Advanced Vectorization of PPML Method for Intel® Xeon® Scalable Processors

Igor Chernykh[1]✉, Igor Kulikov[1], Boris Glinsky[1], Vitaly Vshivkov[1], Lyudmila Vshivkova[1], Vladimir Prigarin[2]

[1]Institute of Computational Mathematics and Mathematical Geophysics SB RAS, 630090 Novosibirsk, Russia
[2]Novosibirsk State Technical University, 630073 Novosibirsk, Russia
chernykh@ssd.sscc.ru, kulikov@ssd.sscc.ru, gbm@sscc.ru, vsh@ssd.sscc.ru, lyudmila.vshivkova@parbz.sscc.ru, vovkaprigarin@gmail.com

Abstract. Piecewise Parabolic Method on a Local Stencil is very useful for numerical simulation of fluid dynamics, astrophysics. The main idea of the PPML method is the use of a piecewise parabolic numerical solution on the previous time step for computing the Riemann problem solving partial differential equations system (PDE). In this paper, we present the new version of PDE solver which is based on the PPML method optimized for Intel Xeon Scalable processor family. The results of performance comparison between different types of AVX-512 compatible Intel Xeon Scalable processors are presented. Special attention is paid to comparing the performance of Intel Xeon Phi (KNL) and Intel Xeon Scalable processors.

Keywords: Massively Parallel Supercomputers · Astrophysics · Code Vectorization.

1 Introduction

For the past decade, the most of research papers in high-performance computing and numerical simulation of different problems using supercomputers are dedicated to parallel algorithms, parallel programming techniques, performance analysis and tests. However, at the same time we can see that the architecture of CPUs evolves not only in number of cores. Modern CPUs have more complex core structure than ten years ago. The most recent processors have many cores/threads and the ability to implement single instructions on an increasingly large data set (SIMD width) [1]. For example, fig.1 shows core architecture of Intel Xeon Scalable Processor. From our point of view, the key factor to utilize all computational power of modern CPUs is a FMA+SIMD code optimization as well as using of OpenMP API within a multicore CPU. Intel Xeon Scalable Processor has 2 FMAs per core and AVX-512 support. Also worth noting that base AVX-512 instruction set for Intel Xeon Scalable Processor is supported by Intel Xeon Phi 7290 (KNL architecture) processor. It means that if you have AVX-512 optimized code for Intel Xeon Phi 7290, Intel Xeon Scalable processor will support your optimizations.

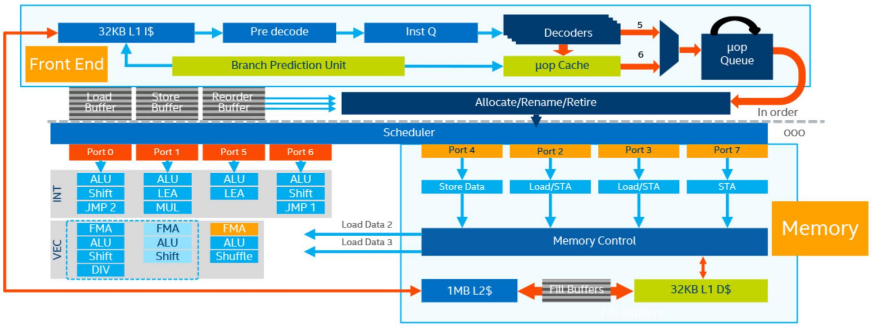


Fig. 1. Intel Xeon Scalable CPU core architecture [2].

Unlike Intel Xeon Phi 7290, Intel Xeon Scalable Processors have cores downclock in case of using AVX-512 instructions. Cores downclock depends on the core's load by AVX-512 instructions. Table 1 shows CPU frequency behavior for three kinds of Intel Xeon Gold Processors which are used for tests in this paper. As you can see from the table, some processors have low base AVX-512 frequency due to TDP value restrictions.

Table 1. Turbo frequencies for Intel Xeon Gold processors. Full load of cores.

Mode	Intel Xeon Gold 6144 (8 cores)	Intel Xeon Gold 6150 (18 cores)	Intel Xeon Gold 6154 (18 cores)
Base (without/with AVX-512) frequency	3.5GHz/2.2GHz	2.7GHz/1.9GHz	3GHz/2.1GHz
Normal turbo	4.1 GHz	3.4 GHz	3.7 GHz
AVX-512 turbo	2.8 GHz	2.5 GHz	2.7 GHz

In the next chapters, we will show the results of performance tests by our solver which is based on PPML method on this three kinds of Intel Xeon Scalable Processors.

