OMPEGAS: Optimized Relativistic Code for Multicore Architecture

Elena N. Akimova 1,2,*, Vladimir E. Misilov 1,2, Igor M. Kulikov 3 and Igor G. Chernykh 3

1 Krasovskii Institute of Mathematics and Mechanics, Ural Branch of RAS, S. Kovalevskaya Street 16, 620108 Ekaterinburg, Russia; v.e.misilov@urfu.ru
2 Department of Information Technologies and Control Systems, Institute of Radioelectronics and Information Technology, Ural Federal University, Mira Street 19, 620002 Ekaterinburg, Russia
3 Institute of Computational Mathematics and Mathematical Geophysics SB RAS, 630090 Novosibirsk, Russia; kulikov@ssd.sscc.ru (I.M.K.); chernykh@parbz.sscc.ru (I.G.C.)

* Correspondence: aen15@yandex.ru

Abstract: The paper presents a new hydrodynamical code, OMPEGAS, for the 3D simulation of astrophysical flows on shared memory architectures. It provides a numerical method for solving the three-dimensional equations of the gravitational hydrodynamics based on Godunov’s method for solving the Riemann problem and the piecewise parabolic approximation with a local stencil. It obtains a high order of accuracy and low dissipation of the solution. The code is implemented for multicore processors with vector instructions using the OpenMP technology, Intel SDLT library, and compiler auto-vectorization tools. The model problem of simulating a star explosion was used to study the developed code. The experiments show that the presented code reproduces the behavior of the explosion correctly. Experiments for the model problem with a grid size of $128 \times 128 \times 128$ were performed on an 16-core Intel Core i9-12900K CPU to study the efficiency and performance of the developed code. By using the autovectorization, we achieved a 3.3-fold increase in speed in comparison with the non-vectorized program on the processor with AVX2 support. By using multithreading with OpenMP, we achieved an increase in speed of 2.6 times on a 16-core processor in comparison with the vectorized single-threaded program. The total increase in speed was up to ninefold.

Keywords: computational astrophysics; type II supernova; high performance computing; OpenMP; SIMD

MSC: 65Y05; 85A30

1. Introduction

Supernovae are the main source of the “elements of life”. In particular, the supernovae are the birthplaces for such elements as carbon, oxygen, and ferrum, which are the basis of all life forms. The great energy output of the supernovae allows them to spread the heavy elements great distances. This contributes to synthesis of complex chemical compounds up to organic ones [1,2].

Together with the various mechanisms of supernova explosions and their material compositions [3–13], it remains important to study the hydrodynamics of supernova explosions. Naturally, because of proportion of the scales between the stars, the supernova remnant, and the core, simulation of such an explosion requires the most powerful supercomputers. Analysis of the top 10 supercomputers in the world shows the prevalence of hybrid architectures, primarily equipped with graphics accelerators and vector processors.

Development of codes for such architectures is a difficult problem. It requires not only application of the appropriate technologies but also a special structure of the numerical method and mathematical models. Several codes adapted for graphics processors (GPU) or the Intel Xeon Phi accelerators have been developed. We review the most interesting solutions.
The WOMBAT code [14] uses a combination of the Message Passing Interface (MPI), Open Multi-Processing (OpenMP), and high-level Single Instruction Multiple Data (SIMD) vectorization technologies. The WOMBAT code is remarkable for its orientation to the target architecture of the supercomputer, which will run it. The code utilizes one-sided MPI communications, which allow one to efficiently implement the exchange of the subdomains overlaps. The specialized Cray MPI is used for optimization of the interprocess communications. The code also uses high-level SIMD vectorization. It appears that the WOMBAT code contains the requirements for modern program implementation of astrophysical codes. The previously developed codes will be used for a long time, but the future is for codes oriented to the specific classes of the supercomputer architectures.

