Student Cluster Competition 2017, Team Tsinghua University: Reproducing Vectorization of the Tersoff Multi-Body Potential on the Intel Skylake and NVIDIA Volta Architectures

Ka Cheong Jason Lau*, Yuxuan Li, Lei Xie, Qian Xie, Beichen Li,
Yu Chen, Guanyu Feng, Jiping Yu, Miao Wang, Wentao Han and Jidong Zhai*

Department of Computer Science and Technology
Tsinghua University, Beijing, China

* i@laukc.com, zhaijidong@tsinghua.edu.cn

ABSTRACT

A paper of SC '16 entitled "*The Vectorization of the Tersoff Multi-Body Potential: An Exercise in Performance Portability*"[4] implemented reduced precision calculation and cross-platform vectorization for Tersoff potential, which the authors claimed as accurate, efficient, scalable and portable. In this report, we focus on recently released computing architectures, Intel Skylake and NVIDIA Volta, to present our results and compare them with that of the Tersoff paper. With new input provided in Porter's paper [6] on our cluster, among the given testcases, our results demonstrate that we fail to reproduce the original publication's claims about performance improvements and scalability. Deeper analysis demonstrate it is the communication bottleneck caused by special characteristics of the new input data that limit the reproducibility.

Keywords

Reproducibility; Vectorization; Molecular dynamics simulation; Tersoff potential; Student cluster competition

1. INTRODUCTION

LAMMPS is a parallel molecular dynamics simulator written in C++ [5], supporting not only MPI-based parallelization, but also OpenMP, USER-INTEL, KOKKOS [3] and GPUs [2] modes. With the improvement of computational power, people expect to make simulations more accurately on higher performance platforms. Therefore, though costly, multi-body potential, in contrast to pair potential, has gained its popularity as well as multi-body potential computation optimization. However, the complexity of multi-body potential kernel limits the range of architectures available with native vectorization porting [4].

Höhnerbach, Ismail and Bientinesi (2016) attempt to overcome these difficulties by implementing a general kernel approach in their SC '16 paper, *The Vectorization of the Tersoff Multi-Body Potential: An Exercise in Performance Portability* [4], which we will refer to as "the Tersoff paper". In this paper, a mechanism named general building blocks are designed, which is consistent in different architectures.

By identifying general building blocks, abstracting corresponding vectorization back-end, applying methods in different modes including the USER-INTEL [1] and the KOKKOS package [3], and reducing precision on well-examined calculations, authors claimed that performance will be gained and scaled well without accuracy reduction in the Tersoff paper. More specifically, three claims are made:

Accuracy Maintenance due to the limited interaction with one given atom, round-off error might not accumulate since only interactions contribute to the force.

Performance Gain speedup by applying the vectorization method can reach 2x to 3x on most CPUs, and between 3x and 5x on accelerators.

Strong Scaling the optimization scales to multi-node circumstances automatically without degrading due to the overhead, e.g. communication in the real world.

We attempt to reproduce the three claims above in this report.

2. REPRODUCTION ENVIRONMENT

Our cluster that is used to reproduce the result of the Tersoff paper is set up as listed in Table 1.

Nodes	Dell EMC PowerEdge R740 × 4
CPU per node	Intel Xeon Platinum 8176 × 2
GPU per node	NVIDIA Tesla V100 × 2
Memory per node	DDR4 2666 MT/s 16 GB × 12
Storage	Intel SSD DC P3700 1.8 TB × 2
Network	Mellanox EDR InfiniBand
Operating System	CentOS Release 7.4.1708
Kernel	Linux kernel 3.10.0

Table 1: Environment Specification

Intel Xeon Platinum 8176 is of 28 cores per socket and AVX-512 ISA including conflict detection extension. With it, we can validate or challenge the claim that higher performance can be achieved with longer vectors, 512 bits, and test if the conflict detection instructions can improve the serialized updating as the authors expected in the "future".

NVIDIA Tesla V100 GPUs provide powerful computation ability. The wrap of CUDA works similarly as $32 \times 64 = 2048$ bits SIMD, which is wide enough to evaluate the effectiveness of vectorization.