2 Mathematical Model and Numerical Method

In our work, we use a multicomponent hydrodynamic model of galaxies considering the chemodynamics of molecular hydrogen and cooling in the following form:

$$\begin{aligned}
&\frac{\partial \rho}{\partial t} + \nabla \cdot \left(\rho \vec{u}\right) = 0 \\
&\frac{\partial \rho_{H_2}}{\partial t} + \nabla \cdot \left(\rho_{H_2} \vec{u}\right) = S\left(\rho, \rho_H, \rho_{H_2}\right) \\
&\frac{\partial \rho_{H}}{\partial t} + \nabla \cdot \left(\rho_{H} \vec{u}\right) = -S\left(\rho, \rho_H, \rho_{H_2}\right) \\
&\frac{\partial \rho \vec{u}}{\partial t} + \nabla \cdot \left(\rho \vec{u} \vec{u}\right) = -\nabla p - \rho \nabla \Phi \\
&\frac{\partial \varepsilon}{\partial t} + \nabla \cdot \left(\varepsilon \vec{u}\right) = -\left(\gamma - 1\right) \varepsilon \nabla \cdot \left(\vec{u}\right) - Q \\
&\frac{\partial E}{\partial t} + \nabla \cdot \left(E \vec{u}\right) = -\nabla \cdot \left(p \vec{u}\right) - \left(\rho \nabla \Phi, \vec{u}\right) - Q \\
&\Delta \Phi = 4\pi G \rho \\
&E = \varepsilon + \frac{\rho \vec{u}}{2} \\
&p = \left(\gamma - 1\right) \varepsilon ,
\end{aligned} \tag{1}$$

where ρ is density, ρ_H is atomic hydrogen density, ρ_{H_2} is molecular hydrogen density, $\vec{u}$ is the velocity vector, ε is internal energy, p is pressure, E is total energy, γ is the ratio of specific heats, Φ is gravity, G is the gravitational constant, S is the formation rate of molecular hydrogen, and Q is a cooling function. A detailed description of this model can be found in [3].

The formation of molecular hydrogen is described by an ordinary differential equation [4]:

$$\frac{dn_{H_2}}{dt} = R_{gr}\left(T\right) n_H \left(n_H + 2n_{H_2}\right) - \left(\xi_H + \xi_{diss}\right) n_{H_2} \tag{2}$$

where n_H is the concentration of atomic hydrogen, n_{H_2} is the concentration of molecular hydrogen, and T is temperature. Detailed descriptions of the *H*2 formation rate R_{gr} and the photodissociation ξ_H, ξ_{diss} of molecular hydrogen, can be found in [5,6]. Chemical kinetics was don with using of CHEMPAK tool [7,8]

The original numerical method based on the combination of the Godunov method, operator splitting approach and piecewise-parabolic method on local stencil was used for numerical solution of the hyperbolic equations [9]. The piecewise-parabolic method on local stencil provides the high-precision order. The equation system is solved in two stages: at the Eulerian stage, the equations are solved without advective terms and at the Lagrangian stage, the advection transport is being performed. At the Eulerian stage, the hydrodynamic equations for both components are written in the non-conservative form and the advection terms are excluded. As the result, such a system has an analytical solution on the two-cell interface. This analytical solution is used to evaluate the flux through the two-cell interface. In order to improve the precision order, the piecewise-parabolic method on the local stencil (PPML) is used. The method is the construction of local parabolas inside the cells for each hydrodynamic quantity. The main difference of the PPML from the classical PPM method is the use of the local stencil for computation. It facilitates the parallel implementation by using only one layer for subdomain overlapping. It simplifies the implementation of the boundary conditions and decreases the number of communications thus improving the scalability. The detailed description of this method can be found in [10]. The same approach is used for the Lagrangian stage. Now the Poisson equation solution is based on Fast Fourier Transform method. This is because the Poisson equation solution takes several percents of the total computation time. After the Poisson equation solution, the hydrodynamic equation system solution is corrected. It should be noticed here that the system is over defined. The correction is performed by means of the original procedure for the full energy conservation and the guaranteed entropy nondecrease. The procedure includes the renormalization of the velocity vector length, its direction remaining the same (on boundary gas-vacuum) and the entropy (or internal energy) and dispersion velocity tensor correction. Such a modification of the method keeps the detailed energy balance and guaranteed non-decrease of entropy.

3 Performance Analysis

We used Intel Advisor [11] for performance analysis of our code. Intel Advisor collects different statistics from each cycle of the code. Statistics collection consists of 3 steps: survey collection, trip count collection, visualization and/or extraction of the collected data into a report.