The CHOLLA code [15–17] was implemented to conduct numerical experiments using GPUs. It is based on the Corner Transport Upwind method, the essence of which is the propagation of the counterflux scheme in the multidimensional case [18,19]. For storing the computational grid in a GPU, the structure of a mesh containing all the hydrodynamical parameters is used. It is stated that such locality of the data allows one to obtain a more efficient application of the global memory of the graphics processor. Calculation of the time step was performed on graphics processors using the NVIDIA Compute Unified Device Architecture (CUDA) extensions.

The GAMER code [20] solves aerodynamic equations applying the Adaptive Mesh Refinement (AMR) approach to graphics accelerators. For solving aerodynamic equations, the Total Variational Diminishing method is used. For solving the Poisson equation, a combination of the fast Fourier transform method and successive over relaxation is used. The main feature of this software is probably of implementation of the AMR approach on GPUs. Thus, the regular structure of the grid naturally falls onto the GPU architecture, while the tree structure requires a special approach. This approach is based on the application of octets to define the grid. Under this, each octet is projected to a separate thread in the GPU. The main problem is in forming the fake meshes for the octets. Note that solving this problem takes around 63% of the time. However, this procedure can be implemented independently for each octet. The GAMER code was further optimized [21] and extended to solving the magnetic hydrodynamic equations [22].

Approaches such as those described above were implemented in the FARGO3D [23] and the SMAUG [24] codes. The authors of the present work also implemented the GPUPEGAS code [6] for supercomputers with graphics processors and the AstroPhi code for the Intel Xeon Phi in the offload mode [25] and in the native mode using low-level vectorization [26].

The necessity of more efficient utilization of the graphics and vector processors and accelerators will result in creation of new codes. One code is presented in this work. We develop the OMPEGAS code for numerically solving the special relativistic hydrodynamics equations. The numerical algorithm is based on Godunov’s method and the piecewise parabolic method on the local stencil [27,28]. This technique has been successfully applied to a number of astrophysical problems [29,30]. Its main advantages are a high order of accuracy and high efficiency of calculations. The code was developed for shared memory multicore systems with vectorization capabilities. This make it possible to perform experiments on a wide range of computers, from laptops to nodes of cluster supercomputers. Additionally, in future, a hybrid implementation is possible for massive parallel systems.

In Section 2, we provide the hydrodynamical model of the astrophysical flows and describe the numerical method for solving the hydrodynamical equations. Section 3 is devoted to detailing the architecture of the OMPEGAS code and to the study of its performance. Naturally, the code development raised several issues for discussion, which are considered in Section 4. Section 5 is devoted to the numerical experiments. Section 6 concludes the paper.
2. Numerical Model

To describe the gas dynamics, we use the numerical model presented in work [31]. The physical (primitive) variables are: the density $\rho$, vector of velocity $\vec{v}$, and pressure $p$. When describing the evolution of gas in the model of special relativistic hydrodynamics, we introduce the concept of special enthalpy $h$, which is determined by the formula

$$ h = 1 + \frac{\gamma}{\gamma - 1} \frac{p}{\rho}, $$

where $\gamma$ is the adiabatic exponent. The relativistic speed of sound $c_s$ is determined by the formula

$$ c_s^2 = \gamma \frac{p}{\rho h}. $$

Let us choose the speed of light to be $c \equiv 1$ as the dimensionless unit speed. In this case, the Lorentz factor $\Gamma$ will be determined by the formula

$$ \Gamma = \frac{1}{\sqrt{1 - (v/c)^2}} = \frac{1}{\sqrt{1 - v^2}}. $$

Thus, the speed module is limited to one. Let us introduce concepts of the conservative variables relativistic density $D = \Gamma \rho$ and the relativistic momentum $M_j = \Gamma^2 \rho \bar{v}_j$, where $\bar{v}_j$ are components of the velocity vector $\bar{v}$ at $j = 1, 2, 3$, and the total relativistic energy $E = \Gamma^2 \rho h - p$.