With hexa channel provided by Xeon 8176, 126 GB/s memory bandwidth can be achieved per socket. Hyperthreading is enabled to overlap memory access and computation. With Tesla V100, applications can run on an exceptional memory bandwidth – 900 GB/s. Therefore, the impact of memory bottleneck is slightly ruled out so as to focus on computational optimization.

Equipped with these recent high-performance hardware, we are able to demonstrate that the original claim "the optimization will be 'future proof' " is tenable and validate the portability of Tersoff paper's results with SIMD architectures that it focused on.

3. COMPILATION AND RUN

To best reproduce the results described, we choose to clone from the Tersoff Git repository[1], attached as an artifact in the Tersoff paper's appendix, since the code in official LAMMPS repository are kept updating and might be mixed with other optimization.

The compilers and running environment is listed as below:

Compilers	Intel C++ Compilers 2018 CUDA 9.0
MPI	Intel MPI Library 2018
Math Kernel Library	Intel MKL 2018

Table 2: Compilers and Environment

By using *icc 2018* and *CUDA 9.0*, we can verify the portability on the compiler side. But we have to do minor modifications on *intel_simd.h* to wrap SIMD data types by construction based on LAMMPS official version 11Aug17, and to change some deprecated options in *Makefile.intel_cpu* after confirming these would not change the underlying behavior. Also, the corresponding GPU arch settings are changed to be "sm_70" to support the Tesla V100.

Using the provided workflow in the Tersoff paper, we can use the scripts *build.sh* and *bench-*.sh* to compile and run the reference CPU version, different CPU versions of the optimized binary with configuration macros, optimized and original KOKKOS versions and reference GPU versions using corresponding packages and parameters.

4. REPRODUCTION WORK

Four pieces of work are done to fully reproduce the Tersoff paper's result: *Accuracy Study* is to validate the reduced precision implementations; *Single-Threaded Execution* gives the purest representation of the speedup obtained by the optimizations; *Single Node Execution* shows the results on an entire node; while *Strong Scalability* scales to a cluster with multiple nodes, which is a real-world scenario.

We evaluate the accuracy using the same *in.tersoff-acc*, and obtain the results of single-threaded execution with the same *in.tersoff* used in the Tersoff paper. For single and multiple nodes executions, besides the original *in.tersoff_bench* [4] file, a new input from the paper: *Empirical bond-order potential description of thermodynamic properties of crystalline silicon* [6] by Porter et al. is also used.

[1]github.com/HPAC/lammps-tersoff-vector

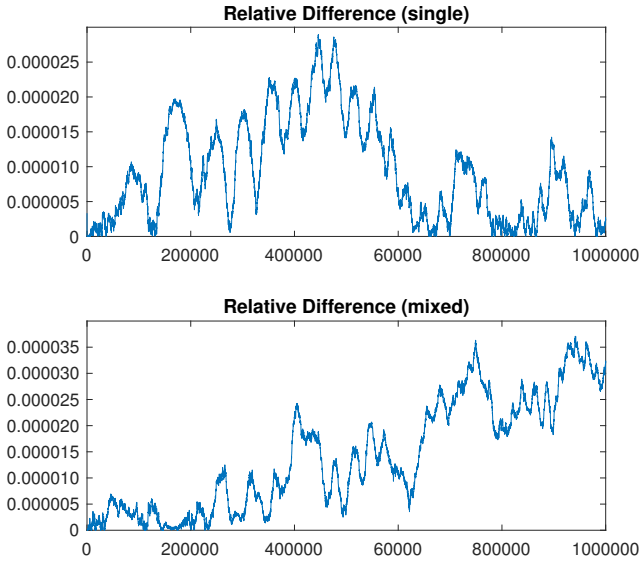

Figure 1: Relative difference between the *single/mixed* and *double* precision solvers for a system of 32 000 atoms for 10^6 timesteps on *16 processes*.

4.1 Accuracy Study

In the Tersoff paper, they created versions that compute Tersoff potential in single and mixed precision. The energy measurement in a long-running simulation was used to validate these reduced precision implementations. Here we repeat the procedure to test relative difference, including single to double and mixed to double, in order to validate that the accuracy can be maintained under our hardware and software setup.