1. Survey collection by command line with advisor:

```
mpirun -n <number of nodes> advixe-cl -collect survey ––
trace-mpi -- ./<app_name>
```

2. Trip count collection by command line with advisor:

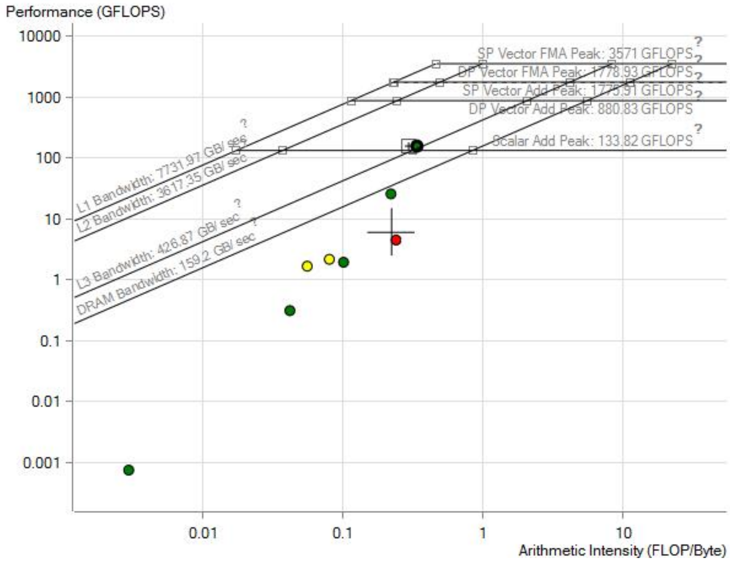
```
mpirun -n <number of nodes> advixe-cl -collect   trip-
counts -flop ––trace-mpi -- ./<app_name>
```

3. Extraction of the data in a report:

```
advixe-cl –report survey –show-all-columns –-format=text
-– report-output report.txt
```

In our research, we used RSC Tornado [12] experimental nodes with Intel Xeon Gold 6144, 6150, 6154 processors. Each node has two CPUs and 192 GB DRAM. All tests are optimized for using only OpenMP parallelization for maximum performance.

Figure 3 shows a very good correlation between results of the performance tests and AVX-512 turbo frequencies of processors being tested. Despite the highest core frequencies of Intel Xeon Gold 6144 total performance is lower than Intel Xeon Gold 6154 because of the 8 cores on a chip. The results of the same tests on Intel Xeon Phi 7290 (KNL) can be found in [13]. At this moment we achieved 202 GFLOPS from RSC Tornado-F node of NKS-1P supercomputer [14] with one Intel Xeon Phi 7290. Two Intel Xeon Gold 6154 processors have 15% better performance than Intel Xeon Phi 7290 but also they are two times more expensive than one KNL.



a)