Under the absence of the relativistic velocities, the Lorentz factor is taken to be unity. Then, the speed of sound of the ideal gas is calculated by the formula

$$ c_s^2 = \gamma \frac{p}{\rho}. $$

The conservative hydrodynamic quantities take the form: $D = \rho$ is density, $M_j = \rho v_j$ is the momentum, and $E = \rho \bar{v}_j^2 + \frac{p}{\gamma - 1}$ is the total energy.

The equations of the Newtonian and relativistic hydrodynamics in the form of the conservation laws are written in the same form

$$ \frac{\partial D}{\partial t} + \frac{\partial (D \bar{v}_k)}{\partial x_k} = 0, $$

$$ \frac{\partial M_j}{\partial t} + \frac{\partial \left( M_j \bar{v}_k + p \delta_{jk} \right)}{\partial x_k} = 0, $$

$$ \frac{\partial E}{\partial t} + \frac{\partial (E + p) \bar{v}_k}{\partial x_k} = 0, $$

where $\delta_{jk}$ is the Kronecker symbol. The exactly general form of the hyperbolic equations is important to us, since we construct a new modification of the numerical scheme for it.

For complex relativistic objects, such as jets, supernovae, and collapsing stars, the correct simulation of the shock waves is of great importance [32–37]. Usually, to solve the equations, the Roe-type, Rusanov-type, or the Harten–Lax-van Leer schemes are used [38–40]. Recently, a unified approach for such problems was proposed [41]. In our code, we use the piecewise parabolic method on a local stencil [29] and the combination of the Roe scheme [31] and the Rusanov scheme [42].
3. Parallel Implementation

In this section, we describe the developed OMPEGAS code for the hydrodynamics simulation. In our implementation, we use OpenMP technology, the Intel SIMD Data Layout Template (SDLT) library, and the auto-vectorization directives of the Intel C++ Compiler Classic.

3.1. Domain Decomposition

In our implementation, we use the following domain decomposition approach. The uniform grid in the Cartesian coordinates for solving the hydrodynamical equations allows one to use arbitrary Cartesian topology for the domain decomposition. This structure of the computation provides great potential for scalability. The OMPEGAS code uses the multilevel multidimensional decomposition of the domain. On one coordinate, the inner fine decomposition is performed by vectorization, while decomposition of two other coordinates is performed by two collapsed OpenMP loops.

Note that our method also allows one to perform additional decomposition by one of the coordinates using the MPI with one layer for subdomains overlapping.

Figure 1 shows the geometric domain decomposition implemented by the OpenMP and vectorization.

![Geometric domain decomposition](image)

Figure 1. Geometric domain decomposition.

3.2. OpenMP + SIMD Implementation

The following algorithm for computing the single time step is used (with the corresponding function names used in the code):

1. Find a minimum time step between all spatial cells (computational_tau()).
2. Construct the local parabolic approximations of primitive variables for each cell, consider the boundary conditions (build_parabola_sdlt()).
3. Solve the Riemann problem for each cell (eulerian_stage()).
4. Process the boundary conditions for the obtained conservative variables.
5. Recover the primitive variables for each cell.
6. Process the boundary conditions for the primitive variables (boundary_mesh()).

3.3. Data Structures

To optimize the data storage for the parabolic coefficients, we used the Intel SDLT library [43]. It allows one to achieve the performance of the vectorized C++ code by automatically converting the Array of Structures to the Structure of Arrays by aligning the data to the SIMD words and cache lines and providing the $n$-dimensional containers. The code is shown in Listing 1.
Listing 1. Declaration of data structures for the parabolic approximations.

```c
#include <sdlt/sdlt.h>

// type of parabola
struct parabola
{
  // parameters of parabola
  real ulr, ull, dq, dq6, ddq;
  // right integral of parabola
  inline real right_parabola(real ksi, real h) const
  // left integral of parabola
  inline real left_parabola(real ksi, real h) const
  // build local parabola
  inline void local_parabola(real uvalue)
  // reconstruct parabola
  inline void reconstruct_parabola(real uvalue)
};

// type of mesh
struct mesh
{
  // value on mesh
  real u;
  // parabols
  parabola parx, pary, parz;
};

// invoking SDLT macros for structures
SDLT_PRIMITIVE(parabola, ulr, ull, dq, dq6, ddq)
SDLT_PRIMITIVE(mesh, u, parx, pary, parz)

// defining the 3D container type
typedef sdlt::n_container<mesh, sdlt::layout::soa<1>, 
  sdlt::n_extent_t<int, int, int>> ContainerMesh3D;
```

