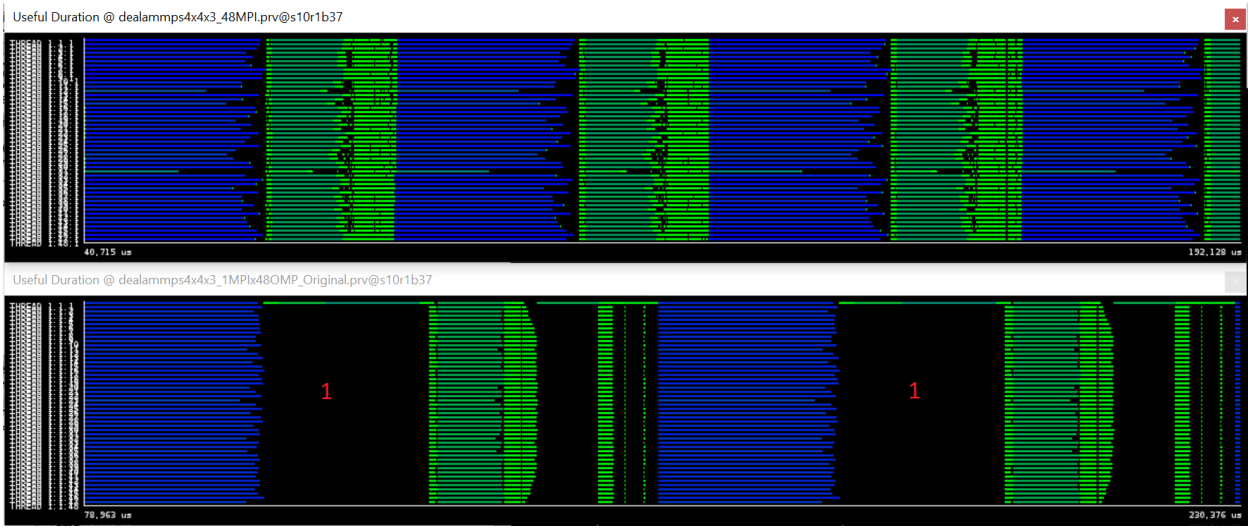


Figure 2: MPI-only (top) vs OpenMP-only (bottom) execution: the load imbalance is alleviated (blue region) but there are significant parts not OpenMP parallelized

4.1. OpenMP taskification

Given the timeline in Figure 2, our first work was to parallelize with OpenMP the biggest sequential part (marked with a red 1 in the timeline). This part corresponds to Neighbor::build_topology() function. The original LAMMPS code is depicted in Listing 1.

```
void Neighbor::build_topology()
{
  if (force->bond) {
    neigh_bond->build();
    ...
  }
  if (force->angle) {
    neigh_angle->build();
    ...
  }
  if (force->dihedral) {
    neigh_dihedral->build();
    ...
  }
  if (force->improper) {
    neigh_improper->build();
    ...
  }
}
```

Listing 1: void Neighbor::build_topology() LAMMPS original code

This function consists in 4 calls to 4 different functions build (each one from a different class). Each of these 4 functions have a very similar structure that consists in a computation phase that ends up with an MPI_Allreduce of an int calculated in this computation phase (among other things). The 4 functions work with different data structures, therefore they are independent of each other and can be run in parallel.

The objective was to annotate these 4 calls to build functions with OpenMP tasks in order to allow their execution in parallel by different threads. Besides, the issue associated with the MPI_Allreduce command remains at the end of each function. The solution consists in moving these communications out of the 4 functions and put them one level above in the Neighbor::build_topology function. The idea, then, is to have at the end 4 tasks with the computation of the 4 build functions and one final task with the 4 communications. In this way, we delay as much as possible MPI communications, preventing unnecessary waiting times if there is imbalance between MPI ranks in some of the 4 computation phases.

In order to allow the MPI communications to be executed at the end of the 4 build computations we need to move them outside of each function. To do that, a change in the signature of the functions is needed: we need them to return an int (the value shared in the MPI_Allreduce) instead of void. Then, in each of the build functions, the calculated int in the computation phase is returned by the function instead of being directly shared with other MPI ranks through the MPI_Allreduce call. These returned values are then used in the new MPI_Allreduce calls located in void Neighbor::build_topology() function.

Code Listing 2 presents a skeleton of the final implementation of Neighbor::build_topology() function. As it can be seen, 4 new local variables are declared: the variables will be used to store the values returned by each of the build functions and, in turn, to honor the dependencies between the 4 computation tasks and the communications task.

```
void Neighbor::build_topology()
{
int nmissing_bond, nmissing_angle, nmissing_dihedral, nmissing_improper, all;
#pragma omp parallel
#pragma omp single
{
#pragma omp task depend(out:nmissing_bond)
  if (force->bond) {
    nmissing_bond = neigh_bond->build();
    ...
  }
#pragma omp task depend(out:nmissing_angle)
  if (force->angle) {
    nmissing_angle = neigh_angle->build();
    ...
  }
#pragma omp task depend(out: nmissing_dihedral)
  if (force->dihedral) {
    nmissing_dihedral = neigh_dihedral->build();
    ...
  }
#pragma omp task depend(out: nmissing_improper)
  if (force->improper) {
    nmissing_improper = neigh_improper->build();
    ...
  }
#pragma omp task depend(in:nmissing_bond,nmissing_angle,nmissing_dihedral,\
nmissing_improper)
{
  MPI_Allreduce() x 4
    ...
} //end task
#pragma omp taskwait
} //end single and parallel
```

```
}
```

Listing 2: Modified void Neighbor::build_topology() code including OpenMP taskification

The present modifications significantly reduce the execution time of the biggest sequential part and work efficiently for a small number of threads. Note, however, that we are generating only 5 OpenMP tasks and only 4 of them can run in parallel as the communications one needs to wait for the execution of the others. So, when moving to the extreme case of using 48 threads (the number of cores on the target machine), more parallelism is needed. To accomplish that, the loops that make the calculations inside each of the 4 build functions have also been taskified: this allows the generation of sufficient tasks to feed all threads.

The results of these modifications can be appreciated in Figure 3 where the red lines mark explicitly the region of code affected by these changes and the reduction in execution time (upper part of the figure corresponds to the original LAMMPS OpenMP implementation and the bottom timeline corresponds to our improved OpenMP version).

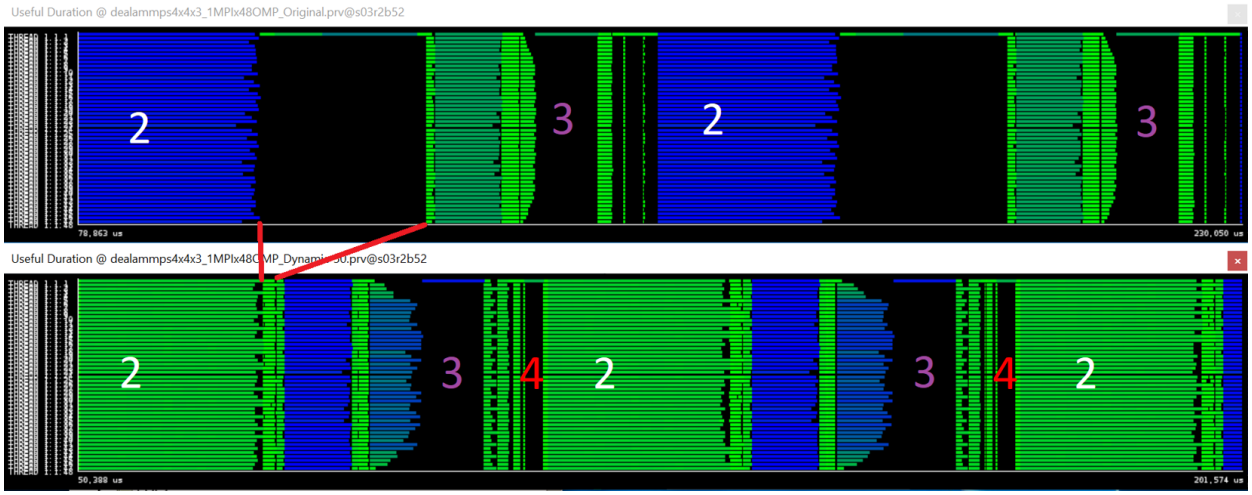


Figure 3: Improved version compared to the LAMMPS original OpenMP implementation.