As Fig. 1 illustrates, in a running of 16 MPI processes, exactly the same numbers of processes the Tersoff paper ran on, the deviation of both single and mixed resides within 0.004% of the double reference, slightly exceed the original results of 0.002% as shown in Fig. 3 of the Tersoff paper, but still in a reasonable range. This can be justified because in the Tersoff paper they ran on accelerators while we do not.

Several extra runs reveal that with MPI process count increasing, the relative difference gets worse. With 32 processes, the difference doubles. Fig. 2 shows that the deviation of single exceed 0.01% with 220 processes. But by comparing to the relative difference accumulated between the results of 220 processes and 16 processes using the same double precision solver, we believe that the error caused by reducing precision can be neglected.

Therefore, a change from double to single or mixed would neither cause a significant nor accumulated error. We successfully reproduced the accuracy experiment.

4.2 Execution Performance

In Tersoff paper, Fig. 4 shows single-threaded performance for all different execution modes: Ref[2], Opt-D, Opt-S

[2]as stated by Höhnerbach in an email reply, LAMMPS run without USER-INTEL package. OMP package is necessary on hyperthreading environment.

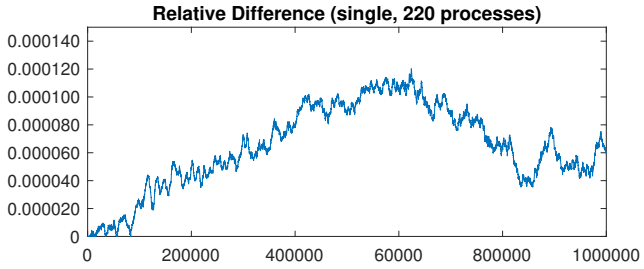

Figure 2: Relative difference between the *single* and *double* precision solvers for a system of 32 000 atoms for 10^6 timesteps on *220 processes*.

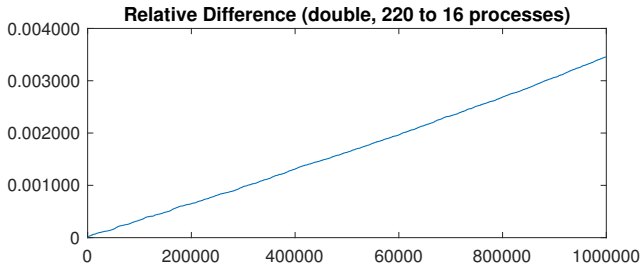

Figure 3: Relative difference between *220 processes* and *16 processes* double precision solvers for a system of 32 000 atoms for 10^6 timesteps.

and Opt-M[3]; Fig. 5 shows realistic single node performances for Ref and Opt-M; Fig. 6 presents single GPU results for two different GPUs: K20x and K40. Architectures including ARM, Kepler, Westmere, Sandy Bridge, Haswell, and Broadwell are tested, which give substance to portability.

To provide more data points as evidence, we choose Skylake (SL) with AVX-512 support to run this experiment. Cloud component provided by CycleCloud gives us a chance to have tests on another architecture, Haswell (HW), during the competition, as a replication of existing results in the Tersoff paper.

4.2.1 Single-Threaded Execution

Since the single-threaded execution can eliminate the effect of parallelism from multiple threads or nodes, it shows purely the essential speedup using vectorization optimization.

Note that the scale of the used test data is tiny, therefore sensitive to interference, we tried to disable all unused daemon on the system, and set a negative[4] nice value to the running application. With 100 data points, we can display more accurate performance metrics.

As shown in Fig. 4, both Skylake and Haswell perform expected speedup comparing Ref with Opt-*, e.g. a factor of 4.8 between Opt-S and Ref on Haswell as described in the Tersoff paper, and 5.9 on Skylake.

4.2.2 Single Node Execution

Single node execution mixes the vectorization and multi-thread speedup, introducing communication overhead and bandwidth problem. Due to the space limitations, we only present the results of Skylake in Fig. 5 with 4.4x speedup on *in.tersoff_bench*, and 0.74x on *in.porter* here. Besides, more experiments showing similar results can also be obtained on Haswell, the CPU architecture on the cloud.