3.4. Auto Vectorization

To study the code for vectorization, we used the Intel Advisor tool [44]. The initial analysis showed that the automatic vectorization was prevented due to the supposed dependencies and function calls. To avoid this, we used the directive ‘#pragma ivdep’ for loops and ‘#pragma forceinline recursive’ for functions. This allowed the compiler to vectorize automatically the relevant loops.

The next recommendation by the Advisor was optimizing the arithmetic operations. The costliest procedure of the Riemann solver performed 750 divisions in the loop body. After using the compiler option ‘\Qprec-div-‘, this number was reduced approximately to 300. Figures 2 and 3 show the result of the Advisor survey before and after applying this option. Note that the vectorization efficiency of the Riemann solver loop (last row) rose from 2.5 times to nearly perfect fourfold.

The procedures for constructing the parabolic approximations [29] used a lot of branching. Thus, their vectorization potential was limited.
3.5. OpenMP Parallelization

Parallelization was performed by distributing the workload between the OpenMP threads. Listing 2 shows the typical usage of ‘#pragma omp for’ directive for the Riemann solver loop. Two outer loops were collapsed and parallelized using the OpenMP. The inner loop was vectorized. The scheduling was set to ‘runtime’, which allowed us to tune the program without recompiling.

Listing 2. Riemann solver loop.

```c
#pragma omp for schedule(runtime) collapse(2)
for (i = 1; i < NX - 1; i++)
  for (k = 1; k < NY - 1; k++)
  {
    // Riemann solver
    #pragma ivdep
    for (l = 1; l < NZ - 1; l++)
      SRHD_Lamberts_Riemann(
```

3.6. Code Research

The numerical experiments were performed on a 16-core Intel Core i9-12900K processor with AVX/AVX2 support and DDR5-6000 memory. A mesh with the size of $128 \times 128 \times 128$ was used to solve the evaluation. The computing times presented below exclude the time spent on the initialization, finalizing, and file operations.

3.6.1. Performance of Single Thread Using Vectorization

Table 1 shows the computing time of a single-threaded program with various compilation settings. To disable auto-vectorization, we used the `#vec` option, and to use the precise or optimized divisions, we used the `#prec-div` or `#precc-div` options, respectively.
3.6.2. Threading Performance

For studying the threading performance, we used the speedup $S_m = T_1 / T_m$ coefficient, where $T_m$ is the calculation time on $m$ threads for the same problem.

<table>
<thead>
<tr>
<th>SIMD Length</th>
<th>Time [Minutes]</th>
<th>Precise Division</th>
<th>Optimized Division</th>
</tr>
</thead>
<tbody>
<tr>
<td>No vectorization</td>
<td>25</td>
<td>14</td>
<td></td>
</tr>
<tr>
<td>4×double (AVX2)</td>
<td>13.3</td>
<td>7.6</td>
<td></td>
</tr>
</tbody>
</table>

Table 2 shows the results of solving the problem with the $128 \times 128 \times 128$ mesh on various numbers of the OpenMP threads. The table contains the computing time of the program and speedup for various numbers of the OpenMP threads. It also shows the memory bandwidth utilization and performance in FLOPs, as well as the maximum values obtained via the Intel Advisor. Note that these two values were affected by the profiling overhead.