4.2. Use of OpenMP dynamic scheduler

The second code modification has been done in the previous mentioned function NPairHalfBinNewtonTri::build(NeighList *list). It has been shown how by simply using the LAMMPS OpenMP implementation the detected imbalance was alleviated. A close look to the source code shows, however, that the LAMMPS OpenMP implementation does a static partition of the workload (like MPI does) so there is still some room for improvement in this part of the code. This static partition of the workload is done through the use of 3 macros defined in npair_omp.h (see Listing 3). These 3 macros are widely used along all the LAMMPS OpenMP code so the same code refactoring done in this section could be done in many other parts of the code.

```
// get access to number of threads and per-thread data structures via FixOMP
#define NPAIR_OMP_INIT                                    \
  const int nthreads = comm->nthreads;                    \
  const int ifix = modify->find_fix("package_omp")

// get thread id and then assign each thread a fixed chunk of atoms
#define NPAIR_OMP_SETUP(num)                              \
  {                                                       \
    const int tid = omp_get_thread_num();                 \
    const int idelta = 1 + num/nthreads;                  \
    const int ifrom = tid*idelta;                         \
```

```
    const int ito   = ((ifrom + idelta) > num)  \
      ? num : (ifrom+idelta);                   \
    FixOMP *fix = static_cast<FixOMP *>(modify->fix[ifix]); \
    ThrData *thr = fix->get_thr(tid);           \
    thr->timer(Timer::START);

#define NPAIR_OMP_CLOSE                         \
      thr->timer(Timer::NEIGH);                 \
    }
```

Listing 3: Macros defined in npair_omp.h

Listing 3 shows that the macro actually performing the workload partition is NPAIR_OMP_SETUP(num). It does so by dividing num among the number of available threads (i.e. in a loop of num iterations, defines the starting and end iteration that must be executed by each thread by setting ifrom and ito variables).

Once understood how these macros work, it is quite straightforward to implement the proposed approach. As it can be seen in Listing 4 the proposed change simply consists in substituting the original for that uses ifrom and ito variables by another one that, instead, starts at 0 and ends at nlocal (i.e. the value used in NPAIR_OMP_SETUP in this case). Of course, the new for is surrounded by a #pragma omp for schedule(dynamic) to do the worksharing (note that a #pragma omp parallel is not needed as it is already present at the beginning of the function in the original code, see Listing 4).

```
void NPairHalfBinNewtonTriOmp::build(NeighList *list)
{
  ...
  NPAIR_OMP_INIT;
#if defined(_OPENMP)
#pragma omp parallel default(none) shared(list)
#endif
  NPAIR_OMP_SETUP(nlocal);
  ...
  #pragma omp for schedule(dynamic,50)
  for (i = 0; i < nlocal; i++) {
//  for (i = ifrom; i < ito; i++) {
        ... (Computation)

  }
  NPAIR_OMP_CLOSE;
}
```

Listing 4: Sketch of NPairHalfBinNewtonTriOmp::build(NeighList *list) code including the dynamic OpenMP schedule

4.3. Enabling more OpenMP parallelism

Looking into the generated traces showed that there was a relatively large portion of code not OpenMP parallelized just before the execution of void NPairHalfBinNewtonTriOmp::build(NeighList *list). This is shown in Figures 3 and 4 marked with a red 4. The bottom timeline of Figure 4 represents (at the same timescale) the same part of the execution after parallelizing some functions in this region of code. The red lines going from one timeline to the other show the reduction in execution time achieved when using the new implemented parallel regions. Three different functions have been parallelized in this section of code but let us focus on the most important in terms of execution time: void NBinStandard::bin_atoms() (see Listing 5).

```
void NBinStandard::bin_atoms()
{
  ...
#if defined(_OPENMP)
#pragma omp parallel for
#endif
  for (i = 0; i < mbins; i++) binhead[i] = -1;

```

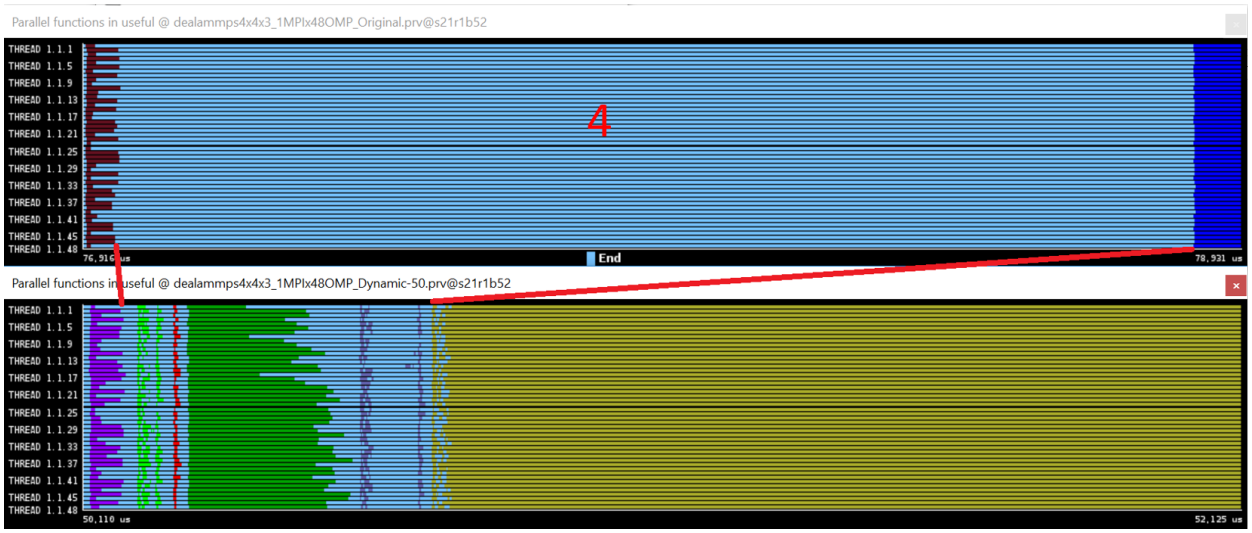


Figure 4: New implemented parallel regions.

```
  ...

  if (includegroup) {
    int bitmask = group->bitmask[includegroup];
    for (i = nall-1; i >= nlocal; i--) {
      if (mask[i] & bitmask) {
        ...
      }
    }
    for (i = atom->nfirst-1; i >= 0; i--) {
      ...
    }

  } else {
#if defined(_OPENMP)
#pragma omp parallel for private(ibin)
#endif
    for (i = nall-1; i >= 0; i--) {
      ibin = coord2bin(x[i]);
      atom2bin[i] = ibin;
      bins[i] = binhead[ibin];
      binhead[ibin] = i;
    }
  }
}
```

Listing 5: Sketch of void NBinStandard::bin_atoms()) code including the new added parallelization

Two parallel for worksharing constructs have been defined: the first one is not very relevant in terms of execution time and it corresponds to the red area in the bottom timeline in Figure 4. The important one is the located at the bottom of the Listing 5 which corresponds to the green region. As it can be seen it is a very simple parallel for that just needs to privatize ibin to work correctly.