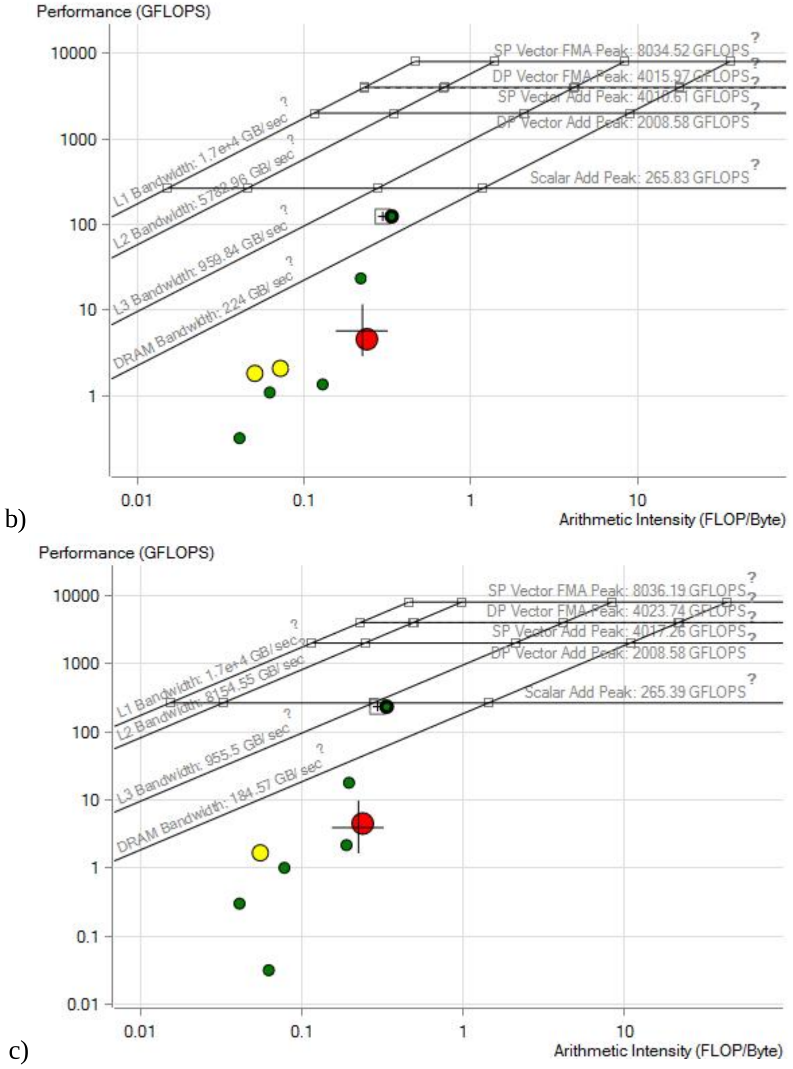


Fig. 3. Roofline chart for our code: a) 2x Intel Xeon Gold 6144 – 158 GFLOPS; b) 2x Intel Xeon Gold 6150 – 125 GFLOPS; c) 2x Intel Xeon Gold 6154 – 234 GFLOPS.

4 Conclusion

Modern processors, such as Intel Xeon Scalable, provide support for vector operations on 512-bit vectors. Applications can pack 32 double precision and 64 single precision floating point operations per second per clock cycle within the 512-bit vectors. This technology significantly expands the possibilities for solving complex scientific problems. But due to the TDP restrictions, Intel Xeon Scalable processors downclocked depending on AVX-512 instructions core load. Author’s astrophysics code which is based on the combination of the Godunov method, operator splitting approach and piecewise-parabolic method on local stencil was used for performance tests of Intel Xeon Scalable processors. This code is AVX-512 instructions set optimized by using AVX-512 intrinsics. We achieve 158 GFLOPS on 2x Intel Xeon Gold 6144 processors, 125 GFLOPS on 2x Intel Xeon Gold 6150 processors, 234 GFLOPS on 2x Intel Xeon Gold 6154 processors. These results are in a good accordance with AVX-512 turbo frequencies of processors being tested. Two Intel Xeon Gold 6154 processors have 15% better performance than one Intel Xeon Phi 7290 which was tested earlier.

Acknowledgments. This work was partially supported by RFBR grants 18-07-00757, 18-01-00166 and 16-07-00434. Methodical work was partially supported by the Grant of the Russian Science Foundation grant 16-11-10028.

References

1. Vectorization: A Key Tool To Improve Performance On Modern CPUs. https://software.intel.com/en-us/articles/vectorization-a-key-tool-to-improve-performance-on-modern-cpus
2. Intel Xeon Scalable Debuts. https://hothardware.com/reviews/intel-xeon-scalable-processor-family-review?page=2
3. Vshivkov V.A., Lazareva G.G., Snytnikov A.V., Kulikov I.M., Tutukov A.V.: Hydrodynamical code for numerical simulation of the gas components of colliding galaxies. Astrophysical Journal. Supplement Series, 194(47), 1-12 (2011)
4. Bergin E.A., Hartmann L.W., Raymond J.C., Ballesteros-Paredes J.: Molecular cloud formation behind shock waves. Astrophys. J.,612, 921-939 (2004)
5. Khoperskov S.A., Vasiliev E.O., Sobolev A.M., Khoperskov A.V.: The simulation of molecular clouds formation in the Milky Way. Monthly Notices of the Royal Astronomical Society, 428 (3), 2311-2320 (2013)
6. Glover S., Mac Low M.: Simulating the formation of molecular clouds. I. Slow formation by gravitational collapse from static initial conditions. Astrophysical Journal. Supplement Series, 169, 239-268 (2006)
7. Chernykh, I., Stoyanovskaya, O., Zasypkina, O.: ChemPAK software package as an environment for kinetics scheme evaluation. Chemical Product and Process Modeling, 4(4) (2009)
8. Snytnikov, V.N., Mischenko, T.I., Snytnikov, V., Chernykh, I.G.: Physicochemical processes in a flow reactor using laser radiation energy for heating reactants. Chemical Engineering Research and Design, 90(11), 1918-1922 (2012)
9. Godunov S.K., Kulikov I.M.: Computation of discontinuous solutions of fluid dynamics equations with entropy nondecrease guarantee. Computational Mathematics and Mathematical Physics, 54, 1012–1024 (2014)
10. Kulikov I., Vorobyov E.: Using the PPML approach for constructing a low-dissipation, operator-splitting scheme for numerical simulations of hydrodynamic flows. J. Comput. Phys., 317, 316—346 (2016)
11. Intel Advisor. https://software.intel.com/en-us/intel-advisor-xe

12. RSC Tornado. http://www.rscgroup.ru/en/our-technologies/267-rsc-tornado-cluster-architecture
13. Glinskiy, B., Kulikov, I., Chernykh, I.: Improving the Performance of an AstroPhi Code for Massively Parallel Supercomputers Using Roofline Analysis. Communications in Computer and Information Science, vol. 793, pp. 400-406 (2017)
14. Siberian Supercomputer Center ICMMG SB RAS. http://www2.sscc.ru