<table>
<thead>
<tr>
<th>Number $m$ of OpenMP Threads</th>
<th>Time $T_m$ [Minutes]</th>
<th>Speedup $S_m$</th>
<th>Memory Bandwidth Utilization [GB/s]</th>
<th>Performance [GFLOPs]</th>
</tr>
</thead>
<tbody>
<tr>
<td>1 (Serial)</td>
<td>7.6</td>
<td>—</td>
<td>19</td>
<td>16</td>
</tr>
<tr>
<td>2</td>
<td>4.8</td>
<td>1.6</td>
<td>31</td>
<td>26</td>
</tr>
<tr>
<td>4</td>
<td>3.4</td>
<td>2.27</td>
<td>48</td>
<td>39</td>
</tr>
<tr>
<td>8</td>
<td>3.0</td>
<td>2.53</td>
<td>59</td>
<td>47</td>
</tr>
<tr>
<td>16</td>
<td>2.9</td>
<td>2.64</td>
<td>70</td>
<td>49</td>
</tr>
<tr>
<td>24</td>
<td>2.9</td>
<td>2.61</td>
<td>65</td>
<td>48</td>
</tr>
<tr>
<td>Hardware maximum</td>
<td></td>
<td></td>
<td>88</td>
<td>856</td>
</tr>
</tbody>
</table>

4. Discussion

In this section, we define several important (as we see it) aspects that were not noted above. In the present case, these issues were not primarily important, but they will be addressed in further works.

1. In this work, we used the statement of the special relativistic hydrodynamical model in the form of the mass, moment, and total mechanical energy conservation laws. This did not take into account the magnetic fields. In the future, we may implement the magnetohydrodynamical model.

2. The Core i9-12900K CPU is based on heterogeneous architecture. It included 8 high-performance cores and 8 power-efficient cores. In future, such systems will probably gain more popularity. One of the problems was the difference in performance between cores resulting in the thread imbalance. For our code changing the OpenMP scheduling from ‘static’ to the ‘dynamic’, reduced the execution time from 185 to 176 s, an improvement of 5%. Note that while the high-performance cores support the Hyper-Threading (allowing the CPU to run 24 hardware threads), the best results are achieved by the 16 OpenMP threads.

3. Another challenge was in the fact that the implementation was memory-bound. We investigated the performance of the most costly procedure, namely, the Riemann solver. Table 3 presents the memory bandwidth utilization and performance in the GFLOPS for the parallel implementation of this procedure with various numbers of OpenMP threads. The values were obtained by Intel Advisor.
The table shows that the performance was limited by the memory bandwidth. The roofline analysis by the Intel Advisor (presented in Figure 4) confirmed that the implementation was bound mostly by the memory.

Table 3. Performance of the Riemann solver.

<table>
<thead>
<tr>
<th>Number (m) of OpenMP Threads</th>
<th>Memory Bandwidth Utilization [GB/s]</th>
<th>Performance [GFLOPs]</th>
</tr>
</thead>
<tbody>
<tr>
<td>1 (Serial)</td>
<td>12</td>
<td>27</td>
</tr>
<tr>
<td>2</td>
<td>24</td>
<td>50</td>
</tr>
<tr>
<td>4</td>
<td>37</td>
<td>78</td>
</tr>
<tr>
<td>8</td>
<td>52</td>
<td>110</td>
</tr>
<tr>
<td>16</td>
<td>70</td>
<td>121</td>
</tr>
<tr>
<td>24</td>
<td>65</td>
<td>114</td>
</tr>
<tr>
<td>Hardware maximum</td>
<td>88</td>
<td>856</td>
</tr>
</tbody>
</table>

Figure 4. Intel Advisor Roofline Analysis graph for the Riemann Solver procedure.

There are several ways to overcome this problem.

- Utilize hardware with a higher memory bandwidth, such as graphics processors. Modern GPUs’ memory bandwidth is usually much higher than that in CPUs.
- Use distributed memory systems. This approach would provide the effective summation of the bandwidths of the individual computers. It would also allow one to solve larger problems that do not fit into a single unit. Our numerical method and the domain decomposition approach provide an efficient way to implement the MPI parallelization.
- The OmpSs programming model [45] may be used to develop parallel code for heterogeneous computing systems with CPUs and GPUs.
- Increase the arithmetic intensity of the code. This may be achieved by using higher-order numerical methods.
- Application of specialized data structures tailored for the task. For example, the Locally Recursive non-Locally Asynchronous algorithms with the Function-
ally Arranged Shadow structure [46] provide better cache reuse and arithmetic intensity for stencil methods than the classic nested loop traversal.