Another important region of code candidate for enabling more OpenMP parallelism is the one marked with a purple 3 in Figure 3. This area corresponds to the execution of PPPM::poisson_ik_triclinic() method. A sketch of the original code can be seen in Listing 6.

```

void PPPM::poisson_ik_triclinic()
{
```

```
  int i,j,k,n;

  // x direction gradient

  n = 0;
  for (i = 0; i < nfft; i++) {
    work2[n] = fkx[i]*work1[n+1];
    work2[n+1] = -fkx[i]*work1[n];
    n += 2;
  }

  fft2->compute(work2,work2,-1);

  n = 0;
  for (k = nzlo_in; k <= nzhi_in; k++)
    for (j = nylo_in; j <= nyhi_in; j++)
      for (i = nxlo_in; i <= nxhi_in; i++) {
        vdx_brick[k][j][i] = work2[n];
        n += 2;
      }

  // y direction gradient

... //(same code for y direction)

  // z direction gradient

... //(same code for z direction)
}
```

Listing 6: Sketch of void PPPM::poisson_ik_triclinic()) original code

The method consists on three differentiated parts (one for each x, y and z direction) with identical structure: an initial loop, a call to fft2->compute() and, last, three nested loops. Unfortunately, variable n prevents a direct parallelization of both the initial and the three nested loops.

The solution for the first loop is to incorporate the management of variable n (initialization and increment) to the control structure of the loop (see Listing 7). Once this is done, a simple pragma omp parallel for will do the work.

```

void PPPM::poisson_ik_triclinic()
{
  int i,j,k,n;

  // x direction gradient
  #pragma omp parallel for
  for (i = 0, n = 0; i < nfft; i++, n = n+2) {
    work2[n] = fkx[i]*work1[n+1];
    work2[n+1] = -fkx[i]*work1[n];
  }

  fft2->compute(work2,work2,-1);

  int BS = (nyhi_in - nylo_in + 1) * (nxhi_in - nxlo_in + 1) * 2;
  #pragma omp parallel
  {
  #pragma omp for private(n,j,i) nowait
  for (k = nzlo_in; k <= nzhi_in; k++) {
    n = (k - nzlo_in) * BS;
    for (j = nylo_in; j <= nyhi_in; j++) {
      for (i = nxlo_in; i <= nxhi_in; i++) {
        vdx_brick[k][j][i] = work2[n];
        n += 2;
      } //i
```

```
    } //j
  } //k

  // y direction gradient
  #pragma omp for
  for (i = 0, n = 0; i < nfft; i++, n=n+2) {
    work2[n] = fky[i]*work1[n+1];
    work2[n+1] = -fky[i]*work1[n];
  }
  } //parallel

  ... //(rest of code omitted)

}
```

Listing 7: Sketch of parallelized version void PPPM::poisson_ik_triclinic()) method

The solution for the second case (the three nested loops) is trickier: we need to privatize variable n and to do that we need to manually calculate the initial value for n at each iteration of the outer-most loop. This is easilly done through the use of the added variable BS that stores the increments of the variable n in the two inner-most loops. Once all of these is done (see Listing 7), the outer-most loop can be parallelized by simply privatizing n, j and i.

As a last comment, the parallel region opened for the nested loops of x direction is used for the first loop of y direction as depicted in Listing 7. The same is done between y direction and z direction (not shown in the Listing).

4.4. Overlapping computation and communication

Last, but not least, we show here how to effectively overlap computation with MPI communication. More precisely, we have worked in remap_3d function, which is called several times in the same region of code mentioned on the last part of previous subsection. As it can be seen in Listing 8, the function consists in 4 differentiated parts:

1. A sequence of MPI_Irecv calls to receive data from other processes.
2. A sequence of (pack, MPI_Send) calls that packs and sends data to other processes.
3. A call to pack and unpack to manage the data of the calling process.
4. A sequence of (MPI_Waitany, unpack) to wait for the corresponding MPI_Irecv to get the data and put it in the required memory location.

```

void remap_3d(FFT_SCALAR *in, FFT_SCALAR *out, FFT_SCALAR *buf,
              struct remap_plan_3d *plan)
{
... // (omitted code)

    // post all recvs into scratch space
    for (irecv = 0; irecv < plan->nrecv; irecv++) {
      MPI_Irecv(&scratch[plan->recv_bufloc[irecv]],plan->recv_size[irecv],
                MPI_FFT_SCALAR,plan->recv_proc[irecv],0,
                plan->comm,&plan->request[irecv]);
    }

    // send all messages to other procs
    for (isend = 0; isend < plan->nsend; isend++) {
      plan->pack(&in[plan->send_offset[isend]],
                 plan->sendbuf,&plan->packplan[isend]);
      MPI_Send(plan->sendbuf,plan->send_size[isend],MPI_FFT_SCALAR,
               plan->send_proc[isend],0,plan->comm);
    }

```

```
    // copy in -> scratch -> out for self data
    if (plan->self) {
      isend = plan->nsend;
      irecv = plan->nrecv;
      plan->pack(&in[plan->send_offset[isend]],
                 &scratch[plan->recv_bufloc[irecv]],
                 &plan->packplan[isend]);
      plan->unpack(&scratch[plan->recv_bufloc[irecv]],
                   &out[plan->recv_offset[irecv]],&plan->unpackplan[irecv]);
    }

    // unpack all messages from scratch -> out
    for (i = 0; i < plan->nrecv; i++) {
      MPI_Waitany(plan->nrecv,plan->request,&irecv,MPI_STATUS_IGNORE);
      plan->unpack(&scratch[plan->recv_bufloc[irecv]],
                   &out[plan->recv_offset[irecv]],&plan->unpackplan[irecv]);
    }

... // (omitted code)
}
```

Listing 8: Sketch of the original code in remap_3d() method

The changes done in the code to allow computation and communication overlapping can be seen in Listing 9 and are summarized in the following items:

1. All code is wrapped by a parallel and a single constructs to create the parallel OpenMP region and allow only one thread to enter the code to create tasks.
2. The loop corresponding to point number 2 of the original code has been split into two loops: one loop doing all the packs and the other doing all the MPI_Send. The loop doing the packs has been moved to the very beginning of the function and each pack has been defined as a task.
3. To allow the previous taskification, plan->sendbuf has been redefined (not shown) and now is a buffer of buffers indexed by isend: this allows for all pack tasks to be independent.
4. As it is independent of the rest of the communications, the self-data management of point 3 of the original code has been moved next and taskified.
5. A taskwait is needed just after the loop with MPI_Irecv because the following MPI_Send needs the tasks with packs defined in point 2 to be finished.
6. Finally, the unpacks of the last loop have been defined as tasks: in this way, the following MPI_Waitany does not need to wait for the previous unpack to finish.

```

void remap_3d(FFT_SCALAR *in, FFT_SCALAR *out, FFT_SCALAR *buf,
              struct remap_plan_3d *plan)
{
... // (omitted code)

#pragma omp parallel
#pragma omp single
{
    for (isend = 0; isend < plan->nsend; isend++) {
      #pragma omp task firstprivate(isend)
      plan->pack(&in[plan->send_offset[isend]],
                 &plan->sendbuf[isend*plan->sendbuf_size],&plan->packplan[isend]);
    }

    // copy in -> scratch -> out for self data
    if (plan->self) {
      isend = plan->nsend;
      irecv = plan->nrecv;
      #pragma omp task firstprivate(isend,irecv)
      {
```

```
      plan->pack(&in[plan->send_offset[isend]],
                 &scratch[plan->recv_bufloc[irecv]],
                 &plan->packplan[isend]);
      plan->unpack(&scratch[plan->recv_bufloc[irecv]],
                   &out[plan->recv_offset[irecv]],&plan->unpackplan[irecv]);
    }
  }

  // post all recvs into scratch space
  for (irecv = 0; irecv < plan->nrecv; irecv++) {
    MPI_Irecv(&scratch[plan->recv_bufloc[irecv]],plan->recv_size[irecv],
              MPI_FFT_SCALAR,plan->recv_proc[irecv],0,
              plan->comm,&plan->request[irecv]);
  }
#pragma omp taskwait
  // send all messages to other procs
  for (isend = 0; isend < plan->nsend; isend++) {
    MPI_Send(&plan->sendbuf[isend*plan->sendbuf_size],plan->send_size[isend],MPI_FFT_SCALAR,
             plan->send_proc[isend],0,plan->comm);
  }

  // unpack all messages from scratch -> out
  for (i = 0; i < plan->nrecv; i++) {
    MPI_Waitany(plan->nrecv,plan->request,&irecv,MPI_STATUS_IGNORE);
    #pragma omp task firstprivate(irecv)
    plan->unpack(&scratch[plan->recv_bufloc[irecv]],
                 &out[plan->recv_offset[irecv]],&plan->unpackplan[irecv]);
  }
} //parallel and single

... // (omitted code)
}
```