[3]run with USER-INTEL package, in the default_vector flavor, set mode parameter with double (D), single (S), or mixed (M)

[4]high priority

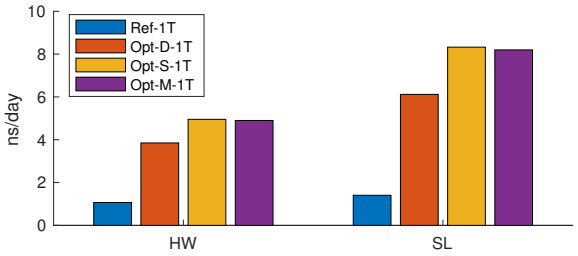

Figure 4: Performance across different CPU architectures; single-thread run on *in.tersoff*. 32 000 atoms.

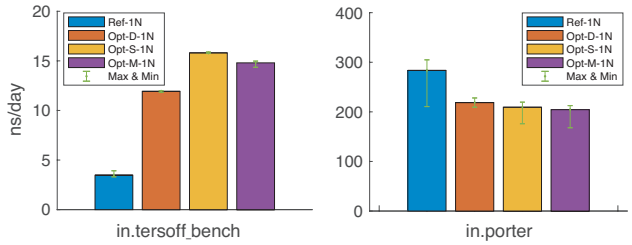

Figure 5: Performance on Skylake (SL) Intel Xeon Platinum 8176 x2; single-node multi-process run on *in.tersoff_bench* and *in.porter*

Part of the performance reduction of Porter's input [6] could be attributed to the nature of its calculation. Porter's work is to calculate the thermodynamic properties of crystalline silicon, while silicon is in a state of solid, therefore the atoms are located densely. For a process, after computing the Newton force, halo calculation need to update f_i and f_j, which might be handled by another process, so communication is required. Since the atoms in the model are dense, a process may have a lot of halo atoms sent to an adjacent MPI process, contributing to the high cost of communication.

With the Allinea MAP profiling tools provided by SCC, we can see in Fig. 6 that most of the time is consumed in MPI communication, while the Tersoff kernel used less than 1% of the total execution time, which proves that the communication cost caused by the property of the input accounts for the slowdown.

In conclusion, we can completely reproduce the expected performance increase on the exact same input provided in the Tersoff paper. But with the new input *in.porter* from Porter's paper, performance dramatically drops, due to the characteristic of the given task that requires lots of communication.

4.2.3 GPU Execution

Additionally, we perform experiments on NVIDIA Tesla V100 to test the portability of the Tersoff paper claimed, as Fig. 7 shown. 6.1x speedup is achieved, greater than the 5x in the paper, which might be attributed to the improved bandwidth.

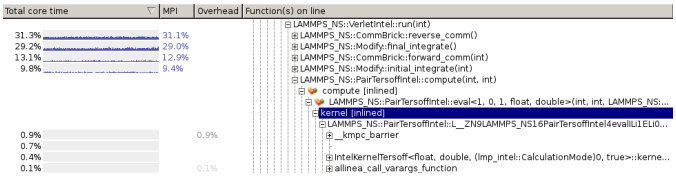

Figure 6: Hotspots measured with Allinea MAP

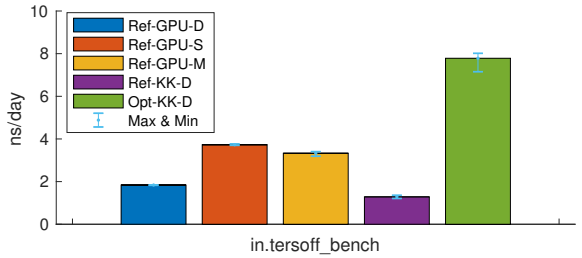

Figure 7: Performance on NVIDIA Tesla V100; single-card run on *in.tersoff_bench*

4.3 Strong Scalability

As we all know, we care more about the performance at scale rather than single thread or single node execution because they rarely appear in realistic simulations. In this section, we measure the optimization effect on our cluster to validate its scalability.