5. Model Problem: Hydrodynamical Simulation of Star Explosion

Let us consider the model problem of a hydrodynamical explosion. At the center of the star (defined by the equilibrium density profile), we induce the point blast with a given dispersion of the density \( \delta \rho / \rho \approx 2\% \). Further, all equations and results of the simulation are presented in the dimensionless form.

As the initial data for the profile of the star, let us take ones of the hydrostatic equilibrium configuration. In spherical coordinates with the gravitational constant \( G \equiv 1 \), the data have the form

\[
\frac{\partial p}{\partial r} = -\frac{M(r)\rho}{r^2}, \quad \frac{\partial M}{\partial r} = 4\pi r^2 \rho.
\]

The initial density distribution is chosen as follows:

\[
\rho_0(r) = \begin{cases} 
1 - r, & r \leq 1, \\
0, & r > 1.
\end{cases}
\]

Thus, the initial distribution of the pressure and gravitational potential have the form

\[
p_0(r) = \begin{cases} 
-\frac{\pi r^2 (9r^2 - 28r + 24)}{36} + \frac{5\pi}{36}, & r \leq 1, \\
0, & r > 1,
\end{cases}
\]

\[
\Phi_0(r) = \begin{cases} 
-\frac{\pi}{3} (r^3 - 2r^2), & r \leq 1, \\
-\frac{\pi}{3r}, & r > 1.
\end{cases}
\]

We shall use this profile to define the initial configuration of the star.

Figure 5 shows the results of the simulation and the density profiles at various instants. Despite introducing the random perturbations of the star, the physics was dominated by the process of the point blast similar to the Sedov problem.
Figure 5. Density of the supernova in the equatorial cut of the star at the dimensionless instants of time 0 (top left), 0.1 (top right), 0.2 (bottom left), and 0.4 (bottom right).

6. Conclusions

We presented a novel hydrodynamic OMPEGAS code for the three-dimensional simulation of astrophysical flows using multicore processors with vector instructions. By using auto-vectorization, we achieved a 3.3-fold increase in speed in comparison with the non-vectorized program on a processor with AVX2 support. By using multithreading with the OpenMP, we achieved an increase in speed of 2.6 times on a 16-core processor in comparison with the vectorized single-threaded program. In the total, an increase in speed of nearly nine times was achieved for the test problem. The numerical experiment was performed for the model problem of the hydrodynamical simulation of a star explosion.


Funding: The work of the third author (I.M.K.) and fourth author (I.G.C.) was supported by the Russian Science Foundation (project no. 18-11-00044). The first author (E.N.A.) and second author (V.E.M.) received no external funding.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The data presented in this study are the model data. Data sharing is not applicable to this article.

Acknowledgments: We would like to thank the editors and reviewers for providing the valuable comments on our article.

Conflicts of Interest: The authors declare no conflicts of interest.
Abbreviations
The following abbreviations and symbols are used in this manuscript:

- AVX: Advanced Vector Extensions
- CPU: Central Processing Unit
- FLOPs: Floating Point Operations per Second
- GPU: Graphics Processing Unit
- MPI: Message Passing Interface
- OpenMP: Open Multi-Processing
- SDLT: SIMD Data Layout Templates
- SIMD: Single Instruction Multiple Data
- \( S_m \): Speedup coefficient
- \( T_m \): Execution time

References
2. Tutukov, A.V.; Dryomova, G.N.; Dryomov, V.V. Hypervelocity stars: theory and observations. Physics-Uspekhi 2021, 64, 967. [https://doi.org/10.3367/UFNe.2020.11.038992]