Listing 9: Sketch of the modified code in remap_3d() method, including the code reordering and the taskification of packs and unpacks

Figure 5 shows how computation and communication have been effectively overlapped. The timelines correspond to a trace of a run with 8 MPI processes with 6 OpenMP threads each. The upper timeline represents the MPI calls (being pink → MPI_Irecv, blue→ MPI_Send, and green → MPI_Waitany) and the bottom timeline represents task execution. Figure 5 is a zoom of just one invocation of remap_3d method for the first 2 processes (12 threads in total) used in the execution. In this case, isend = irecv = 3 so each process executes 3 pack tasks + the pack/unpack task corresponding to the self data (the tasks on the left part), and 3 unpack tasks (on the right). It can be seen how the packs are now overlapped with the MPI_Irecv calls at the beginning and how the MPI_Waitanys at the end do not need to wait for the execution of the previous unpack.

5. Evaluation

5.1. Environment

The experiments have been performed on MareNostrum4 [26]. This supercomputer is based on Intel Xeon Platinum processors from the Skylake generation. It is a Lenovo system composed of SD530 Compute Racks, an Intel Omni-Path high performance network interconnect and running SuSE Linux Enterprise Server as operating system. Compute nodes are equipped with:

- 2 sockets Intel Xeon Platinum 8160 CPU with 24 cores each @ 2.10GHz for a total of 48 cores per node.
- L1d 32K; L1i 32K; L2 cache 1024K; L3 cache 33729K.
- 96 GB of main memory 1.88 GB/core.

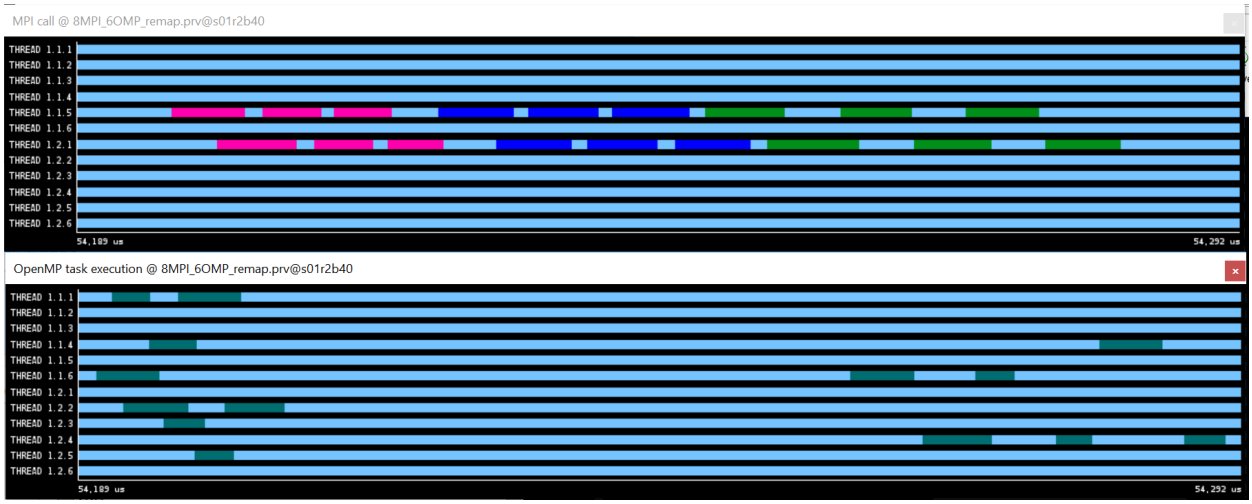


Figure 5: Overlapping computation and communication in remap_3d method.

- 100 Gbit/s Intel Omni-Path HFI Silicon 100 series PCI-E adapter.
- 10 Gbit Ethernet.
- 200 GB local SSD available as temporal storage during jobs.
- The processors support well-known vectorization instructions such as SSE, AVX up to AVX-512.

The software environment used is as follows:

- LAMMPS 20Nov19
- Intel 17.0.4 20170411 compiler

5.2. Benchmark Characterization

Figure 6 shows efficiency metrics for all the testcases and benchmarks studied in this work. These performance metrics correspond to MPI-only executions and using 48 MPI ranks in all cases. Most of them present a poor parallel efficiency, considering that only 48 MPI ranks are considered. Particularly bad are the cases of short-range interactions benchmarks and the CG-GO testcase. The reasons for poor parallel efficiency are diverse. While the main problem of short-range interactions benchmarks is Communication Efficiency, the most limiting factor of the CG-GO benchmark is Load Balance with an extremely low value (in contrast with the rest of the benchmarks). Mid-range interactions benchmarks have very similar characteristics: although with slightly different weights on the two components, they both have the same Parallel Efficiency value. Finally, Rhodopsin is the best performing benchmark. Note that, the Epoxy testcase presents quite different characteristics when compared with CG-GO as discused later.

So with all these benchmarks we cover very different scenarios both in terms of type of simulation and in terms of performance metrics characteristics.

5.3. Execution Time

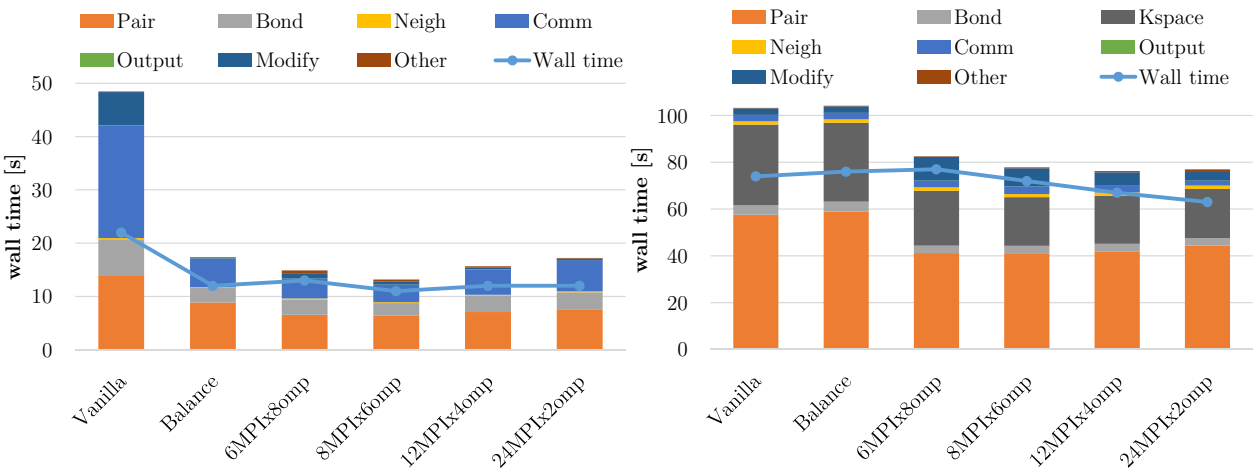
In this section we present the wall time execution of all testcases and benchmarks for different setups including the so-called ”Vanilla” (i.e. the regular MPI-only execution), ”Balance” (i.e. MPI-only execution but including the balancing mechanisms provided by the LAMMPS implementation) and different hybrid configurations. 48 cores are used in all cases and, for the hybrid configurations, only configurations up to the point where using more OpenMP threads translates in worst performance than the MPI-only case are shown. In all benchmarks, the number of timesteps has been increased to have an execution wall time of at least 1 minute for the Vanilla case in order to better appreciate execution time differences (i.e. the wall times reported are the ones provided by LAMMPS and they only have a precission of seconds).