With *in.tersoff_bench* as the input file, the performance on our cluster has a nearly linear improvement as expected while nodes are increased, see Fig. 8. But we cannot reproduce the result when we use *in.porter* as input. This is due to the same communication issue described in Section 4.2.2.

5. UTILIZING AVX512-CD

AVX512-CD (Conflict Detection) is a new feature of AVX512 instruction set, which enables to compute histogram in parallel. It is mentioned that the new feature can be utilized to optimize the Newton force f accumulation by constructing non-conflict set using conflict detection instruction *_mm512_conflict* and broadcasting instruction *_mm512_broadcastm* repeatedly and then updating array f by using the non-conflict set mask. But this idea is not implemented in the Tersoff paper.

Followed by Tersoff paper's idea, we have developed three versions with single, double and mixed precision respectively. Unfortunately, no matter which version we use, we got a lower performance (overall 7% reduction) by applying AVX512-CD to both f_i and f_j update because i index is easy to conflict. Besides, there is still overall 1% performance reduction even only AVX512-CD on f_j update are used. Through more analysis, we found that conflict detection introduces more overhead than benefits because only one addition can be vectorized.

The result shows that AVX512-CD is not suitable for the Tersoff potential computation. To make full use of the new feature, more compute-intensive patterns are necessary.

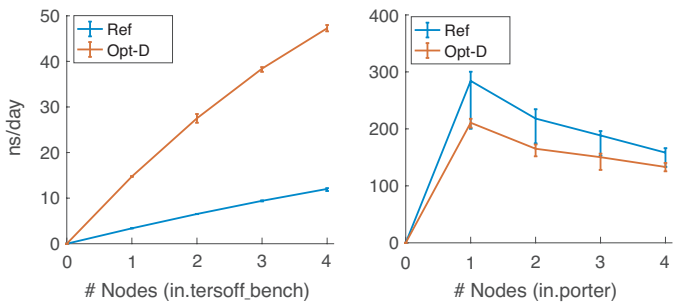

Figure 8: Strong scalability with *in.tersoff_bench* and *in.porter* as input

6. CONCLUSION

Through our effort to reproduce the Tersoff paper's result, we demonstrate that by using the input provided in the Tersoff paper, we can smoothly reproduce the accuracy analysis, the performance improvements and the strong scalability not only on the already tested platforms like Haswell architecture by using the cloud components provided by SCC, but also on our cluster equipped with the most advanced architectures, like Intel Skylake and NVIDIA Volta which was not available when the Tersoff paper released. This proves the portability of the methods proposed in the Tersoff paper.

However, we found that the performance improvements and scalability cannot be reproduced with the tasks provided in Porter's paper. After close scrutiny, it is found that the special characteristic of the task that atoms are dense causes intense communication. Therefore the communication becomes a bottleneck, causing the impossibility to have expected overall speedup with the method specified in the Tersoff paper.

7. REFERENCES

[1] W. M. Brown, J.-M. Y. Carrillo, N. Gavhane, F. M. Thakkar, and S. J. Plimpton. Optimizing legacy molecular dynamics software with directive-based offload. *Computer Physics Communications*, 195:95–101, 2015.

[2] W. M. Brown, P. Wang, S. J. Plimpton, and A. N. Tharrington. Implementing molecular dynamics on hybrid high performance computers–short range forces. *Computer Physics Communications*, 182(4):898–911, 2011.

[3] H. C. Edwards, C. R. Trott, and D. Sunderland. Kokkos: Enabling manycore performance portability through polymorphic memory access patterns. *Journal of Parallel and Distributed Computing*, 74(12):3202–3216, 2014.

[4] M. Höhnerbach, A. E. Ismail, and P. Bientinesi. The vectorization of the tersoff multi-body potential: An exercise in performance portability. In *Proceedings of the International Conference for High Performance Computing, Networking, Storage and Analysis*, SC '16, pages 7:1–7:13, Piscataway, NJ, USA, 2016. IEEE Press.

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

[6] L. J. Porter, S. Yip, M. Yamaguchi, H. Kaburaki, and M. Tang. Empirical bond-order potential description of thermodynamic properties of crystalline silicon. *Journal of applied physics*, 81(1):96–106, 1997.