	Testcases:		Short-range:		Mid-range:		Long-range:
	Epoxy	CG-GO	Granular chute	Polymer chain	EAM	Lennard-Jones	Rhodopsin
Parallel efficiency	0,69	0,40	0,38	0,38	0,58	0,58	0,64
|-- Load balance	0,85	0,42	0,85	0,80	0,81	0,95	0,94
|-- Communication efficiency	0,81	0,95	0,45	0,48	0,72	0,62	0,69

Figure 6: Efficiency metrics of all testcases and benchmarks using 48 MPI ranks.

5.3.1. Testcases

Let us start by the CG-GO testcase as it is the one that shows the greatest load imbalance. Figure 7 (left part) presents the execution times for different setups of this benchmark. The bars represent the maximum time spent by an MPI rank on a given code section as reported by LAMMPS in its performance execution report while the blue line that traverses the figure represents the total wall time of the different executions. So the difference between both heights gives an idea of the load imbalance that affects a given configuration. As it can be seen, this is huge for the Vanilla case. The Balance version reduces this difference a lot (mainly by reducing the "Comm" maximum execution time). Note, however how a wide range of hybrid configurations (from 24 to 8 MPI processes) do this better achieving also lower (i.e. better) wall execution times.



(a) CG-GO performance results

(b) Epoxy performance results

Figure 7: Performance results for test cases inputs

Figure 8 represent 10 timesteps of the executions using the Vanilla, Balance and Hybrid versions at the same timescale. As it can be seen, the Vanilla version presents a heavy load imbalance: this allows the Balance version to achieve a 43% improvement in execution time (timeline in the middle). But, more interestingly, the Hybrid version (bottom timeline) is 12% faster than the Balance version and it achieves a 50% improvement when compared with the Vanilla version.

Figure 9 shows the parallel functions executed by the Hybrid version. The time spent in OpenMP parallel regions in this case is 80,6% of the total execution time.

A very different scenario is shown in the right plot of the figure, that presents the execution time of the Epoxy resin testcase. Note the different characteristics of this testcase when compared with the CG-GO. First, the imbalance is not so relevant for the Vanilla scenario. In fact, it can be seen how the use of the balance mechanisms provided by LAMMPS actually make the execution slower (i.e. it adds overhead without any improvement). Hybrid configurations from 24 to 8 MPI ranks perform better in terms of both load balance and execution time, mainly due to the better performance of the pairing ("Pair", orange in the figure) of the hybrid cases, meaning that the OpenMP parallelization is more efficient than the MPI one.

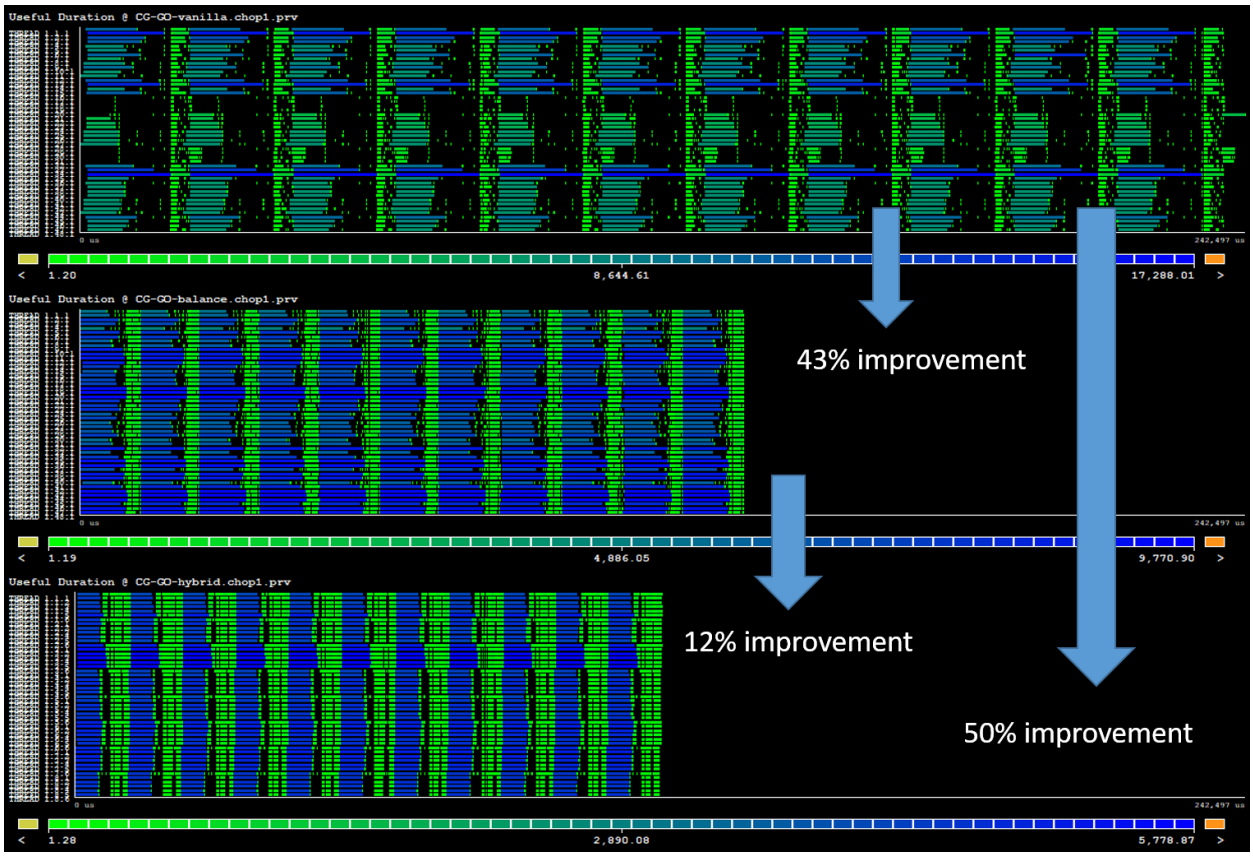


Figure 8: Useful duration timelines for the CG-GO testcase (upper: Vanilla, middle: Balance, lower:Hybrid).

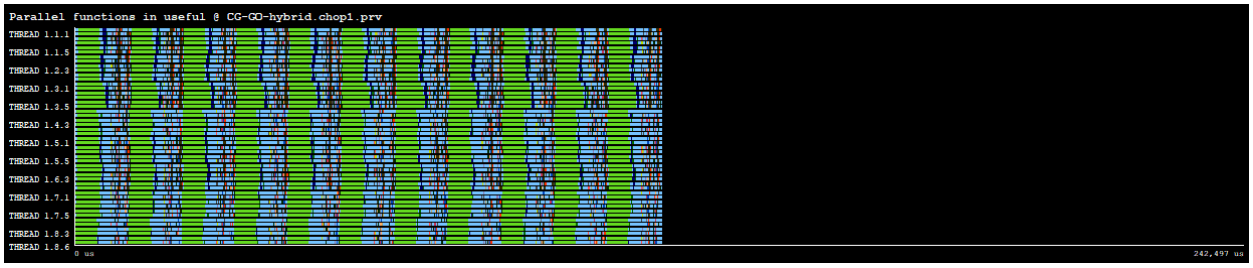
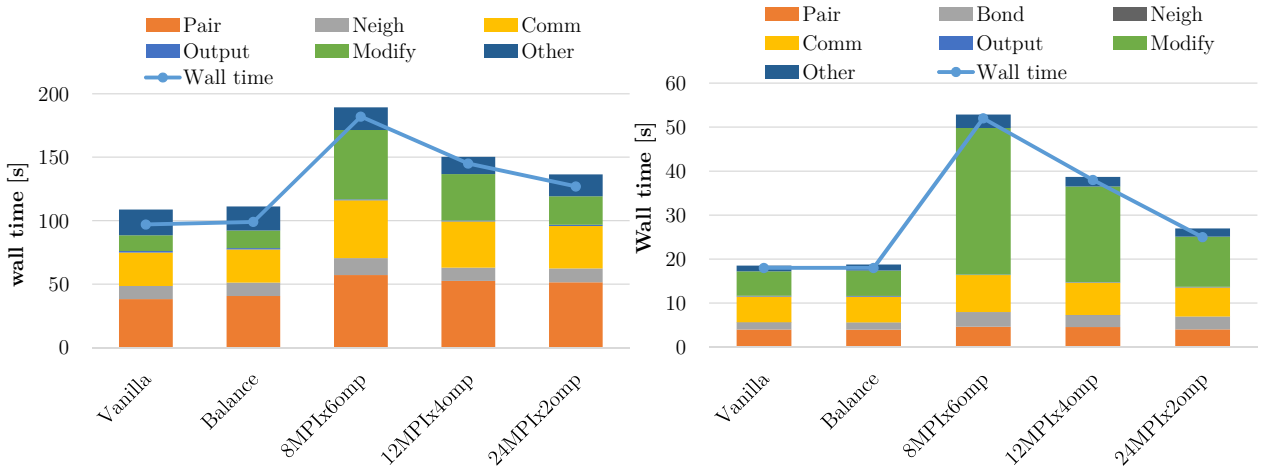


Figure 9: Parallel functions timeline for the CG-GO testcase (Hybrid version).

This testcase clearly shows how the Hybrid version is able to improve the Vanilla even when the Balance version is not, demonstrating that hybridization provides other benefits than just load balance.

5.3.2. Short-range interactions benchmarks

Figure 10 present the results for the short-range interactions benchmarks. For both benchmarks the analysis is actually the same: nor the Balance mechanism nor the Hybrid solution are able to improve the Vanilla setup.



(a) Granular chute flow performance results

(b) Polymer chain melt performance results

Figure 10: Short-range interactions benchmarks results

For the Balance mechanism, the explanation is clear: the problem of the Vanilla setup, if any, is not load imbalance. For the Hybrid configuration, we will take the Polymer chain benchmark as a representative but a similar analysis could be done for the Granular chute. Figure 11 represents 10 timesteps of the Polymer chain benchmark. The timelines, of course at the same timescale, show two main reasons that explain why this short-range interactions cases do not benefit from the use of a Hybrid implementation:

1. The most computational intensive part (dark blue) is not OpenMP parallelized so the execution time is increased.
2. Only very small parts of the less computational intensive part (light green) are parallelized with OpenMP, leading also to an increased execution time. This can be perfectly seen in Figure 12 where the OpenMP parallel regions are depicted (meaning light blue no parallel region at all, i.e. sequential execution). Actually, the percentage of time of the whole execution spent in OpenMP parallel regions is only of 23%. This suggests that there is a lot of room for improvement by parallelizing other parts of the code used by this benchmark in similar ways as explained in Section 3.3.

5.3.3. Mid-range interactions benchmarks

Figure 13 (left) presents the execution times of the EAM benchmark for different configurations. In this case, all the versions perform quite similar. It is noticeable, however that the best performing version is Hybrid for the 24MPIx2omp case: 70 seconds in contrast with the 76 seconds of the Vanilla or the 75 seconds of the Balance version. A similar analysis can be done for the Lennar-Jones benchmark (Figure 13 (right): the only noticeable difference is that in this case Balance is a bit worst than Vanilla (just one second) while the 24MPIx2omp Hybrid configuration is still able to improve by 4 seconds the Vanilla case.

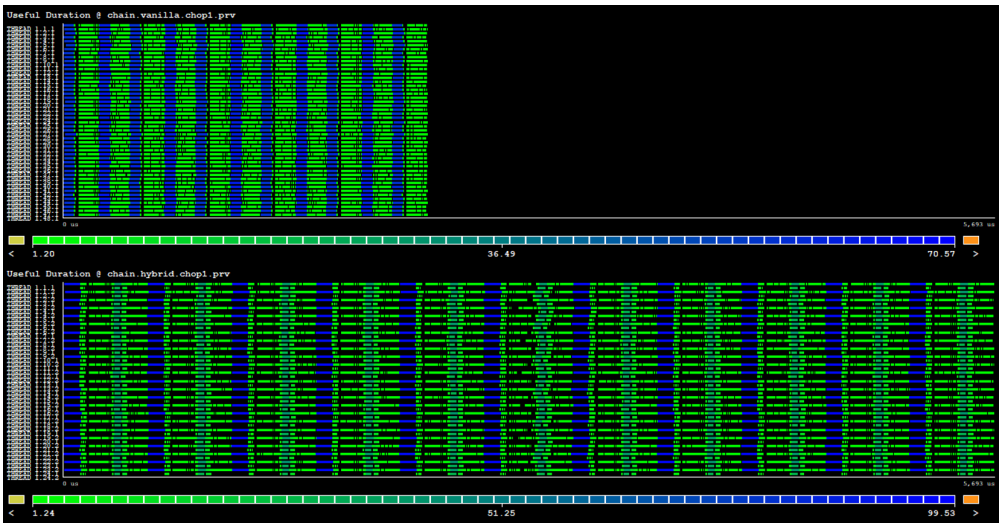


Figure 11: Useful duration timelines for the Polymer chain melt benchmark (Upper part: Vanilla execution, lower part: Hybrid version).

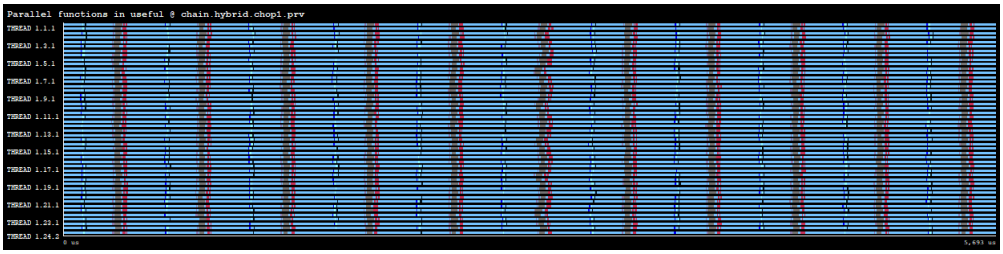
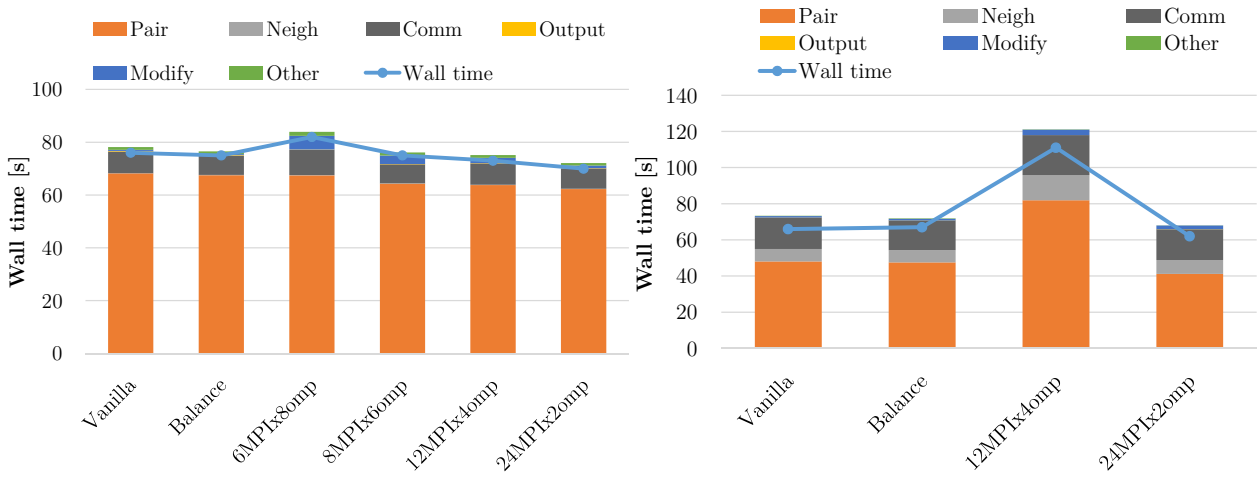


Figure 12: Parallel functions timeline for the Polymer chain melt benchmark (Hybrid version).



(a) EAM metallic solid performance results

(b) Lennard-Jones liquid performance results

Figure 13: Mid-range interactions benchmarks results

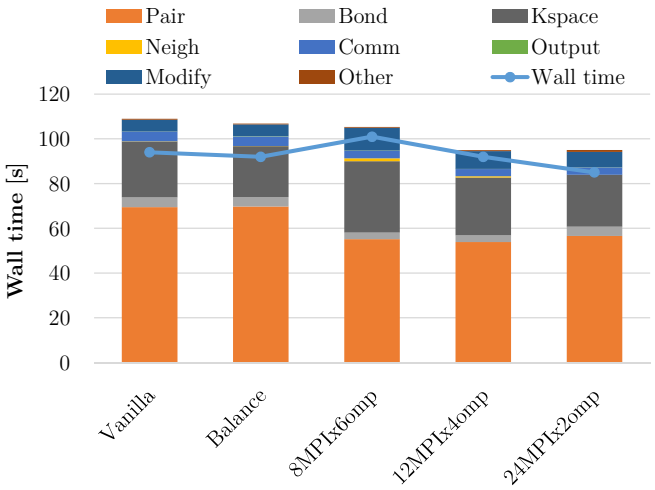


Figure 14: Rhodopsin protein benchmark results.

5.3.4. Long-range interactions benchmarks

Figure 14 presents the execution time of the Rhodopsin benchmark for different configurations. The three Hybrid configurations on the right are able to outperform both Vanilla and Balance versions. The pairing process (orange bar) is much faster in the Hybrid configurations.

Figure 15 represents two timelines of 10 timesteps of the Rhodopsin benchmark execution at the same timescale. As it can be visually noted, the execution of the Hybrid case is significantly faster. This gain in performance comes mainly from the pairing phase (blue sections in the timelines) done in the compute function in the PairLJCharmmCoulLongOMP module of LAMMPS which is faster for the Hybrid case.

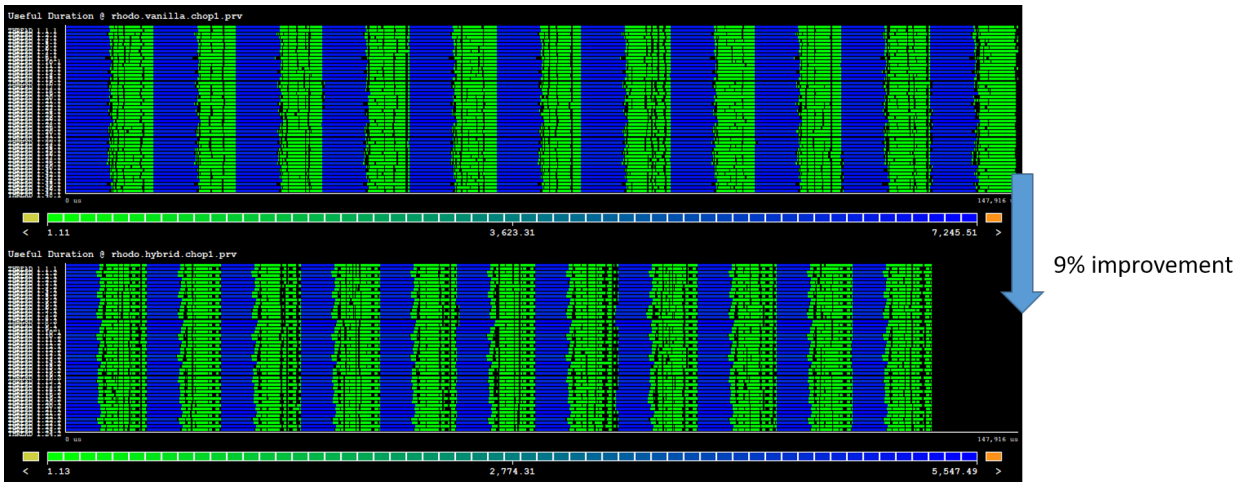


Figure 15: Useful duration timelines for the Rhodopsin protein benchmark (top: Vanilla, bottom: Hybrid).

As it can be seen on the red parts of Figure 16, this function is fully OpenMP parallelized. Note, however, that there are still other parts of the code not parallelized with OpenMP (light blue in Figure 16 and black in bottom timeline of Figure 15), making the execution of the green areas in Figure 15 being slightly faster for the Vanilla case. But even so, the improvement achieved by the OpenMP implementation of compute subroutine is able to compensate by far this loss in performance.

All in all, the ratio of time in parallel regions with respect to the whole execution time is pretty high: 76,6%. We can now compare short-range with long-range interactions benchmarks and explain why Hybrid

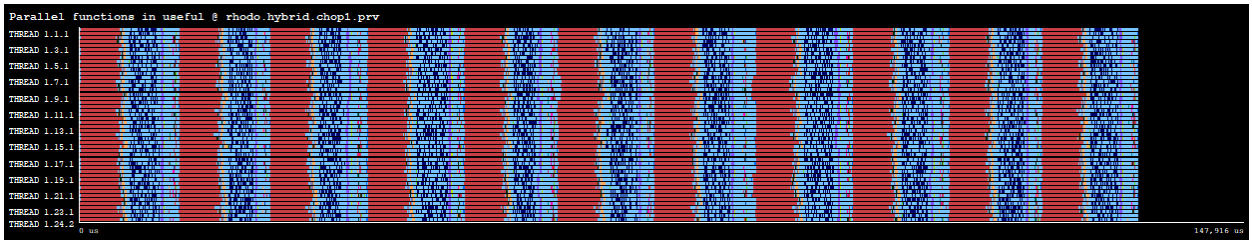


Figure 16: Parallel functions timeline for the Rhodopsin protein benchmark (Hybrid version).

implementation is able to improve the performance of the later but not of the formers. Comparing Figures 12 and 16 one can see how the percentage of time outside any parallel OpenMP region (light blue areas) for both benchmarks is drastically different. Actually, the percentage of time inside OpenMP parallel regions of the Polymer chain benchmark is only 23,7%. This is exacerbated if we focus on the pairing phases: while in the case of Rhodopsin benchmark this corresponds to the red parts in Figure 16, so it represents a high percentage of the whole execution, this is not the case for the Polymer chain benchmark where the pairing (light brown in Figure 12) corresponds to a very residual part of the whole execution (in this benchmark, subroutine compute of PairLJCutOMP module).

So, at the end, the reason that ultimately explains the different behaviour of the two kind of benchmarks is that the weight of the pairing process (which is the most important part parallelized with OpenMP in all benchmarks) in the whole execution is very high for the long-range interactions benchmarks (45,6% in the Rhodopsin benchmark) while it is insignificant for the short-range interactions benchmarks (8,4% for the Polymer chain benchmark).

6. Conclusions

This paper has shown the potential of the proposed hybrid MPI+OpenMP approach to effectively alleviate performance problems such as load imbalance in LAMMPS simulations.

LAMMPS provides a ready-to-use balacing mechanism to partially solve load balancing problems in molecular dynamics simulations with non uniform atom densities. The LAMMPS balancing mechanism shows high effiency compared to MPI-only simulation in case of high load imbalance (CG-GO testcase). We have introduced the use of an MPI+OpenMP hybrid implementation of LAMMPS as a third option. Furthermore, we have complemented the current partial OpenMP implementation of LAMMPS with additions and modifications, driven by the Epoxy textcase.

Our proposed modified version LAMMPS has been extensively compared against the baseline MPI case and against the use of the LAMMPS balance mechanism. Five benchmarks present in the LAMMPS distribution with varying range of interaction (short, mid and long-range) together with two testcases with very different characteristics were used for the comparison.

For the short-range interactions benchmarks, the regular MPI-only version was the best performing. As long as they do not present a load imbalance problem, the balancing mechanism does not provide any benefit in this case. The problem with the hybrid setup in this case is that only a small fraction of the simulations (~20%) runs in OpenMP parallel regions. This suggests that more additions similar to the ones proposed in Section 3.3 could be done to the OpenMP LAMMPS implementation.

For the mid-range interactions benchmarks, the hybrid option was the best in all cases. The balance mechanism only improves a bit the EAM simulation while it is the worst option for the Lennard-Jones benchmark. These results indicate that the hybrid implementation is able to improve performance metrics others than load balance, such as communication efficiency.

In the Rhodopsin benchmark (long-range interactions) the execution time is mainly dominated by the LAMMPS pairing process. The obtained results show that the OpenMP paralllelization of the pairing is much faster than the MPI one, making the hybrid approach the best option also for this benchmark.

Regarding the highly-imbalanced testcase (CG-GO), the balancing mechanism shows its potential achieving a 43% improvement with respect of the regular MPI simulation. Interestingly, the hybrid version is able to improve even further, up to 50%.

In the case of the Epoxy resin testcase (the one that motivated the implementations explained in Section 3.3), the use of the balance mechanism only adds overhead (the execution is slower than for the regular MPI version). The hybrid implementation, on the contrary, is the best option showing again that it is able to improve simulations where the load balance is not the main problem.

So, the overall conclusion is that LAMMPS hybrid setups are able to handle scenarios with very high load imbalance at least as well (if not better) as the LAMMPS balance mechanism while also providing benefits in other scenarios where load balance is not the main performance bottleneck.

Our suggestion to LAMMPS developers, and MD in general, is to put effort into hybridizing the code with an MPI+OpenMP strategy instead of implementing ad-hoc balancing methods. Because the hybrid code not only can address more dynamic load imbalances but also improve the parallel efficiency reducing the communication.

7. Acknowledgements

This work is partially supported by the Spanish Government through Programa Severo Ochoa (SEV-2015-0493), by the Spanish Ministry of Science and Technology (TIN2015-65316-P), by the Generalitat de Catalunya (2017-SGR-1414), and by the European POP CoE (GA n. 824080). This work is also funded as part of the European Union Horizon 2020 research and innovation programme under grant agreement nos. 800925 (VECMA project; www.vecma.eu) and 823712 (CompBioMed2 Centre of Excellence; www.compbiomed.eu), as well as the UK EPSRC for the UK High-End Computing Consortium (grant no. EP/R029598/1).

References

[1] LAMMPS. [link].
URL https://lammps.sandia.gov/

[2] S. Plimpton, Fast parallel algorithms for short-range molecular dynamics, Journal of computational physics 117 (1) (1995) 1–19.

[3] K. D. Devine, E. G. Boman, R. T. Heaphy, B. A. Hendrickson, J. D. Teresco, J. Faik, J. E. Flaherty, L. G. Gervasio, New challenges in dynamic load balancing, Applied Numerical Mathematics 52 (2-3) (2005) 133–152.

[4] Y. Deng, R. F. Peierls, C. Rivera, An adaptive load balancing method for parallel molecular dynamics simulations, Journal of Computational Physics 161 (1) (2000) 250 – 263. doi:https://doi.org/10.1006/jcph.2000.6501.
URL http://www.sciencedirect.com/science/article/pii/S002199910096501X

[5] S. Plimpton, R. Pollock, M. Stevens, Particle-mesh ewald and rrespa for parallel molecular dynamics simulations, Proc. 8th SIAM Conf. on Parallel Processing for Scientific Computing (08 2000).

[6] C. Walshaw, M. Cross, Mesh partitioning: a multilevel balancing and refinement algorithm, SIAM Journal on Scientific Computing 22 (1) (2000) 63–80.

[7] G. Karypis, V. Kumar, A fast and high quality multilevel scheme for partitioning irregular graphs, SIAM Journal on scientific Computing 20 (1) (1998) 359–392.

[8] D. F. Harlacher, H. Klimach, S. Roller, C. Siebert, F. Wolf, Dynamic load balancing for unstructured meshes on space-filling curves, in: 2012 IEEE 26th International Parallel and Distributed Processing Symposium Workshops & PhD Forum, IEEE, 2012, pp. 1661–1669.

[9] K. Schloegel, G. Karypis, V. Kumar, A unified algorithm for load-balancing adaptive scientific simulations, in: SC'00: Proceedings of the 2000 ACM/IEEE Conference on Supercomputing, IEEE, 2000, pp. 59–59.

[10] LAMMPS-Balance. [link].
URL https://lammps.sandia.gov/doc/balance.html

[11] C. Huang, O. Lawlor, L. V. Kale, Adaptive mpi, in: International workshop on languages and compilers for parallel computing, Springer, 2003, pp. 306–322.

[12] B. Acun, A. Gupta, N. Jain, A. Langer, H. Menon, E. Mikida, X. Ni, M. Robson, Y. Sun, E. Totoni, et al., Parallel programming with migratable objects: Charm++ in practice, in: SC'14: Proceedings of the International Conference for High Performance Computing, Networking, Storage and Analysis, IEEE, 2014, pp. 647–658.

[13] M. Etinski, J. Corbalan, J. Labarta, M. Valero, A. Veidenbaum, Power-aware load balancing of large scale mpi applications, in: 2009 IEEE International Symposium on Parallel & Distributed Processing, IEEE, 2009, pp. 1–8.

[14] M. Garcia, J. Corbalan, J. Labarta, LeWI: A Runtime Balancing Algorithm for Nested Parallelism, in: Proceedings of the International Conference on Parallel Processing (ICPP09), 2009.

[15] M. Garcia-Gasulla, F. Mantovani, M. Josep-Fabrego, B. Eguzkitza, G. Houzeaux, Runtime mechanisms to survive new hpc architectures: a use case in human respiratory simulations, The International Journal of High Performance Computing Applications 34 (1) (2020) 42–56.
[16] R. Rabenseifner, G. Hager, G. Jost, Hybrid mpi/openmp parallel programming on clusters of multi-core smp nodes, in: 2009 17th Euromicro international conference on parallel, distributed and network-based processing, IEEE, 2009, pp. 427–436.
[17] R. Rabenseifner, G. Wellein, Communication and optimization aspects of parallel programming models on hybrid architectures, The International Journal of High Performance Computing Applications 17 (1) (2003) 49–62.
[18] Barcelona Supercomputing Center, Paraver.
URL https://tools.bsc.es/paraver
[19] V. Pillet, J. Labarta, T. Cortes, S. Girona, Paraver: A tool to visualize and analyze parallel code, in: Proceedings of WoTUG-18: transputer and occam developments, Vol. 44, 1995, pp. 17–31.
[20] Barcelona Supercomputing Center, Extrae.
URL https://tools.bsc.es/extrae
[21] H. Servat, et al., Framework for a productive performance optimization, Parallel Computing 39 (8) (2013) 336–353.
[22] D. Fincham, Parallel computers and molecular simulation, Molecular Simulation 1 (1-2) (1987) 1–45. arXiv:https://doi.org/10.1080/08927028708080929, doi:10.1080/08927028708080929.
URL https://doi.org/10.1080/08927028708080929
[23] W. Smith, Molecular dynamics on hypercube parallel computers, Computer Physics Communications 62 (2) (1991) 229 – 248. doi:https://doi.org/10.1016/0010-4655(91)90097-5.
URL http://www.sciencedirect.com/science/article/pii/0010465591900975
[24] S. Plimpton, B. Hendrickson, A new parallel method for molecular dynamics simulation of macromolecular systems, Journal of Computational Chemistry 17 (3) (1996) 326–337. doi:https://doi.org/10.1002/(SICI)1096-987X(199602)17:3<326::AID-JCC7>3.0.CO;2-X.
[25] R. Rabenseifner, G. Hager, G. Jost, Hybrid mpi/openmp parallel programming on clusters of multi-core smp nodes, in: 2009 17th Euromicro International Conference on Parallel, Distributed and Network-based Processing, 2009, pp. 427–436. doi:10.1109/PDP.2009.43.
[26] Barcelona Supercomputing Center, Marenostrum4.
URL https://www.bsc.es/marenostrum/marenostrum